ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΑΣΦΑΛΕΙΑ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Κακόβουλη Μηχανική Μάθηση
(Adversarial Machine Learning)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Χριστίνας Κουδέρη

Επιβλέπων καθηγητής: Παναγιώτης Ριζομυλιώτης

Σάμος, Οκτώβριος 2020

Η σελίδα αυτή είναι σκόπιμα λευκή.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή Παναγιώτη Ριζομυλιώτη για τα εποικοδομητικά σχόλια και τις συμβουλές του. Επίσης, θέλω να ευχαριστήσω την οικογένεια και τους φίλους μου για την υπομονή, την κατανόηση και την στήριξή τους κατά τη διάρκεια των μεταπτυχιακών σπουδών μου και την συγγραφή της παρούσας διπλωματικής εργασίας. Τέλος, θερμά θέλω να ευχαριστήσω τους καλούς μου φίλους Ειρήνη και Μωυσή που όλα αυτά τα χρόνια με βοήθησαν σε ότι και αν χρειάστηκα.

Κουδέρη Χριστίνας

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

Ακρωνύμια

T-FGSM	Targeted Fast Gradient Sign Method
T-IGSM	Targeted Iterative Gradient Sign Method
GAN	Generative Adversarial Network
FNN	Feedforward Neural Network
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
RP2	Robust Physical Perturbations
JSMA	Jacobian-based Saliency Map
UPSET	Universal Perturbations for Steering to Exact Targets
ANGRI	Antagonistic Network for Generating Rogue Images
DNN	Deep Neural Network

Περίληψη

Τα συστήματα μηχανικής μάθησης καθώς προσφέρουν μια ιδιαίτερη ευελιξία, χρησιμοποιούνται σε πλήθος εφαρμογών τα τελευταία χρόνια. Και ενώ πολύ συχνά διάφορα μοντέλα μηχανικής μάθησης χρησιμοποιούνται για την αντιμετώπιση επιθέσεων, όπως παραδείγματος χάριν σε συστήματα ανίχνευσης εισβολών, ταυτόχρονα οι ίδιοι αλγόριθμοι μηχανικής μάθησης σε διάφορες μορφές τους μπορούν να αποτελέσουν οι ίδιοι στόχο επίθεσης από κάποιον κακόβουλο χρήστη. Τα προβλήματα που δημιουργούν αυτές οι επιθέσεις στην λειτουργία των συστημάτων μηχανικής μάθησης είναι ποικίλα. Μπορούν να προκαλέσουν απώλεια λειτουργικότητας ή λανθάνουσα λειτουργικότητα στις διάφορες εφαρμογές που στηρίζονται σε μοντέλα μηχανικής μάθησης με αποτέλεσμα ιδιαιτέρως σημαντικά κόστη τόσο στις επιχειρήσεις που τις παρέχουν όσο και στους χρήστες που τις χρησιμοποιούν. Επίσης ένα σημαντικό ζήτημα που προκύπτει από τέτοιου είδους επιθέσεις είναι αυτό της ιδιωτικότητας, καθώς η μηχανική μάθηση στηρίζεται σε πλήθος δεδομένων και στις περισσότερες των περιπτώσεων αυτά τα δεδομένα είναι προσωπικά δεδομένα χρηστών.

Στην παρούσα διπλωματική εργασία θα γίνει μελέτη και αναφορά στις διάφορες επιθέσεις απέναντι στην μηχανική μάθηση σήμερα αλλά και στις τεχνικές άμυνας απέναντι σε αυτές, με ιδιαίτερη έμφαση στις επιθέσεις διαφυγής και τις επιθέσεις δηλητηρίασης δεδομένων, ενώ ταυτόχρονα δίνονται και παραδείγματα από διάφορες επιθέσεις στον πραγματικό κόσμο. Επίσης, αναφορά θα γίνει και στο ιδιαιτέρως σημαντικό ζήτημα της ιδιωτικότητας που προκύπτει και σε διάφορες τεχνικές για την αποφυγή τέτοιων περιπτώσεων απώλειας ή αποκάλυψης προσωπικών δεδομένων.

Λέξεις Κλειδιά: *κακόβουλη μηχανική μάθηση, επιθέσεις διαφυγής, επιθέσεις δηλητηρίασης δεδομένων, ιδιωτικότητα στη μηχανική μάθηση*

Abstract

Machine Learning Systems are being widely used over the last years, as they offer great agility.

Even though most of the machine learning models are being used to protect systems against attacks, as in example for intrusion detection systems, at the same time the same machine learning algorithms are seen as targets of attack from malicious users. Those attacks executed against machine learning algorithms may cause various malfunctions. For example, applications that rely on machine learning models may suffer severe loss of functionality or impairment, resulting in high costs both for the enterprises offering them, and end users who are using them. Additionally, it is essential to mention the privacy issues that arise from such attacks, as machine learning depends on a great size of data and information, which is most of the times characterized as personal or sensitive.

The master thesis examines the various attacks against machine learning systems of our days, as well as the defense techniques developed. Special emphasis is given to evasion attacks and data poisoning attacks, as well as attacks in the real world are used as examples. Last but not least, the thesis also analyzes the very important issue of protection of personal data and privacy resulting from these attacks and the techniques used to avoid such cases of data loss or sensitive data exposure and leakage.

Keywords: *adversarial machine learning, evasion attacks, data poisoning attacks, privacy in machine learning*

1

Εισαγωγή

Η τεχνητή νοημοσύνη (artificial intelligence) καταλαμβάνει όλο και μεγαλύτερο ρόλο στην καθημερινότητά μας μέρα με τη μέρα. Συγκεκριμένα, ανθίζει σε όλο και περισσότερους τομείς, από την αυτόνομη οδήγηση έως και τα συστήματα προτάσεων ταινιών, και από το ρομποτικό εμπόριο έως την έξυπνη διάγνωση. Έκπληξη δεν προκαλεί το γεγονός πως έχει εφαρμοστεί σε τομείς που σχετίζονται με την ασφάλεια συστημάτων, όπως είναι το φιλτράρισμα μηνυμάτων, η αυθεντικοποίηση προσώπου και άλλα.

Η μηχανική μάθηση αποτελεί τη βασική προσέγγιση για την επίτευξη της τεχνητής νοημοσύνης και παρέχει την εύκολη και γρήγορη εμφύτευση νοημοσύνης σε μηχανές με χρήση απλών ετικετών και χωρίς να χρειάζεται το ξεκαθάρισμα της λογικής και των θεωριών πίσω από τα δεδομένα. Εξαιτίας αυτής της ευκολίας άρχισε να χρησιμοποιείται σχεδόν σε κάθε τομέα. Καθημερινά συλλέγονται δεδομένα από χρήστες προκειμένου να εκπαιδευτούν μοντέλα, τα οποία χρησιμοποιούνται για την καλύτερη εξυπηρέτηση των χρηστών.

Τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks), μια κατηγορία αλγορίθμων της μηχανικής μάθησης απέκτησαν μεγάλη επιτυχία σε διάφορους τομείς, εξαιτίας της ακριβείας τους κυρίως, καθιστώντας τα ιδιαιτέρως δημοφιλή. Ενώ πολλά από τα ιδιαίτερα χαρακτηριστικά τους είναι αυτά που ευθύνονται για την επιτυχία τους, τα ίδια είναι που σε πολλές περιπτώσεις τα καθιστούν και ευπαθή απέναντι σε κακόβουλες ενέργειες.

Αυτός ο κίνδυνος που αντιμετωπίζουν τα συστήματα μηχανικής μάθησης αναφέρεται ως «κακόβουλη μηχανική μάθηση», καθώς τα συστήματα τεχνητής νοημοσύνης μπορούν να εξαπατηθούν και να κάνουν εσφαλμένες αξιολογήσεις. Μια επίθεση σε ένα σύστημα μηχανικής μάθησης μπορεί να σημαίνει την εισαγωγή κακόβουλα σχεδιασμένων δεδομένων καθώς το μοντέλο εκπαιδεύεται ή καθώς το μοντέλο λειτουργεί με σκοπό αυτό να κάνει λανθασμένες προβλέψεις.

Ορισμένα μοντέλα μηχανικής μάθησης που χρησιμοποιούνται ήδη σε εφαρμογές είναι πιθανό να είναι ευάλωτα σε επιθέσεις. Για παράδειγμα, με την τοποθέτηση κάποιων αυτοκόλλητων

σε πινακίδες οδικής κυκλοφορίας από ερευνητές έχει αποδειχθεί πως ένα αυτοκινούμενο όχημα οδηγείται σε λάθος κίνηση. Άλλες έρευνες έχουν δείξει ότι ακόμη και ανεπαίσθητες αλλαγές σε μια εικόνα μπορεί να ξεγελάσουν ένα μοντέλο μηχανικής μάθησης ώστε να ταξινομήσει έναν καλοήθη όγκο ως κακοήθη με 100% βεβαιότητα.

Για την μηχανική μάθηση τα δεδομένα είναι το πιο σημαντικό κομμάτι. Πολλά προσωπικά δεδομένα συλλέγονται και βρίσκονται στο νέφος (cloud) σε μορφή απλού κειμένου ώστε να χρησιμοποιηθούν για τη δημιουργία μοντέλων μηχανικής μάθησης. Το πρόβλημα σε αυτή την περίπτωση δεν εντοπίζεται μόνο στο γεγονός πως τέτοιου είδους προσωπικά δεδομένα βρίσκονται εκτεθειμένα σε εσωτερικούς και εξωτερικούς κακόβουλους χρήστες, αλλά ότι ακόμη και εάν αυτά ανωνυμοποιηθούν ο κίνδυνος εξακολουθεί να υπάρχει. Ταυτόχρονα, το απόρρητο των δεδομένων έχει σταματήσει να είναι απλά μια φιλοσοφία και γίνεται υποχρεωτικό με πληθώρα νόμων και κανονισμών, με αποτέλεσμα η συλλογή και η διατήρηση των προσωπικών δεδομένων να αποτελεί μια δύσκολη διαδικασία. Χαρακτηριστικό παράδειγμα είναι και ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR) που στόχο έχει να δώσει στους καταναλωτές μεγαλύτερο έλεγχο στη συλλογή και χρήση προσωπικών δεδομένων με μεγάλες κυρώσεις προς τις επιχειρήσεις που δεν τον εφαρμόζουν.

Οι αδυναμίες της μηχανικής μάθησης σε συνδυασμό με την πληθώρα δεδομένων που απαιτεί αλλά και το μεγάλο εύρος τομέων που κάνουν χρήση μοντέλων μηχανικής μάθησης οδηγούν πολλούς επιτιθέμενους κατά αυτής. Ένα μεγάλο πλήθος διαφορετικών επιθέσεων εφαρμόζεται κατά της μηχανικής μάθησης με πολλούς διαφορετικούς τρόπους, στόχους και αποτελέσματα, και ενώ η μηχανική μάθηση εξελίσσεται και ως προς το κομμάτι της ασφάλειας το ίδιο συμβαίνει και με τις επιθέσεις.

1.1 Αντικείμενο διπλωματικής

Τα όσα αναφέρθηκαν παραπάνω αποτέλεσαν και το κίνητρο της παρούσας διπλωματικής εργασίας. Αντικείμενο αυτής της διπλωματικής εργασίας είναι η βιβλιογραφική διερεύνηση νέων τύπων επιθέσεων κατά της μηχανικής μάθησης είτε αυτές επιτυγχάνονται μέσω της μεταβολής των δεδομένων εκπαίδευσης είτε μέσω της μεταβολής της εισόδου απόφασης. Κυρίως ερευνήθηκαν δυο τύποι επιθέσεων, οι επιθέσεις διαφυγής και οι επιθέσεις δηλητηρίασης δεδομένων αλλά και οι τεχνικές άμυνας απέναντι σε αυτές. Επίσης, η έρευνα στράφηκε και στα διάφορα ζητήματα ιδιωτικότητας που προκύπτουν από τις επιθέσεις κατά της μηχανικής μάθησης αλλά και στις τεχνικές άμυνας απέναντι σε αυτές.

1.2 Δομή της διπλωματικής

Στο κεφάλαιο 2 αναφέρονται διάφοροι τύποι αλγορίθμων μηχανικής μάθησης και πώς αυτοί λειτουργούν αλλά και που χρησιμοποιούνται συνήθως. Επίσης, δίνεται ο ορισμός της βαθιάς μηχανικής μάθησης, γίνεται αναφορά στο πως αυτή λειτουργεί και στους διάφορους τύπους νευρωνικών δικτύων καθώς και στο που αυτή εφαρμόζεται.

Στο κεφάλαιο 3 γίνεται μια προσπάθεια κατηγοριοποίησης των επιθέσεων κατά των μοντέλων μηχανικής μάθησης με εστίαση κυρίως στις επιθέσεις διαφυγής και στις επιθέσεις

δηλητηρίασης δεδομένων, ενώ δίνονται και διάφορα παραδείγματα επιθέσεων από τον πραγματικό κόσμο. Επίσης, παρουσιάζονται τεχνικές άμυνας έναντι αυτών των ειδών επιθέσεων.

Στο κεφάλαιο 4 αναφέρονται τα σημαντικότερα θέματα απειλών της ιδιωτικότητας που προκύπτουν από τις επιθέσεις κατά της μηχανικής μάθησης και παρουσιάζονται και διάφορες τεχνικές άμυνας για την αποφυγή τέτοιων θεμάτων.

2

Θεωρητικό υπόβαθρο

2.1 Αλγόριθμοι μηχανικής μάθησης

Υπάρχουν διάφορες παραλλαγές στον τρόπο κατηγοριοποίησης των αλγορίθμων μηχανικής μάθησης, αλλά συνήθως μπορούν να χωριστούν σε κατηγορίες βάσει του τρόπου με τον οποίο λαμβάνεται η μάθηση ή του τρόπου με τον οποίο δίνεται ανάδραση στην εκμάθηση στο ανεπτυγμένο σύστημα. Στο παρόν κεφάλαιο αναφέρονται οι βασικοί τύποι αλγορίθμων μηχανικής μάθησης, βάσει του τρόπου με τον οποίο λαμβάνεται η μάθηση.

2.1.1 Επιβλεπόμενη μάθηση (Supervised learning)

Ως επιβλεπόμενη μάθηση, ορίζεται η διαδικασία μάθησης μιας συνάρτησης που αντιστοιχίζει μια είσοδο σε μια έξοδο, βάσει παραδειγμάτων ζευγαριών εισόδου - εξόδου. Κάθε παράδειγμα είναι ένα ζεύγος που αποτελείται από μια είσοδο (συνήθως διάνυσμα) και μια σωστή έξοδο. Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος αναζητά μοτίβα στα δεδομένα, που σχετίζονται με τις επιθυμητές εξόδους, ώστε και σε τυχαίες εισόδους μελλοντικά (μετά το πέρας της εκπαίδευσης) να κάνει σωστές προβλέψεις. Μετά την εκπαίδευση, ο αλγόριθμος θα λάβει νέες εισόδους και θα καθορίσει με ποιες ετικέτες (labels) θα ταξινομηθούν (η ετικέτα βοηθάει να ξεχωρίσουμε κάποια δεδομένα από κάποια άλλα και μπορεί να είναι οτιδήποτε), βάσει των προηγούμενων δεδομένων εκπαίδευσης. Στόχος είναι η πρόβλεψη της σωστής ετικέτας για νέες εισόδους. [1][6]

Η επιβλεπόμενη μηχανική μάθηση, περιλαμβάνει δυο βασικές διαδικασίες: ταξινόμηση (classification) και παλινδρόμηση (regression). [1] [10]

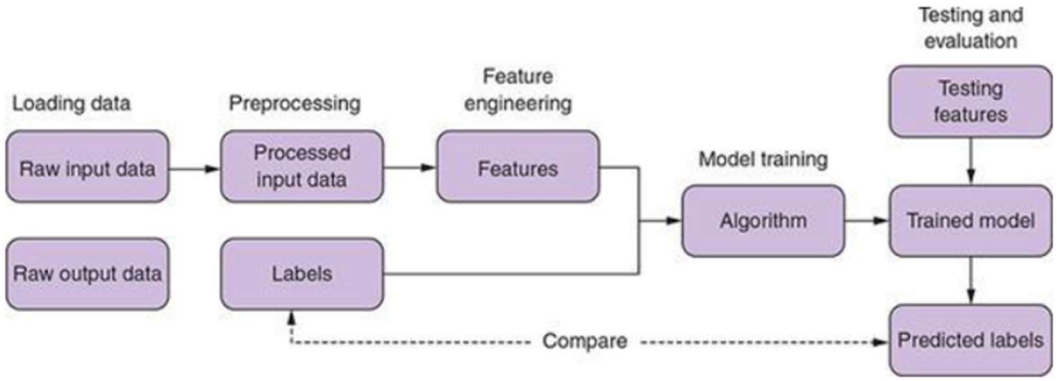
- **Ταξινόμηση (classification):** Η ταξινόμηση χρησιμοποιείται για την πρόβλεψη μιας διακριτής κλάσης ή ετικέτας. Περιλαμβάνει την αντιστοίχιση νέων μεταβλητών εισόδου σε κάποια τάξη (class), στην οποία κατά πάσα πιθανότητα ανήκουν, βάσει ενός μοντέλου ταξινόμησης που δημιουργήθηκε από τα δεδομένα εκπαίδευσης (training data) στα οποία είχε ήδη δοθεί κάποια ετικέτα (labeled data). Αυτά τα δεδομένα χρησιμοποιούνται για την εκπαίδευση ενός ταξινομητή (classifier), έτσι ώστε ο αλγόριθμος να αποδίδει καλά σε δεδομένα που δεν διαθέτουν ετικέτα (που δηλαδή δεν έχουν ακόμη επισημανθεί). Η επανάληψη αυτής της διαδικασίας εκπαίδευσης του ταξινομητή στα ήδη επισημασμένα δεδομένα, είναι γνωστή ως "μάθηση". [2][3][4]

Η ταξινόμηση χωρίζεται σε τρεις κατηγορίες, οι οποίες είναι: δυαδική ταξινόμηση (binary classification), ταξινόμηση πολλαπλών κατηγοριών (multi-class classification) και ταξινόμηση πολλαπλών ετικετών (multi-label classification).

Υπάρχουν διάφοροι αλγόριθμοι ταξινόμησης, που χρησιμοποιούνται για προβλέψεις, όπως νευρωνικά δίκτυα (neural networks), δέντρα αποφάσεων (decision trees), αλγόριθμοι τυχαίου δάσους (random forest algorithms) και άλλοι. [1][4]

- **Παλινδρόμηση (Regression):** Η παλινδρόμηση χρησιμοποιείται για τη πρόβλεψη μιας συνεχόμενης τιμής. Σκοπός είναι η πρόβλεψη της τιμής όσο πιο κοντά στην πραγματική τιμή της εξόδου, όπως το μοντέλο μπορεί, και στη συνέχεια η αξιολόγηση γίνεται με τον υπολογισμό της τιμής του λάθους. Όσο πιο μικρό το λάθος τόσο μεγαλύτερη η ακρίβεια του μοντέλου παλινδρόμησης (regression). [2][3][4]

  Τύποι παλινδρόμησης: Γραμμική παλινδρόμηση (Linear Regression), Πολυωνυμική παλινδρόμηση (Polynomial Regression), Παλινδρόμηση διανυσμάτων υποστήριξης (Support Vector Regression), Παλινδρόμηση δέντρων απόφασης (Decision Tree Regression), Παλινδρόμηση τυχαίων δασών (Random Forest Regression). [1][4]



Εικόνα 1: Η διαδικασία της μηχανικής μάθησης για την επιβλεπόμενη μάθηση [6]

Μελέτες περιπτώσεων χρήσης των αλγορίθμων επιβλεπόμενης μάθησης:

Τα πιο συνηθισμένα πεδία χρήσης για την επιβλεπόμενη μάθηση είναι η πρόβλεψη των τιμών και των τάσεων στις πωλήσεις. Και στις δυο περιπτώσεις, ένας αλγόριθμος χρησιμοποιεί εισερχόμενα δεδομένα για να αξιολογήσει τη δυνατότητα και να υπολογίσει τα πιθανά αποτελέσματα.

Οι περιπτώσεις όπου η επιβλεπόμενη μάθηση χρησιμοποιείται στον επιχειρηματικό κόσμο περιλαμβάνουν τεχνολογίες του κλάδου της διαφήμισης, όπως είναι η ακολουθία προβολής περιεχομένου διαφημίσεων.

2.1.2 Μη επιβλεπόμενη μάθηση (Unsupervised learning)

Η μη επιβλεπόμενη μάθηση (unsupervised learning) είναι το είδος της εκπαίδευσης στο οποίο χρησιμοποιούνται πληροφορίες που δεν είναι ούτε ταξινομημένες ούτε επισημασμένες και έτσι δίνεται η δυνατότητα στον αλγόριθμο να ενεργεί σε αυτές τις πληροφορίες χωρίς καθοδήγηση.

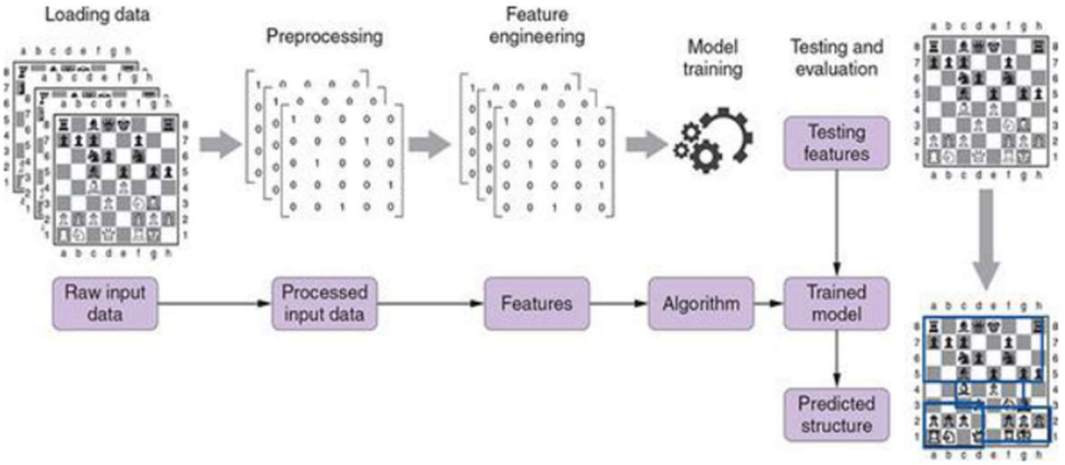
Σκοπός είναι να ομαδοποιούνται ασαφείς πληροφορίες σύμφωνα με ομοιότητες, μοτίβα και διαφορές, χωρίς προηγούμενη εκπαίδευση σε δεδομένα. [1][6]

Σε αντίθεση με την επιβλεπόμενη μάθηση, δεν παρέχεται κάποιου είδους εκπαίδευση στη μηχανή. Συνεπώς, η μηχανή περιορίζεται στο να βρει τη κρυφή δομή σε μη επισημασμένα δεδομένα από μόνη της.

Οι αλγόριθμοι μη επιβλεπόμενης μάθησης εφαρμόζουν τις παρακάτω τεχνικές για να περιγράψουν τα δεδομένα:

- **Ομαδοποίηση** (Clustering): Πραγματοποιείται μια διερεύνηση των δεδομένων, προκειμένου να διαχωριστούν σε ομάδες (clusters), βάσει κάποιων μοτίβων χωρίς προηγούμενη γνώση των χαρακτηριστικών της κάθε ομάδας. Τα χαρακτηριστικά προσδιορίζονται από την ομοιότητα των επιμέρους δεδομένων αλλά και από τις διαφορές από τα υπόλοιπα (μπορεί να χρησιμοποιηθεί για την ανίχνευση ανωμαλιών).
- **Μείωση διάστασης** (Dimensionality reduction): Εξαιτίας της ύπαρξης «θορύβου» στα εισερχόμενα δεδομένα, οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν τη μείωση των διαστάσεων για την εξάλειψη αυτού του θορύβου κατά τον διαχωρισμό των σχετικών πληροφοριών. [5]

Οι πιο ευρέως χρησιμοποιούμενοι αλγόριθμοι είναι: K-μέσων συσταδοποίησης (k-means clustering), t-SNE (t-Distributed Stochastic Neighbor Embedding), PCA (Principal Component Analysis), Κανόνες συσχετίσεων (Association rules) [1][4]



Εικόνα 2: : Η διαδικασία μη επιβλεπόμενης μηχανικής μάθησης για τον εντοπισμό συστάδων πιονιών στο σκάκι [6]

Μελέτες περιπτώσεων χρήσης των αλγορίθμων μη επιβλεπόμενης μάθησης:

Το ψηφιακό marketing και η τεχνολογία της διαφήμισης είναι κάποια από τα πεδία στα οποία η μάθηση χωρίς επίβλεψη χρησιμοποιείται. Επιπλέον, αυτός ο αλγόριθμος χρησιμοποιείται για να διερευνήσει τις πληροφορίες σχετικά με τις προτιμήσεις των πελατών και να προσαρμόσει αναλόγως την υπηρεσία.

Η μη επιβλεπόμενη μάθηση μπορεί να χρησιμοποιηθεί για τον εντοπισμό ομάδων - στόχων βάσει συγκεκριμένων χαρακτηριστικών (μπορεί να είναι δεδομένα συμπεριφοράς, στοιχεία

προσωπικών δεδομένων, συγκεκριμένες ρυθμίσεις λογισμικού) και έτσι μπορεί να χρησιμοποιηθεί για την ανάπτυξη αποτελεσματικότερης στόχευσης του διαφημιστικού περιεχομένου.

2.1.3 Ημί-επιβλεπόμενη μάθηση (Semi-supervised learning)

Οι ημί-επιβλεπόμενοι αλγόριθμοι μάθησης αντιπροσωπεύουν ένα μεσαίο πεδίο μεταξύ των επιβλεπόμενων και των μη επιβλεπόμενων αλγορίθμων. Στην ουσία, το συγκεκριμένο μοντέλο περιλαμβάνει στοιχεία και των δυο. [1]

Παρακάτω περιγράφεται ο τρόπος με τον οποίο ο ημί-επιβλεπόμενος αλγόριθμος λειτουργεί:

1) Χρησιμοποιεί ένα περιορισμένο σύνολο επισημασμένων δεδομένων για να διαμορφώσει τις απαιτήσεις της λειτουργίας
2) Ο παραπάνω περιορισμός έχει ως αποτέλεσμα ένα μερικώς εκπαιδευμένο μοντέλο, το οποίο αναλαμβάνει αργότερα να επισημάνει τα μη επισημασμένα δεδομένα. Εξαιτίας του περιορισμένου δείγματος δεδομένων, τα αποτελέσματα θεωρούνται ψευδο-επισημασμένα.
3) Τέλος, συνδυάζονται τα επισημασμένα και τα ψευδο-επισημασμένα δεδομένα, τα οποία δημιουργούν ένα ξεχωριστό αλγόριθμο, ο οποίος συνδυάζει στοιχεία της επιβλεπόμενης και της μη επιβλεπόμενης μάθησης.

Μελέτες περιπτώσεων χρήσης των αλγορίθμων ημί-επιβλεπόμενης μάθησης:

Πλήθος διάφορων τύπων βιομηχανιών διαχειρίζεται την ανάλυση εικόνας και ομιλίας με τη βοήθεια της ημί-επιβλεπόμενης μάθησης.

Στην περίπτωση της ανάλυσης εικόνας και ομιλίας, ένας αλγόριθμος αποδίδει ετικέτες, έτσι ώστε να παρέχει ένα βιώσιμο μοντέλο ανάλυσης εικόνας ή ομιλίας. Για παράδειγμα, μπορεί να είναι μια μαγνητική ή αξονική τομογραφία. Με μια μικρή σειρά υποδειγματικών σαρώσεων, είναι δυνατόν να παρέχεται ένα συνεκτικό μοντέλο, ικανό να αναγνωρίζει ανωμαλίες στις εικόνες.

2.1.4 Ενισχυμένη μάθηση (Reinforcement learning)

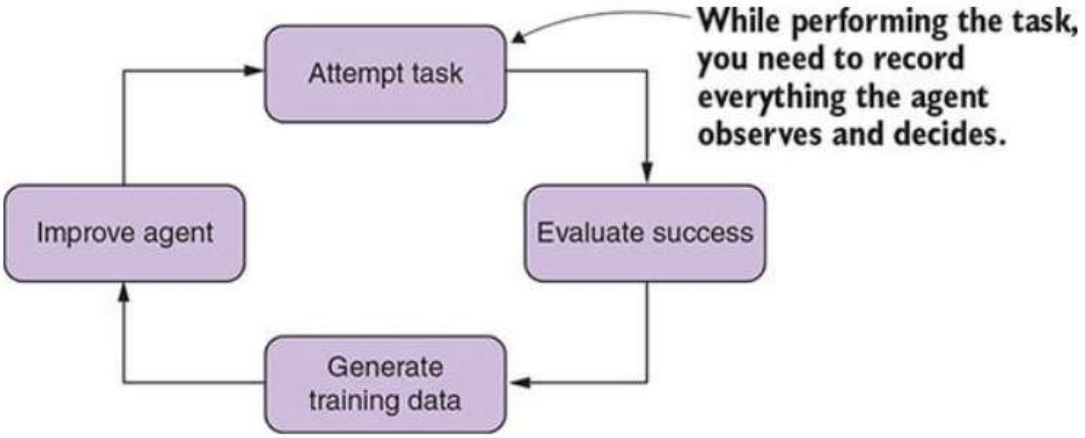
Η ενισχυμένη μάθηση αντιπροσωπεύει αυτό που είναι κοινώς κατανοητό ως μηχανική μάθηση τεχνητής νοημοσύνης.

Στην ουσία, η ενισχυμένη μάθηση αφορά στην ανάπτυξη ενός αυτοσυντηρούμενου συστήματος το οποίο, σε συνεχείς αλληλουχίες προσπαθεί και αποτυγχάνει, βελτιώνεται με βάση το συνδυασμό των δεδομένων και των αλληλεπιδράσεων με τα εισερχόμενα δεδομένα.

Η ενισχυμένη μάθηση χρησιμοποιεί την τεχνική που ονομάζεται εξερεύνηση / εκμετάλλευση. Η λειτουργία είναι απλή, καθώς η δράση λαμβάνει χώρα, παρατηρούνται συνέπειες και η επόμενη ενέργεια εξετάζει τα αποτελέσματα της πρώτης δράσης. [6]

Βασικό στοιχείο των αλγορίθμων της ενισχυμένης μάθησης είναι τα σήματα ανταμοιβής που εμφανίζονται κατά την εκτέλεση συγκεκριμένων εργασιών. Κατά κάποιο τρόπο, τα σήματα ανταμοιβής χρησιμεύουν ως εργαλείο πλοήγησης για τους αλγόριθμους ενίσχυσης. Παρέχουν μια κατανόηση της ορθής και λανθασμένης πορείας δράσης. [1]

Οι πιο συνήθεις αλγόριθμοι ενίσχυσης είναι: Q-Learning, Temporal Difference (TD), Monte-Carlo Tree Search (MCTS), Asynchronous Actor-Critic Agents (A3C)



Εικόνα 3: Στην ενισχυμένη μάθηση οι πράκτορες (agents) μαθαίνουν να αλληλοεπιδρούν με το περιβάλλον τους με δοκιμές και λάθη. Επανειλημμένως επιχειρούν τις διεργασίες τους προκειμένου να πάρουν μια μορφή ανταπόδοσης και μέσω αυτού να μάθουν. Με κάθε κύκλο πραγματοποιούν βελτιώσεις. [6]

Μελέτες περιπτώσεων χρήσης των αλγορίθμων ενισχυμένης μάθησης:

Η ενισχυμένη μάθηση είναι κατάλληλη για περιπτώσεις περιορισμένων ή μη ολοκληρωμένων διαθέσιμων πληροφοριών. Σε αυτή την περίπτωση, ένας αλγόριθμος μπορεί να διαμορφώσει τις λειτουργικές του διαδικασίες βάσει αλληλεπιδράσεων με δεδομένα και σχετικές διαδικασίες.

Τα σύγχρονα βιντεοπαιχνίδια χρησιμοποιούν σε πολύ μεγάλο βαθμό το συγκεκριμένο είδος μοντέλου μηχανικής μάθησης. Η ενισχυμένη μάθηση παρέχει ευελιξία στις αντιδράσεις του συστήματος τεχνητής νοημοσύνης ως προς τη δράση του παίκτη, παρέχοντας έτσι βιώσιμες προκλήσεις.

Τα αυτοκινούμενα οχήματα βασίζονται επίσης σε αλγορίθμους ενισχυμένης μάθησης. Για παράδειγμα, αν ένα αυτοκινούμενο όχημα ανιχνεύσει τον δρόμο προς τα αριστερά μπορεί να ενεργοποιήσει το σενάριο «στρίψε αριστερά» και ούτω καθεξής.

Από την άλλη πλευρά, οι λειτουργίες μάρκετινγκ και διαφήμισης χρησιμοποιούν επίσης την ενισχυμένη μάθηση.

Επίσης, η ενισχυμένη μάθηση χρησιμοποιείται για να ενισχύσει και να προσαρμόσει την επεξεργασία φυσικής γλώσσας (NLP) και τη δημιουργία διαλόγου για bot συνομιλίας (chatbots).

2.1.5 Εξελικτική μάθηση (Evolutionary learning)

Η εξελικτική μάθηση βασίζεται στο πως οι οργανισμοί εξελίσσονται με το πέρασμα των χρόνων. Οι διαδικασίες της εξελικτικής μάθησης έχουν ως βάση βιολογικές διαδικασίες όπως η μετάλλαξη, η αναπροσαρμογή, η επιλογή και η αναπαραγωγή. [86] Μια από τις βασικές μεθόδους υλοποίησης της εξελικτικής μάθησης είναι οι γενετικοί αλγόριθμοι, η λειτουργία των οποίων θυμίζει τους βιολογικούς οργανισμούς, καθώς αυτοί εξελίσσονται και προσαρμόζονται στο περιβάλλον. [87] Για τη δημιουργία και την μοντελοποίηση των γενετικών αλγορίθμων ακολουθούνται τα παρακάτω βήματα – στάδια:

1) Κατά το πρώτο βήμα δημιουργείται ένας τυχαίος αρχικός πληθυσμός.

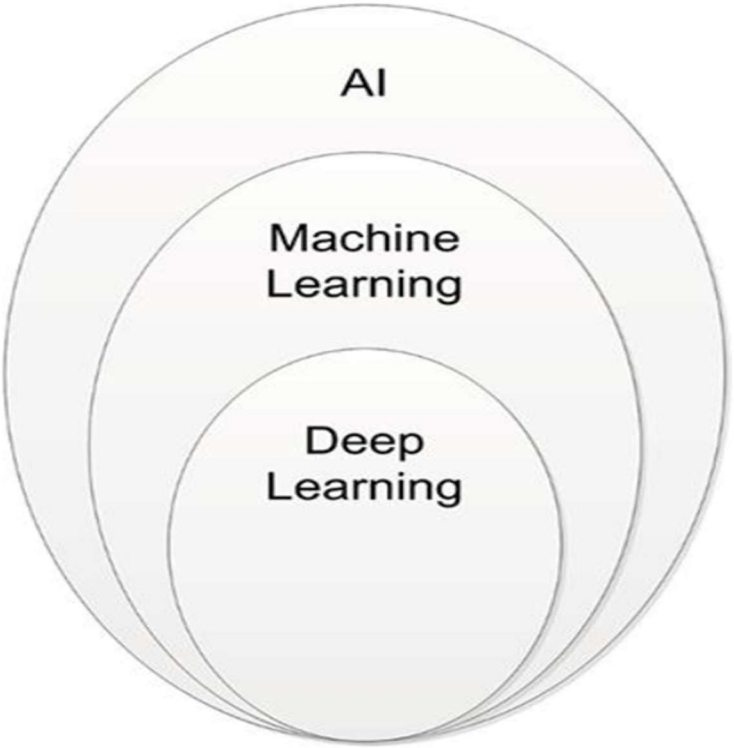
2) Σε αυτό το βήμα γίνεται προσπάθεια αξιολόγησης της προσαρμοστικότητας στο περιβάλλον του κάθε στοιχείου που δημιουργήθηκε στο προηγούμενο βήμα.
3) Σε αυτή τη φάση ακολουθούνται τα βήματα της αναπαραγωγής. Πρώτα επιλέγονται τα στοιχεία με τη καλύτερη προσαρμοστικότητα (γονείς), αμέσως μετά πραγματοποιείται η δημιουργία του ζευγαριού ώστε να γίνει η παραγωγή απογόνων. Και τέλος αξιολογούνται ως προς τη προσαρμοστικότητα στο περιβάλλον τα νέα στοιχεία που δημιουργήθηκαν.
4) Πραγματοποιείται η αντικατάσταση των λιγότερο προσαρμοστικών στοιχείων με τα περισσότερο προσαρμοστικά.
5) Επανάληψη των βημάτων από το στάδιο 2.

Και η επανάληψη αυτή γίνεται όσες φορές είναι απαραίτητο ώστε να βρεθεί ο πιο κατάλληλος πληθυσμός. [88]

Εικόνα 4: Ταξινόμηση αλγορίθμων μηχανικής μάθησης [95]

2.2 Βαθιά μάθηση (Deep Learning)

Η βαθιά μάθηση αποτελεί τμήμα της τεχνητής νοημοσύνης (Artificial Intelligence) και πιο συγκεκριμένα τμήμα της μηχανικής μάθησης, όπως φαίνεται και στην εικόνα παρακάτω.



Εικόνα 5: Τεχνητή νοημοσύνη, μηχανική μάθηση, βαθιά μάθηση

Όπως αναφέρθηκε και προηγουμένως, η βαθιά μάθηση αποτελεί πεδίο της μηχανικής μάθησης. Οι αλγόριθμοι μαθαίνουν και στις δύο αυτές περιπτώσεις από μεγάλα σετ δεδομένων (σε μερικές περιπτώσεις μπορεί και από μικρά). Παρακάτω αναφέρονται κάποιες διαφορές ανάμεσα στη μηχανική και στη βαθιά μάθηση, προκειμένου να κατανοήσουμε κάποιες πτυχές της δεύτερης:

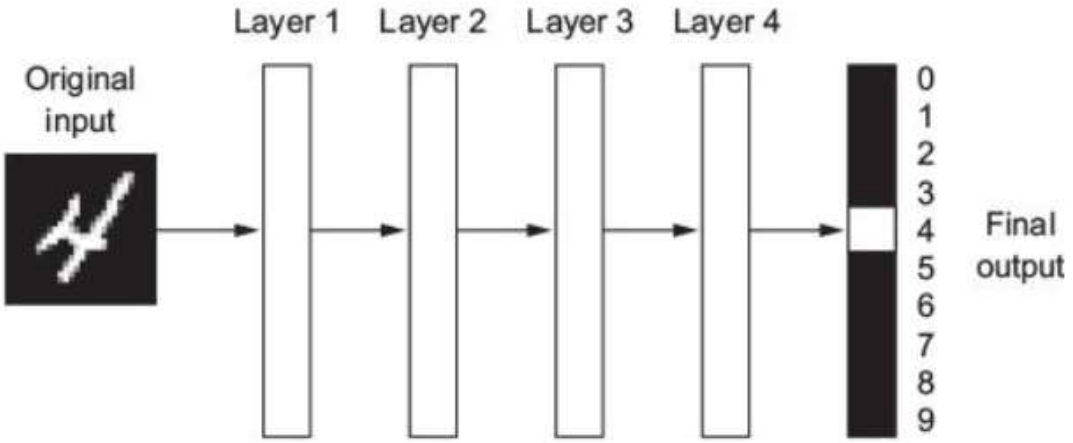
- Η μηχανική μάθηση θα μπορούσε να οριστεί και ως ένα σύνολο διάφορων τεχνικών, οι οποίες είναι ικανές να κάνουν έναν υπολογιστή να μάθει από δεδομένα και κατόπιν να χρησιμοποιήσει ότι έχει μάθει για να προσφέρει απαντήσεις, με τη μορφή προβλέψεων. Μερικές από αυτές τις τεχνικές είναι η στατιστική ανάλυση, η αναζήτηση αναλογιών στα δεδομένα, η χρήση λογικής και άλλες. Αντίθετα με τη μηχανική μάθηση, που χρησιμοποιεί πλήθος τεχνικών, μόνο μια τεχνική χρησιμοποιείται από τη βαθιά μάθηση. Η συγκεκριμένη τεχνική μιμείται τη λειτουργία του ανθρώπινου εγκεφάλου. Για την επεξεργασία των δεδομένων γίνεται χρήση μονάδων υπολογισμού (computing units), που ονομάζονται νευρώνες (neurons), και είναι τοποθετημένες σε διατεταγμένα τμήματα, που λέγονται στρώματα (layers). Η τεχνική αυτή είναι το νευρωνικό δίκτυο (neural network).
- Οι λύσεις μηχανικής μάθησης, προσφέρουν διάφορες προσαρμογές (knobs), προκειμένου να υπάρξει βελτιστοποίηση των αλγορίθμων, οι οποίες ονομάζονται υπερπαράμετροι (hyperparameters). Αντίστοιχα, και στη βαθιά μάθηση χρησιμοποιούνται υπερπαράμετροι, αλλά ταυτόχρονα υπάρχει η δυνατότητα πολλαπλών επιπέδων διαμόρφωσης από τον χρήστη (ο χρήστης καθορίζει τον αριθμό και τον τύπο). Ανάλογα με το νευρωνικό δίκτυο που προκύπτει, το πλήθος των στρωμάτων μπορεί να είναι πολύ μεγάλο και κατ' αυτόν τον τρόπο να προκύπτουν μεμονωμένα νευρωνικά δίκτυα ικανά για εξειδικευμένη μάθηση: κάποια μπορεί να μάθουν να αναγνωρίζουν εικόνες και άλλα να ανιχνεύουν και να αναλύουν φωνητικές εντολές.

- Οι λύσεις μηχανικής μάθησης απαιτούν την παρέμβαση του ανθρώπου προκειμένου να δώσουν τα επιθυμητά αποτελέσματα. Από την άλλη πλευρά στη βαθιά μάθηση δεν απαιτείται σε τέτοιο βαθμό η ανθρώπινη παρέμβαση. [8]

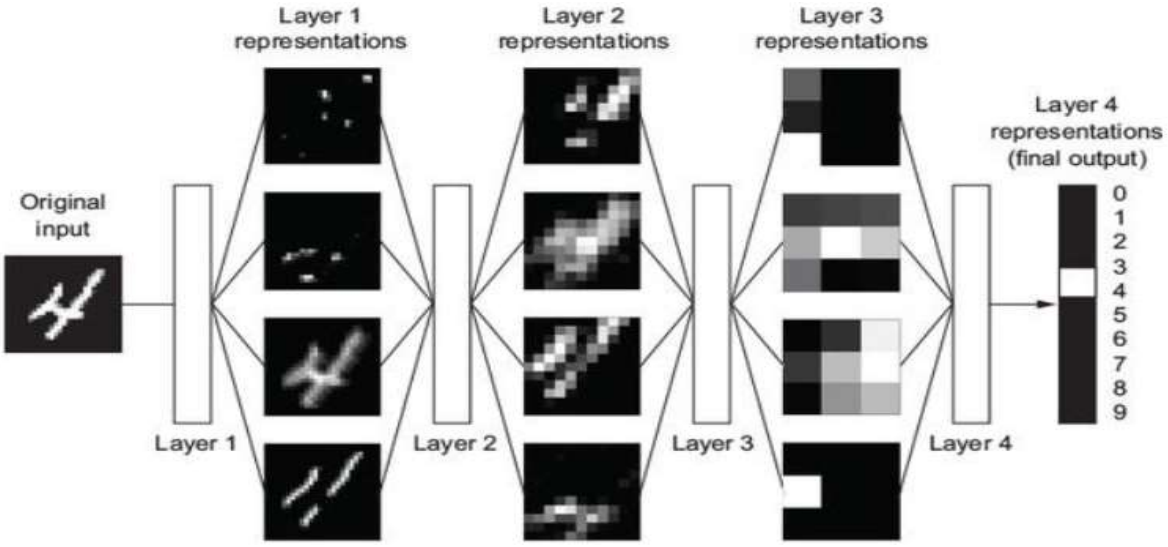
2.2.1 Ορίζοντας τη βαθιά μάθηση

Η βαθιά μάθηση αποτελεί μια νέα προσέγγιση στην εκμάθηση από δεδομένα. Χρησιμοποιεί μια συγκεκριμένη οικογένεια μοντέλων: ακολουθίες απλών συναρτήσεων συνδεδεμένες μεταξύ τους. Αυτές οι αλυσίδες συναρτήσεων αποτελούν τα νευρωνικά δίκτυα. Αυτές οι ακολουθίες συναρτήσεων μπορούν να αναλύσουν μια πολύπλοκη ιδέα σε μια ιεραρχία απλούστερων. Κάθε στρώμα οργανώνει το προηγούμενο στρώμα σε πιο προηγμένες και αφηρημένες έννοιες. [6]

Στις παρακάτω εικόνες παρουσιάζεται πώς ένα δίκτυο με πολλά στρώματα μετασχηματίζει την εικόνα ενός ψηφίου προκειμένου να αναγνωρίσει ποιο είναι.



Εικόνα 6: Βαθύ νευρωνικό δίκτυο για ταξινόμηση ψηφίου [7]

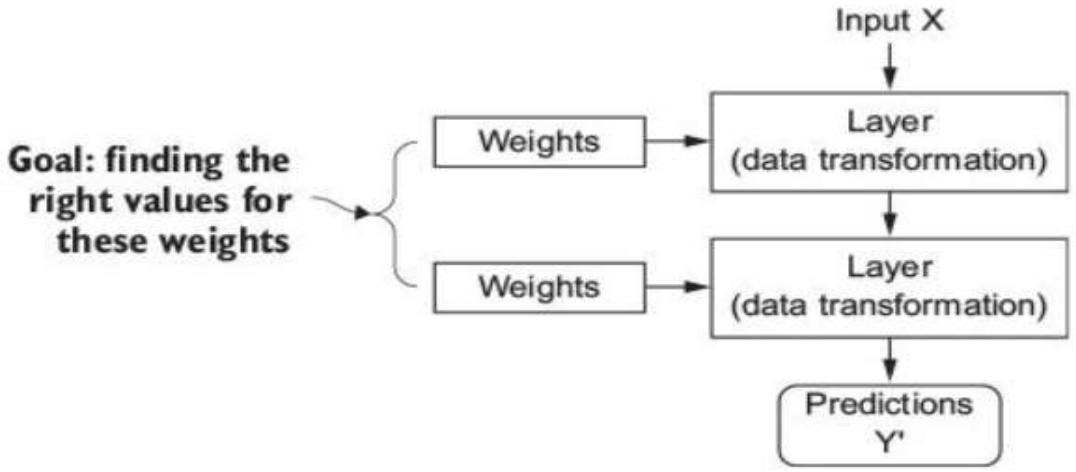


Εικόνα 7: Τα στρώματα αναπαραστάσεων ενός μοντέλου βαθιάς μάθησης για την ταξινόμηση ψηφίου [7]

2.2.2 Πώς λειτουργεί η βαθιά μάθηση

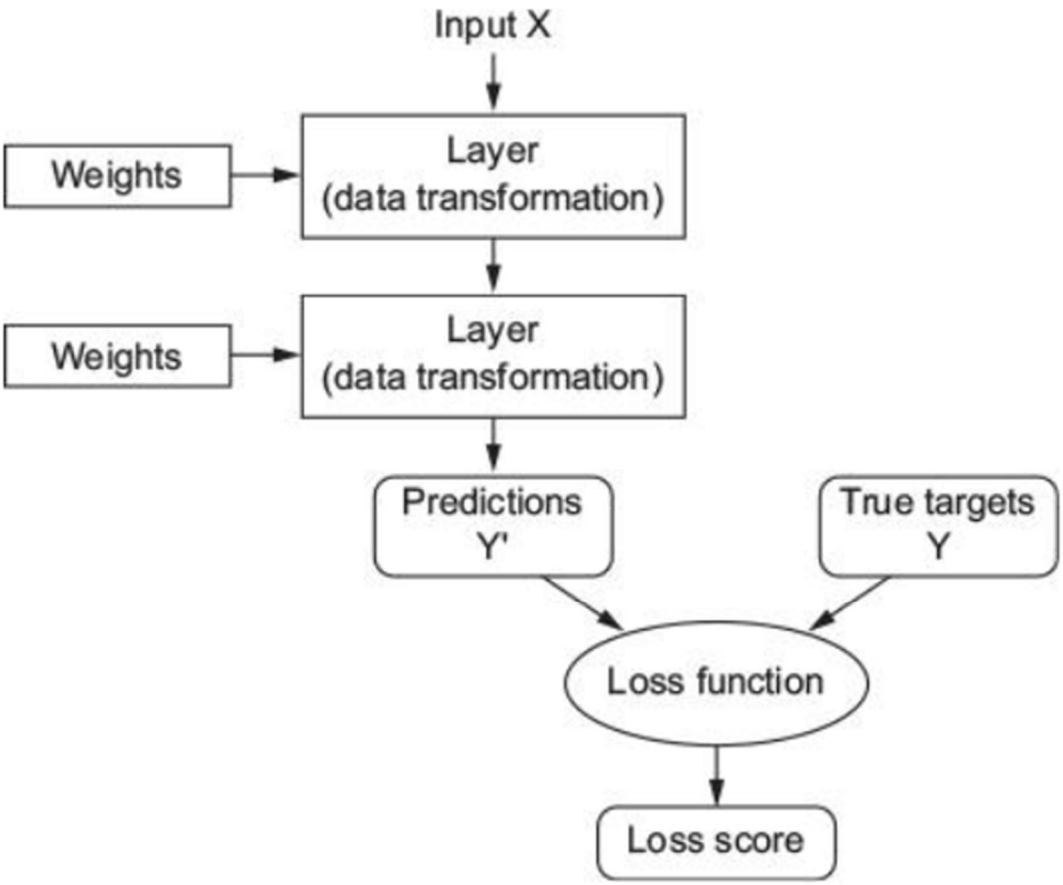
Τα βαθιά νευρωνικά δίκτυα (deep neural networks) πραγματοποιούν τη χαρτογράφηση εισόδου – εξόδου (στόχου) μέσω μιας βαθιάς αλληλουχίας απλών μετασχηματισμών δεδομένων (layers) και αυτοί οι μετασχηματισμοί δεδομένων μαθαίνονται μέσω της έκθεσης σε παραδείγματα.

Πιο συγκεκριμένα, οι προδιαγραφές του τι κάνει ένα στρώμα στα δεδομένα εισόδου του αποθηκεύονται στα βάρη του (layer's weights), τα οποία αποτελούν μια δέσμη αριθμών. Από τεχνικής πλευράς, ο μετασχηματισμός που εφαρμόζεται από ένα στρώμα παραμετροποιείται από τα βάρη του. Η μάθηση σημαίνει την εύρεση των τιμών για τα βάρη του συνόλου των στρωμάτων σε ένα δίκτυο, έτσι ώστε αυτό να αντιστοιχεί σωστά τις εισόδους στις εξόδους – στόχους. Αποτελεί ιδιαιτέρως δύσκολο έργο, καθώς η τροποποίηση της τιμής μιας παραμέτρου επηρεάζει τη συμπεριφορά των άλλων. [7]



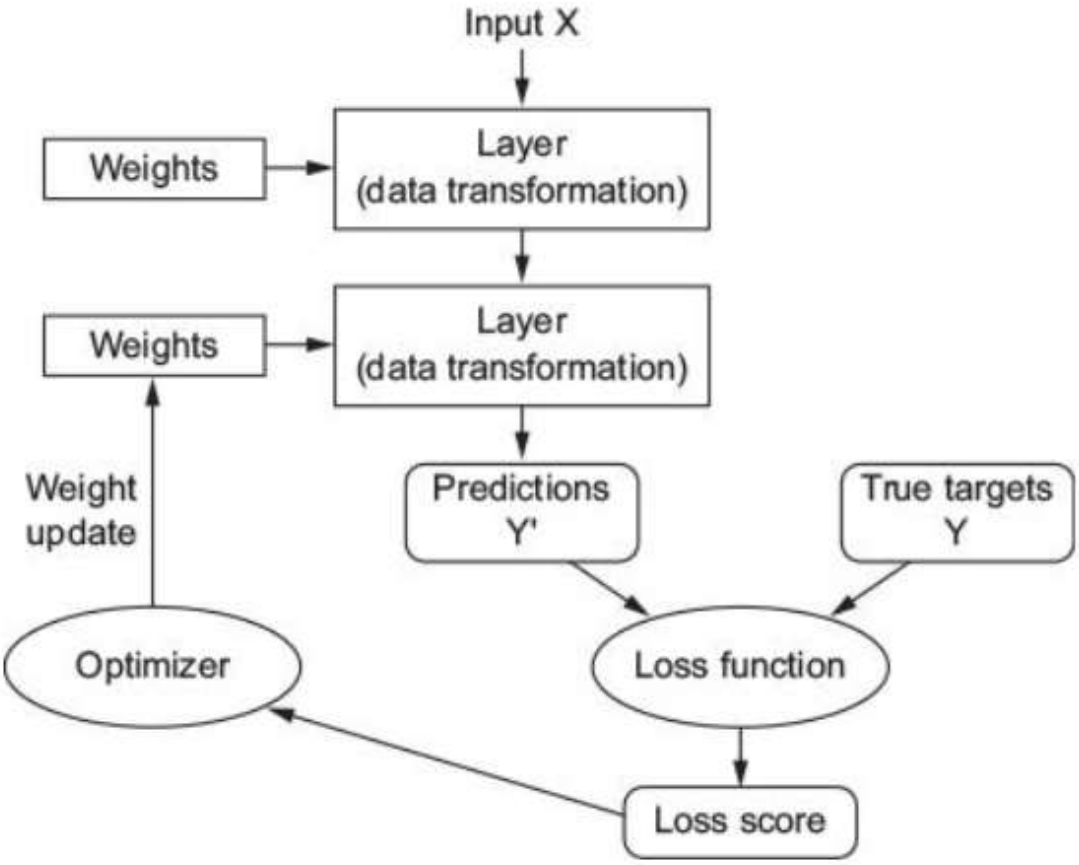
Εικόνα 8: Ένα νευρωνικό δίκτυο παραμετροποείται από τα βάρη του [7]

Προκειμένου ένα νευρωνικό δίκτυο να ελέγξει την έξοδο πρέπει να είναι σε θέση να μετρά την απόκλιση της τιμής κάθε εξόδου από την αναμενόμενη. Η συγκεκριμένη διεργασία επιτελείται από τη συνάρτηση απώλειας (loss function) του δικτύου. Η συγκεκριμένη συνάρτηση πραγματοποιεί τον υπολογισμό της απόκλισης μεταξύ των προβλέψεων που κάνει το δίκτυο, και των πραγματικών στόχων (true targets), καταγράφοντας την απόδοση του δικτύου. [7]



Εικόνα 9: Μια συνάρτηση απώλειας υπολογίζει την ποιότητα της εξόδου του δικτύου [7]

Το πόσο πετυχημένα λειτούργησε το δίκτυο ή όχι, όπως αυτό προέκυψε από την τιμή της συνάρτησης απώλειας, χρησιμοποιείται σαν ανατροφοδότηση (feedback) ώστε οι τιμές των βαρών να αλλάξουν για να μειωθεί ο βαθμός απώλειας (loss score). Αυτή η τροποποίηση των βαρών εκτελείται από τον βελτιστοποιητή (optimizer), ο οποίος επί της ουσίας εκτελεί τον αλγόριθμο οπισθοδιάδοσης (back propagation). [7]

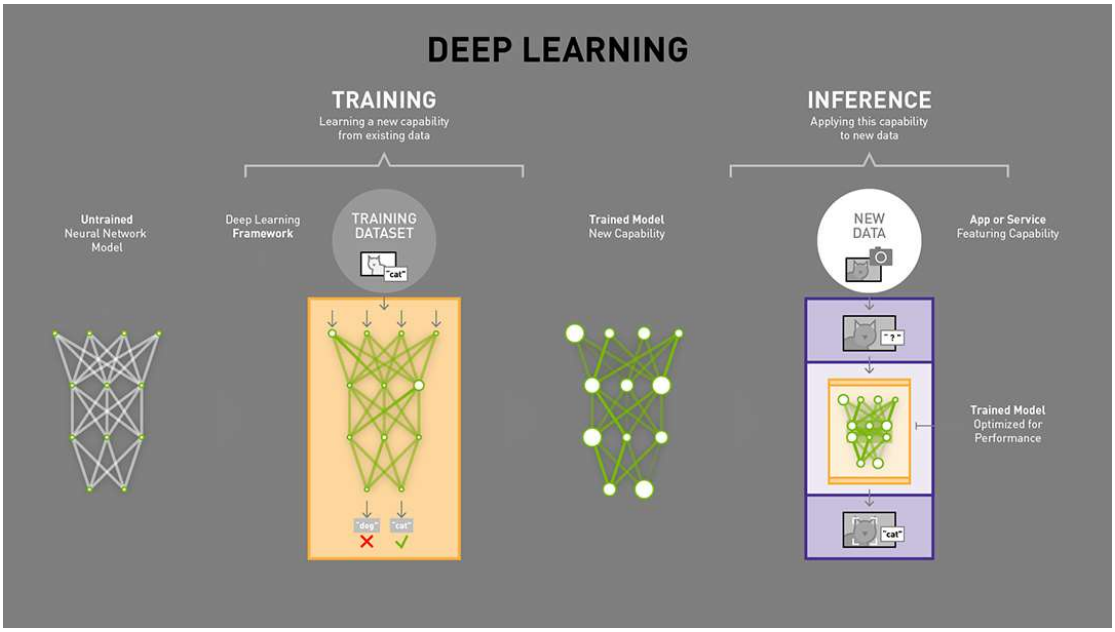


Εικόνα 10: Ο βαθμός απώλειας χρησιμοποιείται σαν ανατροφοδότηση για την εκ νέου παραμετροποίηση των βαρών [7]

Καθώς αρχικά στα βάρη έχουν δοθεί τυχαίες τιμές το δίκτυο υλοποιεί μια σειρά τυχαίων μετασχηματισμών και ο βαθμός απώλειας είναι πολύ υψηλός. Με κάθε παράδειγμα το δίκτυο όλο και προσαρμόζει καλύτερα τα βάρη και μειώνει τον βαθμό απώλειας μέχρις ότου να προκύψει ένα δίκτυο με ελάχιστη απώλεια, ένα εκπαιδευμένο δίκτυο.

2.2.3 Ροή εργασίας της βαθιάς μάθησης

Η διαδικασία της βαθιάς μάθησης περιλαμβάνει δυο φάσεις: την εκπαίδευση (training) και τα συμπεράσματα (inference). Τα βαθιά νευρωνικά δίκτυα μαθαίνουν νέες δεξιότητες κατά τη φάση της εκπαίδευσης (training phase) από τα υπάρχοντα δεδομένα, και αυτές οι δεξιότητες εφαρμόζονται σε άγνωστα δεδομένα στη φάση των συμπερασμάτων (inference phase). [9]



Εικόνα 11: Βαθιά μάθηση: Εκπαίδευση και συμπεράσματα [11]

Εκπαίδευση:

Είναι η φάση κατά την οποία το δίκτυο προσπαθεί να μάθει από τα δεδομένα. Κατά την εκπαίδευση, σε κάθε επίπεδο δεδομένων δίνονται τυχαία βάρη και ο ταξινομητής εκτελεί ένα πέρασμα προς τα εμπρός στα δεδομένα, προβλέποντας τις ετικέτες κλάσης και τις βαθμολογίες χρησιμοποιώντας αυτά τα βάρη. Στη συνέχεια, οι βαθμολογίες της κατηγορίας συγκρίνονται με τις πραγματικές ετικέτες και υπολογίζεται ένα σφάλμα μέσω μια συνάρτησης απώλειας. Αυτό το σφάλμα στη συνέχεια αναπαράγεται μέσω του δικτύου και τα βάρη ενημερώνονται αναλόγως μέσω κάποιου αλγορίθμου.

Συμπεράσματα:

Είναι το στάδιο στο οποίο χρησιμοποιείται ένα εκπαιδευμένο μοντέλο για την εξαγωγή/πρόβλεψη των δειγμάτων δοκιμής και περιλαμβάνει ένα παρόμοιο πέρασμα προς τα εμπρός, όπως στην εκπαίδευση, για την πρόβλεψη τιμών. Σε αντίθεση με την εκπαίδευση, δεν περιλαμβάνει ένα πέρασμα και προς τα πίσω για να υπολογίσει το σφάλμα και να ενημερώσει τα βάρη. Είναι συνήθως μια φάση παραγωγής, όπου το μοντέλο αναπτύσσεται για να προβλέψει δεδομένα του πραγματικού κόσμου.

2.2.4 Τύποι νευρωνικών δικτύων που χρησιμοποιούνται από τη τεχνητή νοημοσύνη

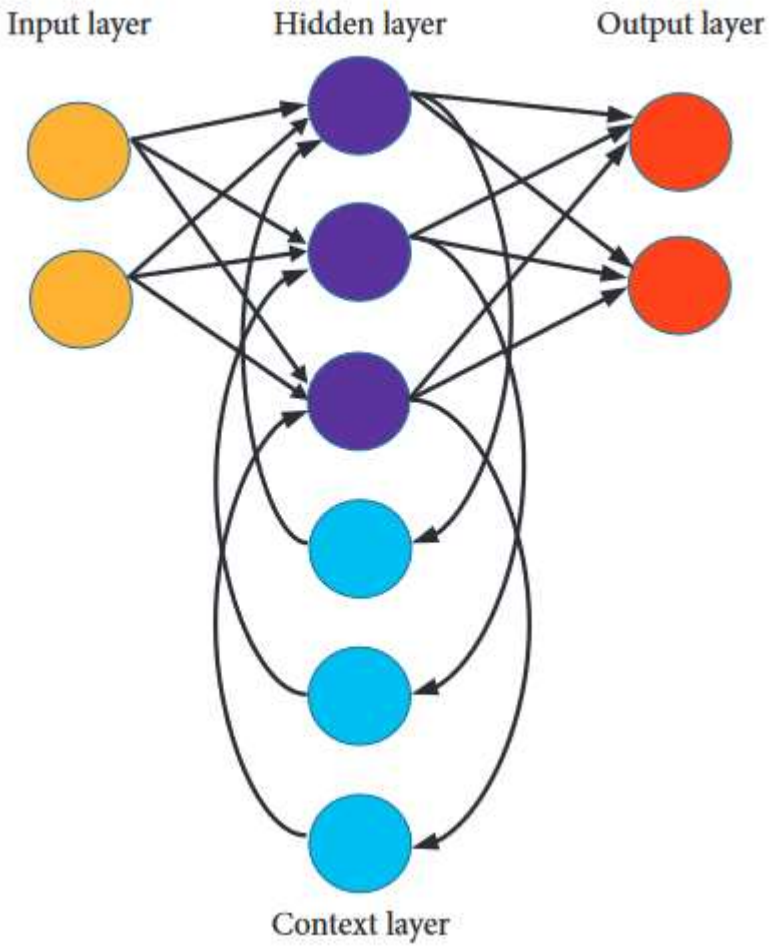
Λαμβάνοντας υπόψιν πως η έρευνα της τεχνητής νοημοσύνης αποσκοπεί στη δημιουργία λειτουργικότητας σε μηχανές που θυμίζει αυτή του ανθρώπινου εγκεφάλου, είναι αυτονόητο πως οι ερευνητές εμπνέονται από τη δομή του ανθρώπινου εγκεφάλου κατά τη δημιουργία μοντέλων τεχνητής νοημοσύνης. Έτσι, η δημιουργία τεχνητών νευρωνικών δικτύων αποτελεί μια προσπάθεια αναπαραγωγής των νευρωνικών δικτύων στο του εγκεφάλου. Συστήματα τεχνητής νοημοσύνης καταφέρνουν να πραγματοποιήσουν ανθρώπινες ενέργειες όπως η κατανόηση της φυσικής

γλώσσας, άλλα και να ξεπεράσουν ειδικούς σε ανθρώπινες ενέργειες που απαιτούν ανάλυση και αναγνώριση προτύπων όπως είναι ο προσδιορισμός του περιεχομένου μιας συγκεκριμένης εικόνας.

Όπως ο ανθρώπινος εγκέφαλος έχει διαφορετικά μέρη που επιτρέπουν διαφορετικές λειτουργίες, έτσι διαφορετικά είδη νευρωνικών δικτύων αναπτύσσονται για την επίλυση διαφορετικών ειδών προβλημάτων. Υπάρχουν πολλοί τύποι νευρωνικών δικτύων που αναπτύσσονται και χρησιμοποιούνται από ερευνητές αλλά μερικοί από αυτούς έχουν βρει μεγαλύτερη δυνατότητα εφαρμογής και ως εκ τούτου είναι περισσότερο δημοφιλείς. Αυτοί αναλύονται παρακάτω.

Ο πρώτος τύπος νευρωνικών δικτύων, αλλά και ο πιο δημοφιλής σήμερα, είναι τα τροφοδοτικά (feedforward) νευρωνικά δίκτυα (FNN). Η ονομασία τους προέρχεται από το γεγονός πως η πληροφορίας ρέει προς μια μόνο κατεύθυνση και δεν υπάρχουν βρόχοι. Τα συγκεκριμένα νευρωνικά δίκτυα μπορούν να ταξινομηθούν ανάλογα με το πλήθος των ενδιάμεσων κρυμμένων στρωμάτων, σε αυτά του ενός στρώματος, τα οποία περιλαμβάνουν μόνο το στρώμα εισόδου και το στρώμα εξόδου (single-layered networks) ή σε πολλών στρωμάτων, τα οποία περιλαμβάνουν πέραν των στρωμάτων εισόδου και εξόδου και άλλα κρυμμένα στρώματα (multilayered networks). Σε αυτού του είδους τα δίκτυα οι συνδέσεις γίνονται μόνο από μονάδες του προηγούμενου προς τις μονάδες του στρώματος που ακολουθεί και όχι ανάμεσα σε μονάδες του ίδιου στρώματος. Τα επεξεργασμένα δεδομένα δεν επιστρέφουν σε προηγούμενο στρώμα. Όσο πιο περίπλοκη είναι η διεργασία που πρέπει να εκτελεστεί από το δίκτυο τόσο πιο πολλά τα επίπεδα. [8] Πιο συχνά χρησιμοποιούνται σε συστήματα αναγνώρισης αντικειμένων και αναγνώρισης ομιλίας.

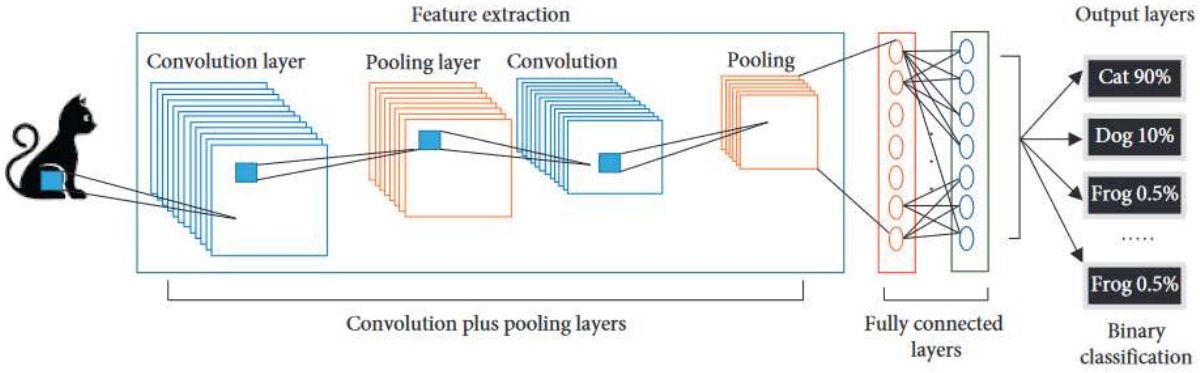
Ένας ακόμη τύπος νευρωνικών δικτύων είναι τα επαναλαμβανόμενα (recurrent) νευρωνικά δίκτυα (RNN), τα οποία όπως και το όνομα τους υποδηλώνει περιλαμβάνουν την επανάληψη λειτουργιών με τη μορφή βρόχων. Σαφώς αποτελούν πιο περίπλοκα δίκτυα από τα προηγούμενα και ταυτόχρονα εξαιτίας αυτής της περίπλοκης λειτουργικότητας τους μπορούν να εκτελέσουν πιο περίπλοκες διεργασίες από την απλή αναγνώριση φωνής. Ενώ στα τροφοδοτικά νευρωνικά δίκτυα οι συνδέσεις οδηγούν από έναν νευρώνα μόνο σε νευρώνες επόμενου στρώματος, στα επαναλαμβανόμενα νευρωνικά δίκτυα υπάρχουν συνδέσεις ανατροφοδότησης που επιτρέπουν την ύπαρξη εσωτερικών καταστάσεων (internal states). Αυτό σημαίνει την ύπαρξη μνήμης που μπορεί να κρατά πληροφορίες για προηγούμενες εισόδους. [65] Βέβαια αυτό το στοιχείο φέρνει και διάφορους περιορισμούς όσον αφορά την εκπαίδευση και τη λειτουργικότητα των συμβατικών RNN δικτύων, καθώς η μνήμη τους είναι βραχυπρόθεσμη. Για να ξεπεραστεί το συγκεκριμένο πρόβλημα μνήμης χρησιμοποιείται μια νεότερη μορφή αυτών των δικτύων, τα οποία ονομάζονται μακριάς βραχυπρόθεσμης μνήμης δίκτυα (long short term memory networks). Η συγκεκριμένη μορφή αυτών των δικτύων επιτρέπει μέσω της επέκτασης της μνήμης την πραγματοποίηση διεργασιών που απαιτούν μεγαλύτερη μνήμη, και τα συμβατικά RNN δίκτυα δε θα μπορούσαν να εκτελέσουν. [66] Συνήθως τα RNN δίκτυα χρησιμοποιούνται για προβλήματα επεξεργασίας φυσικών γλωσσών όπως ομιλία και αναγνώριση κειμένου, πρόβλεψη κειμένου και δημιουργία φυσικής γλώσσας.



Εικόνα 12: Δομή του RNN [98]

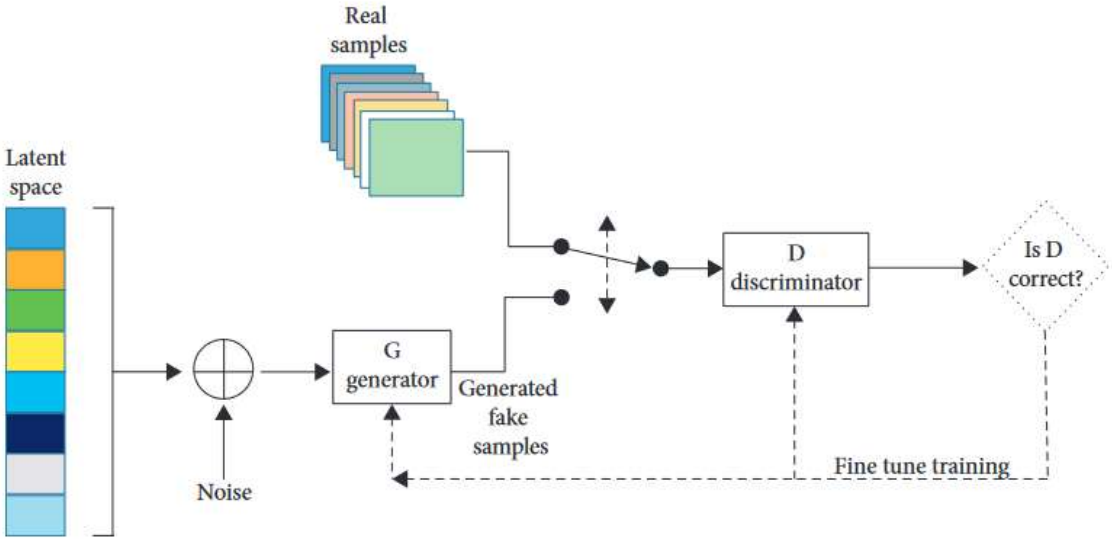
Τα συνελικτικά (convolutional) νευρωνικά δίκτυα (CNN) αποτελούν και αυτά έναν τύπο νευρωνικών δικτύων. Έχουν συνδεθεί κυρίως με εφαρμογές ηλεκτρονικής όρασης και αυτό συμβαίνει εξαιτίας της αρχιτεκτονικής τους που είναι η κατάλληλη για την εκτέλεση σύνθετων οπτικών αναλύσεων. Μια πρώτη έρευνα πάνω στον συγκεκριμένο τύπο δικτύων εστίασε στην χρήση τους για την αναγνώριση χειρόγραφων χαρακτήρων. [67] Τα CNN δίκτυα κατασκευάζονται από πολλά στρώματα και σκοπός των συνδέσεων είναι η εκμάθηση της ιεραρχικής αναπαράστασης χαρακτηριστικών. Η αρχιτεκτονική ενός CNN δικτύου ορίζεται από μια τρισδιάστατη διάταξη νευρώνων αντί της τυπικής δισδιάστατης διάταξης. Τα CNN δίκτυα συνήθως αποτελούνται από τρεις τύπους στρωμάτων, τα στρώματα συνέλιξης (convolution), τα στρώματα ομαδοποίησης (pooling) και τα πλήρως συνδεδεμένα στρώματα (fully connected layers). Τα στρώματα της συνέλιξης και της ομαδοποίησης αναλαμβάνουν την εξαγωγή χαρακτηριστικών ενώ το πλήρως συνδεδεμένο στρώμα αναλαμβάνει την χαρτογράφηση των εξαγόμενων χαρακτηριστικών στη

τελική έξοδο. Τα CNN δίκτυα χρησιμοποιούνται κυρίως σε εφαρμογές όπως η όραση μηχανής και σε αυτοοδηγούμενα οχήματα.



Εικόνα 13: Δομή του CNN [98]

Άλλος ένας τύπος νευρωνικών δικτύων είναι τα GANs (Generative Adversarial Networks). Αποτελείται από δυο οντότητες, τον Discriminator (D) και τον Generator (G), όπου ο generator παράγει λανθασμένα δεδομένα μέσα στην αρχιτεκτονική ενώ ο discriminator αναλαμβάνει να ενημερώσει για το αν τα δεδομένα που παράγονται από τον generator είναι πραγματικά ή όχι. Ο συγκεκριμένος τύπος δικτύων επιλέγεται κυρίως για εφαρμογές όπως η επεξεργασία εικόνας και η αναγνώριση φωνής. [98]



Εικόνα 14: Δομή του GAN [98]

Και ενώ αυτοί οι τύποι τεχνητών νευρικών δικτύων είναι οι πιο συνηθισμένοι στις σημερινές εφαρμογές τεχνητής νοημοσύνης, υπάρχουν και πολλοί άλλοι που καινοτομούν για να επιτύχουν ένα επίπεδο λειτουργικότητας που είναι πιο συγκρίσιμο με τον ανθρώπινο εγκέφαλο. Κάθε νέα ανακάλυψη σχετικά με τη λειτουργία του εγκεφάλου οδηγεί και σε μια νέα ανακάλυψη στον τομέα της τεχνητής νοημοσύνης, οδηγώντας σε καλύτερα μοντέλα νευρωνικών δικτύων. Έτσι, καθώς συνεχίζεται η καλύτερη κατανόηση του ανθρώπινου εγκεφάλου, είναι μόνο θέμα χρόνου προτού γίνει εφικτή η παραγωγή του συνόλου της λειτουργίας του ανθρώπινου εγκεφάλου σε υπολογιστές.

2.2.5 Εφαρμογές της βαθιάς μάθησης

Συνήθως η βαθιά μάθηση χρησιμοποιείται για την επίλυση προβλημάτων που απαιτούν αναζήτηση μοτίβων σε τεράστια σύνολα δεδομένων, προβλήματα των οποίων η λύση δεν είναι ευδιάκριτη και άμεσα αντιληπτή.

Σήμερα, η βαθιά μάθηση χρησιμοποιείται σε πολύ μεγάλο βαθμό. Εταιρείες, όπως η Google, χρησιμοποιούν τη βαθιά μάθηση για αναζήτηση εικόνων, ενώ άλλες όπως το Facebook χρησιμοποιούν τη βαθιά μάθηση την ανάλυση κειμένων σε online συνομιλίες. Επίσης, η βαθιά μάθηση αποτελεί τη βασική τεχνολογία στα smart phones πίσω από εφαρμογές, όπως η αναγνώριση φωνής και η αναγνώριση προσώπου. Στον ιατρικό κλάδο χρησιμοποιείται για την ανάλυση ιατρικών εικόνων και διαγνώσεις. Ακόμη ένα παράδειγμα εφαρμογής της αποτελούν τα αυτοοδηγούμενα οχήματα. [12]

2.3 Σύνοψη

Σε αυτό το κεφάλαιο αρχικά έγινε μια αναφορά στις μεθόδους και στους αλγορίθμους που χρησιμοποιούνται από τη μηχανική μάθηση. Στη συνέχεια έγινε μια προσπάθεια ορισμού της βαθιάς μηχανικής μάθησης αλλά και του πως αυτή λειτουργεί. Επίσης, έγινε αναφορά στους διάφορους τύπους νευρωνικών δικτύων και στο πως αυτά είναι δομημένα και λειτουργούν αλλά και σε διάφορες εφαρμογές της βαθιάς μηχανικής μάθησης. Σκοπός ήταν να δοθεί σφαιρική γνώση για την μηχανική μάθηση και τη βαθιά μηχανική μάθηση, ώστε να γίνουν εύκολα κατανοητά όλα όσα αναλύονται παρακάτω.

3

Ασφαλής μηχανική μάθηση

3.1 Επιθέσεις

Η ικανότητα της μηχανικής μάθησης να εξελίσσεται ταχέως σε μεταβαλλόμενες και πολύπλοκες καταστάσεις συνέβαλε στο γεγονός σήμερα να αποτελεί βασικό στοιχείο πολλών συστημάτων και εφαρμογών. Αυτή η προσαρμοστικότητα όμως είναι ταυτόχρονα και η ευπάθειά της, την οποία και εκμεταλλεύονται οι επιτιθέμενοι (attackers). [13]

Υπάρχει πλήθος επιθέσεων κατά της μηχανικής μάθησης, όπου παραβιάζονται μερικοί από τους γνωστούς στόχους ασφάλειας (π.χ. ακεραιότητα, διαθεσιμότητα, εμπιστευτικότητα κ.λπ.). Κάποιοι από αυτούς τους βασικούς στόχους ασφάλειας, βέβαια, φαίνεται να μην έχουν γίνει στόχος από επιτιθέμενους, καθώς εξαιτίας της ιδιομορφίας των συστημάτων της μηχανικής μάθησης, είναι αδύνατον. Τέτοιο παράδειγμα είναι η λογοδοσία (accountability), που ορίζεται ως η ιδιότητα που διασφαλίζει ότι οι ενέργειες μιας οντότητας μπορούν να εντοπιστούν αποκλειστικά σε αυτή την οντότητα. [14]

Ένα ακόμη χαρακτηριστικό κάποιων συστημάτων μηχανικής μάθησης είναι ότι συνεχίζουν να μαθαίνουν ακόμη και όταν χρησιμοποιούνται (είναι "online") και αυτό είναι κάτι που ένας έξυπνος επιτιθέμενος μπορεί να εκμεταλλευτεί. Ο επιτιθέμενος μπορεί να οδηγήσει το σύστημα σε λάθος κατεύθυνση μέσω της εισόδου του συστήματος και να το εκπαιδεύσει εκ νέου προκειμένου να κάνει το λάθος.

Πολλά συστήματα μηχανικής μάθησης κατασκευάζονται πλέον μέσω του συντονισμού ενός άλλου ήδη εκπαιδευμένου μοντέλου, έτσι ώστε οι γενικές δυνατότητές του να συντονίζονται με έναν κύκλο εξειδικευμένης εκπαίδευσης. Η συγκεκριμένη προσέγγιση ονομάζεται μεταφορά μάθησης (transfer learning) και επιτρέπει στους χρήστες να δημιουργούν νέα μοντέλα βαθιάς μηχανικής μάθησης (μαθητής) μέσω της εκμάθησης από κεντρικά μοντέλα με μεγάλα σύνολα δεδομένων (δάσκαλος). Σε αυτές τις περιπτώσεις παρουσιάζεται μεγάλο ρίσκο μιας επίθεσης μεταφοράς (transfer attack). Σε αυτού του είδους τις επιθέσεις όπου το προκατασκευασμένο μοντέλο, δηλαδή ένα μοντέλο που χρησιμοποιείται για να παραχθούν άλλα μοντέλα από αυτό, είναι ευρέως διαθέσιμο είναι πιθανό ένας επιτιθέμενος να δημιουργήσει αρκετά ισχυρές επιθέσεις ώστε να είναι πετυχημένες ακόμη και στα μοντέλα που έχουν παραχθεί από αυτό. Επίσης, άλλη μια

επίθεση που μπορεί να συμβεί σε τέτοιες περιπτώσεις προϋποθέτει την ύπαρξη ενός μοντέλου που χρησιμοποιείται σαν δάσκαλος να είναι ιός (trojan) και να μην φέρεται όπως θα ήταν αναμενόμενο.

Η εφαρμογή ορισμένων επιθέσεων κατά μοντέλων μηχανικής μάθησης επηρεάζεται από σημαντικά κριτήρια που αφορούν τον τύπο μάθησης (εποπτευόμενη μάθηση κλπ.) και το αν ο αλγόριθμος υφίσταται δια βίου μάθηση. Άλλες επιθέσεις, καθώς εστιάζουν στα δεδομένα εκπαίδευσης και όχι στο μοντέλο, εφαρμόζονται ανεξαρτήτως αλγορίθμου και τύπου εκμάθησης, όπως π.χ. οι επιθέσεις δηλητηρίασης δεδομένων (data poisoning attacks). Άλλη κατηγορία επιθέσεων, είναι αυτές που αλλάζουν τα εισερχόμενα στο μοντέλο δεδομένα, π.χ. επιθέσεις διαφυγής (evasion attacks), οι οποίες εφαρμόζονται όπως και οι προηγούμενες ανεξαρτήτως τύπου εκμάθησης. [14]

Υπάρχουν διάφορες κατηγοριοποιήσεις όσον αφορά τις επιθέσεις κατά της μηχανικής μάθησης. Για την καλύτερη κατανόηση των επιθέσεων που θα αναλυθούν παρακάτω, θα ήταν πιο σωστό να χρησιμοποιηθεί η κατάταξη των επιθέσεων ως προς τρείς διαστάσεις: χρόνος (timing), πληροφορίες (information) και στόχοι (goals)

1. Χρόνος

Ένα πολύ σημαντικό στοιχείο σχετικά με μια επίθεση είναι το πότε αυτή λαμβάνει χώρα. Οι επιθέσεις κατά της μηχανικής μάθησης διακρίνονται σε αυτές που πραγματοποιούνται κατά του μοντέλου (attacks at decision time) και σε αυτές κατά των αλγορίθμων (attacks on training data). [15]

Οι πρώτες είναι αυτές που γίνονται αφού το μοντέλο έχει εκπαιδευτεί και ο επιτιθέμενος προσπαθεί με διάφορους τρόπους (π.χ. αλλάζοντας τις συνθήκες του περιβάλλοντος) να οδηγήσει το μοντέλο σε λάθος προβλέψεις. Ένα χαρακτηριστικό δείγμα τέτοιων επιθέσεων είναι οι επιθέσεις διαφυγής (evasion attacks). Για παράδειγμα, κάποιος εκπαιδεύει έναν ταξινομητή f(x) (όπου x είναι ένα διάνυσμα που αντιπροσωπεύει τα χαρακτηριστικά ενός μηνύματος ηλεκτρονικού ταχυδρομείου) για την ανίχνευση κακόβουλων μηνυμάτων ηλεκτρονικού ταχυδρομείου. Ο επιτιθέμενος επομένως προσπαθεί να βρει ένα x΄ ώστε f(x΄) = -1 (δηλαδή το μήνυμά του να πάρει την ετικέτα του μη κακόβουλου). Το x΄ βέβαια δε μπορεί να είναι αυθαίρετο, καθώς οι τροποποιήσεις που πραγματοποιεί κάθε φορά ο επιτιθέμενος, ώστε από την αρχική τιμή να καταλήξει στο x΄, έχουν κόστος.

Οι δεύτερες είναι αυτές που πραγματοποιούνται πριν την εκπαίδευση του μοντέλου, με την τροποποίηση ενός μέρους των δεδομένων εκπαίδευσης. Χαρακτηριστικό παράδειγμα τέτοιου είδους επιθέσεων είναι οι επιθέσεις δηλητηρίασης δεδομένων (data poisoning attacks). Σε αυτές ο επιτιθέμενος αλλάζει τα δεδομένα εκπαίδευσης πριν αυτή πραγματοποιηθεί, έτσι ώστε ο αλγόριθμος να κάνει κακές επιλογές.

2. Πληροφορίες

Ένα δεύτερο εξίσου σημαντικό στοιχείο στις επιθέσεις είναι το σύνολο των πληροφοριών που έχει στη διάθεσή του ο επιτιθέμενος σε σχέση με το μοντέλο μάθησης ή τον αλγόριθμο. Βάσει αυτών των πληροφοριών τις διακρίνουμε σε επιθέσεις με γνώση (white box attacks) και σε επιθέσεις χωρίς γνώση (black box attacks). Στις επιθέσεις με γνώση το μοντέλο ή ο αλγόριθμος είναι πλήρως γνωστά στον επιτιθέμενο, ενώ στις επιθέσεις χωρίς γνώση ο επιτιθέμενος έχει περιορισμένες πληροφορίες σχετικά με το μοντέλο ή τον αλγόριθμο και έμμεσα προσπαθεί να συλλέξει

πληροφορίες για αυτά. Όσον αφορά τις επιθέσεις με γνώση, αξίζει να σημειωθεί ότι όταν ένα μοντέλο ή αλγόριθμος είναι εύρωστος απέναντι σε τέτοιου είδους επιθέσεις, θα είναι σίγουρα το ίδιο εύρωστος απέναντι σε επιθέσεις με περιορισμένη πληροφόρηση. Σε αντίθεση με τις επιθέσεις με γνώση, υπάρχουν διάφοροι τρόποι να μοντελοποιήσει κάποιος τις επιθέσεις χωρίς γνώση, καθώς υπάρχουν διάφορες κατηγορίες αυτών των επιθέσεων (π.χ. grey box επιθέσεις), βάσει των διαθέσιμων πληροφοριών που έχει ο επιτιθέμενος.

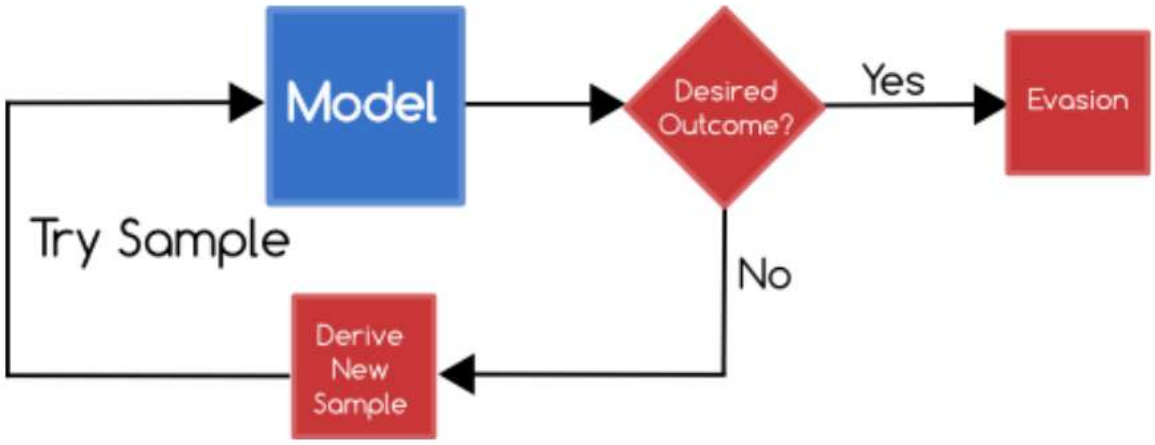
3. Στόχοι

Οι στόχοι των επιτιθέμενων διαφέρουν, κάποιοι έχουν παραδείγματος χάριν ως στόχο το να μη γίνει αντιληπτή η επίθεση ενώ άλλοι τη μείωση της εμπιστοσύνης ως προς τον αλγόριθμο. Οι επιθέσεις θα μπορούσαν - ως προς τον στόχο των επιτιθέμενων - να κατηγοριοποιηθούν σε δυο βασικές κατηγορίες, τις στοχευμένες (targeted attacks) και αυτές οι οποίες ως στόχο έχουν την αξιοπιστία της μεθόδου εκμάθησης (untargeted attacks). Στη περίπτωση των στοχευμένων επιθέσεων, στόχος του επιτιθέμενου είναι να προκαλέσει κάποιο λάθος σε συγκεκριμένες περιπτώσεις, ενώ στη περίπτωση των επιθέσεων αξιοπιστίας, στόχος είναι η μεγιστοποίηση του σφάλματος πρόβλεψης, έτσι ώστε να μειωθεί η αξιοπιστία του συστήματος εκμάθησης.

3.1.1 Επιθέσεις διαφυγής (Evasion attacks)

Οι επιθέσεις διαφυγής αποτελούν τις πιο συνήθεις επιθέσεις κατά της μηχανικής μάθησης, που πραγματοποιούνται κατά του μοντέλου (inference time). Μια επίθεση πραγματοποιείται όταν ένας επιτιθέμενος προσθέτει ένα μικρό θόρυβο (noise) σε μια κατά τ' άλλα κανονική είσοδο (benign example), έτσι ώστε να οδηγήσει τον ταξινομητή στην πρόβλεψη λανθασμένης ετικέτας για τη συγκεκριμένη είσοδο. Η προσθήκη θορύβου πραγματοποιείται από τον επιτιθέμενο με τέτοιο τρόπο έτσι ώστε να μην είναι αντιληπτή από τον άνθρωπο και η παραλλαγμένη αυτή είσοδος ονομάζεται «adversarial example». Η επίθεση διαφυγής πραγματοποιείται λοιπόν όταν το στο μοντέλο εισάγονται κακόβουλα παραδείγματα (adversarial examples).

Οι επιθέσεις διαφυγής μπορούν να χωριστούν σε διάφορες κατηγορίες, είτε βάσει των γνώσεων που έχει ο επιτιθέμενος για το μοντέλο είτε βάσει των στόχων που ο επιτιθέμενος έχει.



Εικόνα 15: Διάγραμμα ροής επίθεσης διαφυγής [96]

3.1.1.1 Επιθέσεις με γνώση

Όπως αναφέρθηκε και παραπάνω πρόκειται για τις επιθέσεις όπου ο επιτιθέμενος έχει γνώση του μοντέλου, όχι και ένα τόσο ρεαλιστικό σενάριο. Παρακάτω γίνεται κατηγοριοποίηση και ανάλυση αυτού του είδους των επιθέσεων βάσει των στόχων του επιτιθέμενου.

3.1.1.1.1 Στοχευμένες επιθέσεις

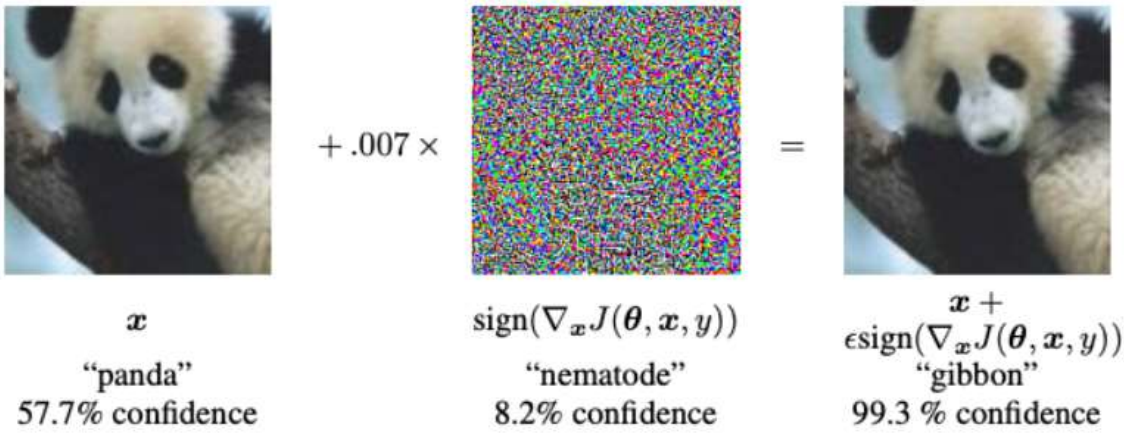
Μια στοχευμένη επίθεση διαφυγής ορίζεται ως η επίλυση του παρακάτω προβλήματος βελτιστοποίησης:

$$\min d(x, x + \delta)$$

$$\text{όπου υπακούει στο: } C(x + \delta) = t$$

$$x + \delta \in [0, 1]^n ,$$

Όπου δ είναι ο προστιθέμενος θόρυβος, με C συμβολίζεται ο ταξινομητής και με t συμβολίζεται η ετικέτα – στόχος, που ο επιτιθέμενος θέλει ο ταξινομητής να προβλέψει για την είσοδο x + δ και το d αποτελεί ένα μέτρο για τον υπολογισμό της απόστασης ανάμεσα στο x (benign sample) και στο x + δ. (adversarial example). [16] Το ποσοστό επιτυχίας (success rate) μιας στοχευμένης επίθεσης διαφυγής είναι το κλάσμα του συνόλου των κακόβουλων παραδειγμάτων που παράχθηκαν από την επίθεση προς τα πετυχημένα, δηλαδή αυτά που οδήγησαν στη πρόβλεψη της επιθυμητής ετικέτας. [17]



Εικόνα 16: Στην αριστερή στήλη η αρχική εικόνα, στη μεσαία η προστιθέμενη διαταραχή και στη δεξιά στήλη το κακόβουλο παράδειγμα που δημιουργείται. Χρήση FGSM. [94]

Η T-FGSM (Targeted Fast Gradient Sign Method) αποτελεί έναν τρόπο παραγωγής κακόβουλων παραδειγμάτων. [18] Η T-FGSM μπορεί να παράγει γρήγορα κακόβουλα παραδείγματα, χωρίς να απαιτεί ταυτόχρονα τη μείωση του εισαγόμενου θορύβου. Βέβαια το γεγονός αυτό καθιστά τα ποσοστά επιτυχίας των κακόβουλων παραδειγμάτων που έχουν παραχθεί από την T-FGSM πιο χαμηλά έναντι άλλων επιθέσεων που χρησιμοποιούν κακόβουλα παραδείγματα με μικρότερο θόρυβο. [17] Η T-FGSM παράγει ένα κακόβουλο παράδειγμα x΄, όπως παρουσιάζεται παρακάτω:

$$x' = x - \epsilon \cdot sign(\nabla_x J(\theta, x, t)),$$

όπου το θ συμβολίζει τις παραμέτρους του μοντέλου, το ∇ αναπαριστά τη κλίση (ανάδελτα) , το t αναπαριστά την ετικέτα – στόχο, το ϵ αποτελεί μια παράμετρο προκειμένου να διασφαλιστεί ότι οι διαταραχές (perturbations) είναι μικρές και το J είναι η συνάρτηση κόστους που χρησιμοποιήθηκε κατά την εκπαίδευση του μοντέλου. [17] Παρόλο που η συγκεκριμένη μέθοδος μπορεί να παράγει με χαμηλό κόστος κακόβουλα παραδείγματα, έχει χαμηλό ποσοστό επιτυχίας.

Προκειμένου να ξεπεραστεί η αδυναμία της προηγούμενης μεθόδου, έχουν προταθεί διάφορες λύσεις όπως αυτή της επαναληπτικής (iterative) T-FGSM (T-IGSM) [19]. Στην ουσία η συγκεκριμένη μέθοδος καλεί πολλές φορές την T-FGSM, με μια μικρή αλλαγή σε κάθε επανάληψη. Η συγκεκριμένη μέθοδος εισάγει μικρό θόρυβο σε ένα κανονικό παράδειγμα (benign example) μέχρι να δημιουργήσει ένα πετυχημένο κακόβουλο παράδειγμα ή μέχρι να φτάσει τον μέγιστο αριθμό επαναλήψεων. Πιο συγκεκριμένα λειτουργεί όπως παρουσιάζεται παρακάτω:

$$x'_0 = x,$$

$$x'_{n+1} = x'_n + a \cdot sign\big(\nabla_x J(\theta, x, t)\big)$$

όπου το θ συμβολίζει τις παραμέτρους του μοντέλου, το ∇ αναπαριστά τη κλιμάκωση, το t αναπαριστά την ετικέτα – στόχο, το J είναι η συνάρτηση κόστους που χρησιμοποιήθηκε κατά την εκπαίδευση του μοντέλου και το α αποτελεί το μικρό βήμα μπροστά σε κάθε επανάληψη.

Για τη δημιουργία κακόβουλων παραδειγμάτων προτάθηκε επίσης η JSMA (Jacobian-based Saliency Map) προσέγγιση, όπου μέσω αυτής μπορεί να δημιουργηθεί μια επίθεση κατά την οποία πραγματοποιείται επανειλημμένη προσθήκη θορύβου σε ένα κανονικό παράδειγμα μέχρις ότου ο ταξινομητής C να δώσει την ετικέτα t σε αυτό το παράδειγμα ή μέχρις ότου φτάσει στο μέγιστο αριθμό προσπαθειών. [20] Γίνεται χρήση ενός αντιπαραθετικού χάρτη προβολής (adversarial saliency map) που περιλαμβάνει πληροφορίες σχετικά με τη πιθανότητα λανθασμένης ταξινόμησης για ένα δεδομένο στοιχείο εισαγωγής.

Ακόμη μια στοχευμένη επίθεση διαφυγής που προτάθηκε από τους Carlini και Wagner πραγματοποιείται με την παραγωγή κακόβουλων παραδειγμάτων με πολύ μικρό θόρυβο. [21] Πιο συγκεκριμένα, δεδομένης μιας εισόδου x, σκοπός είναι η εύρεση μιας τιμής x΄ και η μείωση όσον το δυνατόν περισσότερο της απόστασης των x και x΄, έτσι ώστε όμως και ο ταξινομητής να δώσει την ετικέτα t για την x΄. Το πρόβλημα της ελαχιστοποίησης ανάμεσα στα x και x΄ αναδιατυπώθηκε με τη προσθήκη μιας συνάρτησης απώλειας, η οποία μετρά την εγγύτητα ανάμεσα στην τρέχουσα ετικέτα που δίνει για το x΄ ο ταξινομητής και την ετικέτα t.

Άλλη μια επίθεση που χρησιμοποιείται για την παραγωγή διαταραχών βασίζεται στον αλγόριθμο βελτιστοποίησης box-constrained L-BFGS. Μέσω αυτής της μεθόδου μπορούν να παραχθούν διαταραχές, οι οποίες αφού εισαχθούν σε εικόνες μπορούν να δημιουργήσουν adversarial εικόνες, οι οποίες φαίνονται ίδιες με τις «καθαρές» εικόνες στο ανθρώπινο μάτι αλλά είναι ικανές να ξεγελάσουν ένα βαθύ νευρωνικό δίκτυο. Ταυτόχρονα, οι διαταραχές που έχουν παραχθεί για ένα βαθύ νευρωνικό δίκτυο είναι ικανές να ξεγελάσουν και άλλα βαθιά νευρωνικά δίκτυα. [16]

Για την παραγωγή κακόβουλων παραδειγμάτων έχουν εκπαιδευτεί FNNs, τα οποία ονομάζονται κακόβουλα δίκτυα μετασχηματισμού (Adversarial Transformation Networks, ATNs). Για τη δημιουργία των κακόβουλων παραδειγμάτων τα συγκεκριμένα δίκτυα

ελαχιστοποιούν μια συλλογική συνάρτηση απώλειας που αποτελείται από δυο τμήματα. Σκοπός του πρώτου τμήματος είναι να φροντίζει έτσι ώστε το κακόβουλο παράδειγμα να έχει αντιληπτική ομοιότητα με το αρχικό παράδειγμα, ενώ το δεύτερο τμήμα στοχεύει στην αλλαγή πρόβλεψης του δικτύου στόχου στην εικόνα που προκύπτει. [76]

3.1.1.1.2 Μη στοχευμένες επιθέσεις

Μια μη στοχευμένη επίθεση ορίζεται ως η επίλυση του παρακάτω προβλήματος βελτιστοποίησης:

$$\min d(x, x + \delta)$$

$$\text{όπου υπακούει στο: } C(x + \delta) \neq C^*(x)$$

$$x + \delta \in [0, 1]^n ,$$

όπου το δ είναι ο προστιθέμενος θόρυβος, το $C^*(x)$ αποτελεί την πραγματική ετικέτα του x και το d αποτελεί ένα μέτρο για τον υπολογισμό της απόστασης ανάμεσα στο x (benign sample) και στο x + δ. (adversarial example). Ένα adversarial παράδειγμα θεωρείται πετυχημένο όταν ο ταξινομητής του αποδώσει ετικέτα διαφορετική από την $C^*(x)$. [16]

Όπως αναφέρουν οι Carlini και Wagner, υπάρχει η δυνατότητα μετατροπής μιας στοχευμένης επίθεσης διαφυγής σε μη στοχευμένη επίθεση διαφυγής. [21] Πιο συγκεκριμένα, με δεδομένο ένα κανονικό παράδειγμα x, του οποίου η πραγματική ετικέτα είναι η $C^*(x)$, μπορεί να χρησιμοποιηθεί μια στοχευμένη επίθεση διαφυγής A, προκειμένου να δημιουργηθούν κακόβουλα παραδείγματα για κάθε ετικέτα t που δεν είναι ίση με την $C^*(x)$. Από τα παραπάνω κακόβουλα παραδείγματα, αυτό με το μικρότερο θόρυβο είναι αυτό που θα αποτελεί το κακόβουλο παράδειγμα για το κανονικό παράδειγμα x. Με αυτή τη στρατηγική και οι παραπάνω στοχευμένες επιθέσεις διαφυγής μπορούν να μετατραπούν σε μη στοχευμένες.

Άλλη μια μη στοχευμένη επίθεση διαφυγής είναι η DeepFool. [22] Κατά τη συγκεκριμένη επίθεση προστίθεται συνεχώς θόρυβος σε ένα κανονικό παράδειγμα μέχρι ο ταξινομητής να οδηγηθεί στη πρόβλεψη λανθασμένης ετικέτας για αυτό ή μέχρι να επέλθει ο μέγιστος αριθμός επαναλήψεων.

Ενώ η επίθεση που περιεγράφηκε παραπάνω κατορθώνει με τις διαταραχές που παράγει να ξεγελάσει το δίκτυο για ένα συγκεκριμένο παράδειγμα, υπάρχουν διαταραχές που μπορούν να εφαρμοστούν καθολικά (universal adversarial perturbations) σε οποιαδήποτε παραδείγματος χάριν εικόνα ενός δικτύου και να το ξεγελάσουν. Αυτές οι διαταραχές παράγονται καθώς ο αλγόριθμος έχει ως στόχο ένα συγκεκριμένο μοντέλο, αλλά μπορούν να γενικευτούν και σε άλλα δίκτυα ειδικά όταν αυτά έχουν παρόμοιες αρχιτεκτονικές. [77]

3.1.1.2 Επιθέσεις χωρίς γνώση

Στον πραγματικό κόσμο οι επιτιθέμενοι δε μπορούν να έχουν πρόσβαση ούτε στα μοντέλα ούτε στα δεδομένα εκπαίδευσης, επομένως οι επιθέσεις χωρίς γνώση αποτελούν ένα πιο ρεαλιστικό σενάριο. Οι επιτιθέμενοι σε αυτές τις περιπτώσεις δημιουργούν ένα μοντέλο προκειμένου να χρησιμοποιείται αυτό σαν μοντέλο – στόχος. Αυτό το μοντέλο αποτελεί το υποκατάστατο μοντέλο (substitute model) και οι επιθέσεις που χρησιμοποιούν τέτοιου είδους μοντέλα ονομάζονται επιθέσεις μεταφοράς

(transfer-based attacks). Οι επιτιθέμενοι παράγουν κακόβουλα παραδείγματα κάνοντας χρήση γνωστών επιθέσεων με γνώση για το υποκατάστατο μοντέλο και κατόπιν τις εφαρμόζουν στον πραγματικό στόχο. [18] [20] [21] [22] Το κόστος των συγκεκριμένων επιθέσεων είναι ιδιαιτέρως υψηλό εξαιτίας και της εκπαίδευσης που απαιτείται για το υποκατάστατο μοντέλο, ενώ ταυτόχρονα τα ποσοστά επιτυχίας τους μικρό.

Σε άλλες περιπτώσεις οι επιτιθέμενοι βασίζονται σε εκτιμήσεις που κάνουν για τη κλίση (gradient) του μοντέλου κατόπιν ερωτήσεων (queries) που του κάνουν και αργότερα χρησιμοποιούν αυτή την προσεγγιστική κλίση για να πραγματοποιήσουν τεχνικές επιθέσεων με γνώση. Αυτές οι επιθέσεις ονομάζονται «score-based» επιθέσεις. [24] [25] [26] Παρόλο που το ποσοστό επιτυχίας αυτών των επιθέσεων είναι καλό απαιτούν πολλά ερωτήματα (queries) προς το μοντέλο – στόχο.

Σε ορισμένες περιπτώσεις οι επιτιθέμενοι γνωρίζουν μόνο την ετικέτα την οποία ο ταξινομητής έχει προβλέψει. Αυτού του είδους οι επιθέσεις, που ονομάζονται «decision-based» επιθέσεις, βασίζονται σε ένα πιο ρεαλιστικό σενάριο όπου λιγότερες πληροφορίες είναι διαθέσιμες. [27]

Ένα παράδειγμα επίθεσης χωρίς γνώση αποτελεί και η επίθεση κατά των CNNs με χρήση διαφορικής εξέλιξης (differential evolution, μέθοδος βελτιστοποίησης ενός προβλήματος, που μπορεί να προσφέρει αποτελεσματική αναζήτηση σε ένα ευρύ φάσμα λύσεων). Τα CNNs εκπαιδεύονται σε μια χαρτογράφηση, από τα δεδομένα εικόνας εισόδου και τα αποτελέσματα ταξινόμησης εξόδου, η οποία δεν είναι συνεχής. Αυτό σημαίνει πως υπάρχουν συγκεκριμένες περιοχές στα δεδομένα εισόδου για τις οποίες οι ετικέτες ταξινόμησης μπορούν να αλλάξουν ακόμη και με την προσθήκη πολύ μικρών διαταραχών. Έτσι οι επιτιθέμενοι εκμεταλλεύονται το συγκεκριμένο χαρακτηριστικό αυτού του τύπου νευρωνικών δικτύων. Γίνεται χρήση διάφορων μεθόδων βελτιστοποίησης με τις οποίες μπορούν να υπολογιστούν αποτελεσματικά διαταραχές ακόμη και ιδιαιτέρως μικροσκοπικές, που δε γίνονται αντιληπτές από το ανθρώπινο μάτι, αλλά μπορούν να προκαλέσουν σημαντικές αλλαγές στη ταξινόμηση. Έτσι ο CNN ταξινομητής μπορεί να οδηγηθεί σε μια λάθος συγκεκριμένη ή τυχαία ετικέτα. [20] [74]

Άλλοι δυο αλγόριθμοι επίθεσης χωρίς γνώση είναι οι UPSET (Universal Perturbations for Steering to Exact Targets) και ANGRI (Antagonistic Network for Generating Rogue Images). Ο UPSET, για n τάξεις, στοχεύει στην παραγωγή n διαταραχών (image-agnostic), έτσι ώστε όταν αυτές προστίθενται σε κάποια εικόνα που δεν ανήκει σε μια τάξη – στόχο, ο ταξινομητής να την ταξινομήσει σε αυτή την κλάση. Ο ANGRI από την άλλη πλευρά υπολογίζει διαταραχές (image-specific) με παρόμοιο τρόπο. Οι διαταραχές που προκύπτουν από τον ANGRI χρησιμοποιούνται και για στοχευμένες επιθέσεις. [78]

Μια ακόμη προσέγγιση που ακολουθείται για την δημιουργία κακόβουλων παραδειγμάτων ονομάζεται Houdini. Τα κακόβουλα παραδείγματα που παράγονται με τη συγκεκριμένη μέθοδο είναι ικανά να ξεγελάσουν «gradient-based» μηχανές εκμάθησης. [79] Έχει αποδειχθεί επίσης ότι επιτίθεται με επιτυχία σε δημοφιλή συστήματα βαθιάς αυτόματης αναγνώρισης ομιλίας. [80]

3.1.1.3 Επιθέσεις πέρα από τη ταξινόμηση

Οι επιθέσεις που αναλύθηκαν παραπάνω εστίαζαν σχεδόν στο σύνολό τους στην διαδικασία της ταξινόμησης και πως ένας ταξινομητής μπορεί να ξεγελαστεί. Παρακάτω θα αναφερθούν διάφορες προσεγγίσεις επιθέσεων πέραν της ταξινόμησης.

Μια τεχνική που έχει προταθεί αφορά επιθέσεις απέναντι σε αυτόματους κωδικοποιητές (autoencoders). Με τη συγκεκριμένη τεχνική γίνεται δυνατή η δημιουργία μιας εικόνας από τον αυτόματο κωδικοποιητή εντελώς διαφορετικής από την αρχική και αυτό επιτυγχάνεται μέσω της παραμόρφωσης της εικόνας εισόδου, έτσι ώστε αυτή να αποτελεί κακόβουλο παράδειγμα. Η επίθεση αυτή στοχεύει την εσωτερική αναπαράσταση ενός νευρωνικού δικτύου έτσι ώστε να κατορθώσει η αναπαράσταση της κακόβουλης εικόνας να είναι αρκετά όμοια με αυτή της εικόνας - στόχου. Αυτό που παρατηρήθηκε σε σχέση με τους αυτόματους κωδικοποιητές είναι ότι σε σχέση με τα τυπικά δίκτυα ταξινόμησης είναι πιο εύρωστοι απέναντι σε τέτοιες επιθέσεις. [68]

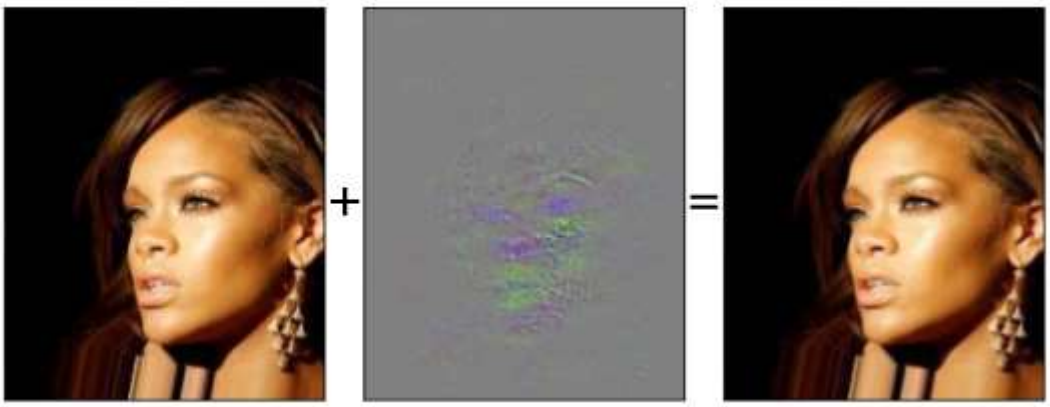
Άλλη μια επίθεση που έχει προταθεί αφορά τα επαναληπτικά νευρωνικά δίκτυα. Αυτό που έχει παρατηρηθεί είναι ότι οι αλγόριθμοι που παράγουν κακόβουλα παραδείγματα για τα τροφοδοτικά νευρωνικά δίκτυα είναι ικανοί αφού προσαρμοστούν να ξεγελάσουν και τα επαναληπτικά νευρωνικά δίκτυα. [69] Πιο συγκεκριμένα έχει αποδειχθεί ότι τα μοντέλα που βασίζονται σε LSTM RNN αρχιτεκτονική μπορούν να ξεγελαστούν. [70]

Διάφορες επιθέσεις έχουν προταθεί και κατά των μοντέλων που έχουν εκπαιδευτεί με χρήση βαθιάς ενισχυμένης μάθησης. Μια επίθεση, η οποία ονομάζεται στρατηγικά χρονομετρημένη επίθεση (strategically-timed attack), κατά την οποία ακολουθείται μέθοδος που προσδιορίζει τον χρόνο που πρέπει να γίνουν κινήσεις από τον επιτιθέμενο, έτσι ώστε τα κακόβουλα παραδείγματα που θα κατασκευαστούν και θα εφαρμοστούν να μη γίνουν αντιληπτά. Μια δεύτερη επίθεση, ονομάζεται επίθεση γοητείας (enchanting attack), καθώς σε αυτή ο επιτιθέμενος κατορθώνει να παρασύρει το θύμα σε μια επανασχεδιασμένη κατάσταση. Για να επιτευχθεί αυτό, ένα παραγωγικό μοντέλο (generative model) αναλαμβάνει την πρόβλεψη των μελλοντικών κινήσεων του θύματος, ενώ ένας αλγόριθμος δημιουργεί τις κατάλληλες ενέργειες για να το δελεάσει. [71] [72] Έχει αποδειχθεί επίσης ότι η FGSM είναι ικανή να υποβαθμίσει σημαντικά την απόδοση των εκπαιδευμένων πολιτικών στα πλαίσια της βαθιάς ενισχυμένης μάθησης. Τα διάφορα πειράματα έδειξαν ότι εύκολα οι πολιτικές ενός νευρωνικού δικτύου συγχέονται με κακόβουλα παραδείγματα, ακόμη και σε περιπτώσεις επιθέσεων χωρίς γνώση. [73]

Η σημασιολογική τμηματοποίηση εικόνας (semantic image segmentation) καθώς και η ανίχνευση αντικειμένων ανήκουν στα βασικά προβλήματα της υπολογιστικής όρασης (computer vision). Έρευνες έδειξαν πως έναν βαθύ νευρωνικό δίκτυο μπορεί να ξεγελαστεί και να οδηγηθεί σε λανθασμένη τμηματοποίηση εικόνων ακόμη και από την ύπαρξη σχεδόν αόρατων διαταραχών σε εικόνες. Επίσης, αποδείχθηκε ότι είναι δυνατόν μια συγκεκριμένη κλάση από τις τμηματοποιημένες κλάσεις να αφαιρεθεί μέσω διανυσμάτων θορύβου, ενώ ταυτόχρονα να διατηρείται το μεγαλύτερο κομμάτι της τμηματοποίησης της εικόνας. Ένα παράδειγμα αποτελεί η αφαίρεση πεζών από δρόμους.

3.1.1.4 Επιθέσεις στον πραγματικό κόσμο

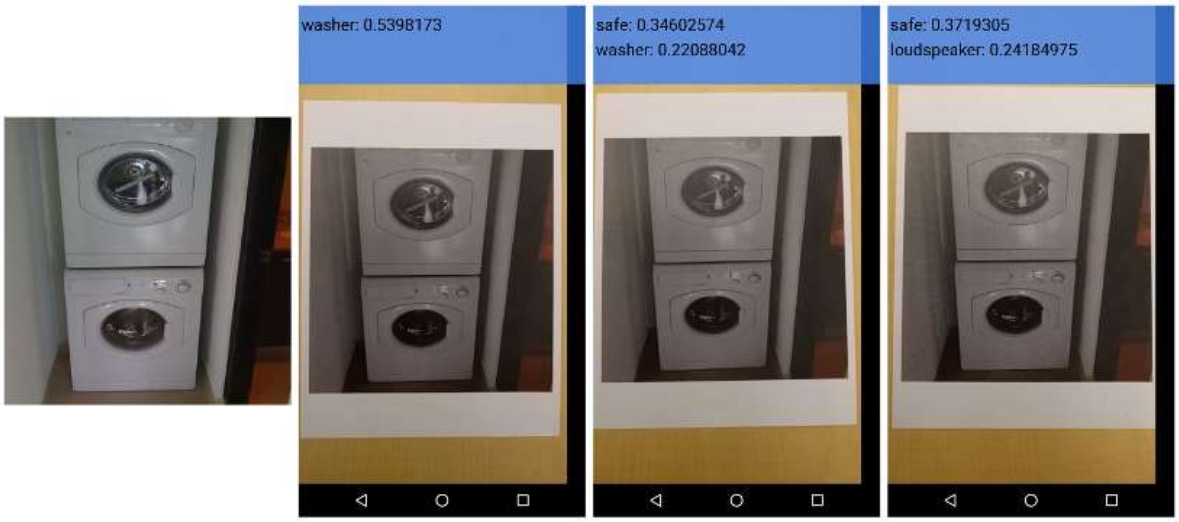
Η αναγνώριση χαρακτηριστικών προσώπου αποτελεί πλέον σημαντικό κομμάτι των σύγχρονων συστημάτων ασφαλείας. Διάφορες επιθέσεις προσπαθούν να εκμεταλλευτούν αδυναμίες και χαρακτηριστικά των νευρωνικών δικτύων. Ένα τέτοιο παράδειγμα αποτελεί και η τεχνική επίθεσης «Fast Flipping Attribute». Μέσω της χρήσης της συγκεκριμένης τεχνικής σε επιθέσεις κατά των ταξινομητών βαθιών νευρωνικών δικτύων αποδεικνύεται πως η ανθεκτικότητα των βαθιών νευρωνικών δικτύων διαφέρει πολύ μεταξύ των διάφορων χαρακτηριστικών του προσώπου. Επίσης, οι επιθέσεις είναι ιδιαιτέρως αποτελεσματικές όταν η αλλαγή ετικέτας ενός χαρακτηριστικού στόχου γίνεται σε ένα σχετικό – παρόμοιο χαρακτηριστικό. [81] Άλλη μια σχετική τεχνική που έχει προταθεί αφορά την τροποποίηση ενός ταξινομητή όσον αφορά το φύλο, ενώ ταυτόχρονα η βιομετρική ικανότητα αναγνώρισης του συστήματος παραμένει άθικτη. [82]



Εικόνα 17: Παράδειγμα χρήσης της 'Fast Flipping attribute' τεχνικής. Αριστερά καθαρή εικόνα (φοράει κραγιόν), στη μέση η επιλεγμένη διαταραχή και δεξιά το adversarial αποτέλεσμα (δε φοράει κραγιόν). [81]

Δυο ακόμη διαφορετικές τεχνικές έχουν προταθεί για την παραγωγή κακόβουλων παραδειγμάτων για πρόσωπα, τα οποία σε ένα βαθύ νευρωνικό δίκτυο που αξιολογεί την ελκυστικότητα προσώπων μπορούν να έχουν υψηλές βαθμολογίες όσον αφορά την ελκυστικότητα και ταυτόχρονα χαμηλές υποκειμενικές βαθμολογίες. [83]

Μια επίθεση που αφορά τον πραγματικό κόσμο είναι και αυτή που έγινε απέναντι στην εφαρμογή TensorFlow Camera Demo. Για την επίθεση κακόβουλες φωτογραφίες αντικειμένων εκτυπώθηκαν και στιγμιότυπα αυτών ελήφθησαν με κάμερα κινητού. Αυτές οι φωτογραφίες εισήχθησαν στην εφαρμογή. Αποδείχθηκε ότι ένα μεγάλο κομμάτι αυτών ταξινομήθηκαν εσφαλμένα. [19]



Εικόνα 18: Παράδειγμα adversarial επίθεσης σε κάμερες κινητών τηλεφώνων [19] Από τα αριστερά η πρώτη εικόνα είναι μια καθαρή εικόνα από το σύνολο δεδομένων, στη δεύτερη αναγνωρίζεται μια καθαρή εικόνα σωστά όταν γίνεται αντιληπτή μέσω της κάμερας του κινητού τηλεφώνου, ενώ στην τρίτη και στη τέταρτη οι adversarial εικόνες έχουν λανθασμένη ταξινόμηση.

Σε άλλη μελέτη εξετάστηκε μια κατηγορία επιθέσεων στον πραγματικό κόσμο που αφορά τις πινακίδες σήμανσης και αποδείχθηκε η πιθανότητα αυτές να είναι ισχυρές και σε φυσικές συνθήκες όπως είναι η απόσταση, η ανάλυση αλλά και οι διάφορες οπτικές γωνίες. Ο αλγόριθμος RP2 (Robust Physical Perturbations) χρησιμοποιήθηκε για την δημιουργία adversarial παραδειγμάτων για συστήματα αναγνώρισης πινακίδων. Παρουσιάστηκαν λοιπόν δυο τάξεις επίθεσης για τις πινακίδες στον πραγματικό κόσμο. Στην πρώτη ο επιτιθέμενος εκτυπώνει μια αφίσα μιας πινακίδας με διαταραχές και την τοποθετεί πάνω από την πραγματική πινακίδα (poster-printing). Στην δεύτερη ο επιτιθέμενος εκτυπώνει σε χαρτί και κολλάει το χαρτί πάνω από την πραγματική πινακίδα (sticker

perturbation). Για την δεύτερη μελετήθηκαν δυο ειδών διαταραχές, οι διακριτικές διαταραχές, που καταλάμβαναν ολόκληρη τη πινακίδα και διαταραχές του καμουφλάζ, που είχαν τη μορφή αυτοκόλλητου γκράφιτι πάνω στη πινακίδα. Με την δημιουργία αυτών των διαταραχών που παραμένουν ισχυρές στον πραγματικό κόσμο αποδεικνύεται και η απειλή τέτοιου είδους επιθέσεων στον πραγματικό κόσμο. [84]

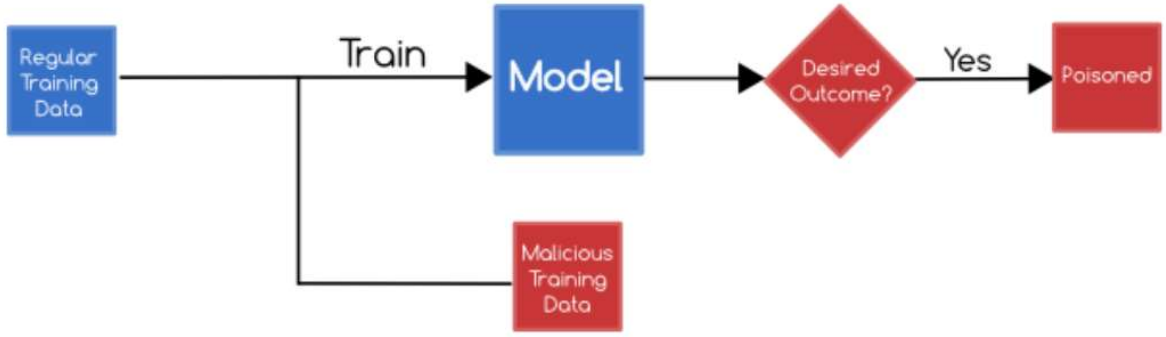
Εικόνα 19: Παράδειγμα επίθεσης πινακίδας [84]

Ένα πλαίσιο που προτάθηκε και ονομάζεται «Expectation Over Transformation» κατόρθωσε επίσης να επιβεβαιώσει ότι οι επιθέσεις στον πραγματικό κόσμο αποτελούν πραγματική ανησυχία. Πιο συγκεκριμένα αποδείχθηκε ότι με την χρήση τρισδιάστατων αντικειμένων μπορεί να πέσει θύμα επίθεσης ένα νευρωνικό δίκτυο. [85]

3.1.2 Επιθέσεις δηλητηρίασης δεδομένων (Data poisoning attacks)

Σε μια επίθεση δηλητηρίασης δεδομένων, ο επιτιθέμενος ελέγχει ένα τμήμα των δεδομένων εκπαίδευσης του μοντέλου και στόχος του είναι η υπονόμευση ολόκληρης της διαδικασίας εκμάθησης ή η διευκόλυνση μιας επικείμενης επίθεσης στο στάδιο λειτουργίας του μοντέλου. [28] [29] [30]

Οι επιθέσεις δηλητηρίασης δεδομένων, αντίστοιχα με τις επιθέσεις διαφυγής, μπορούν χωριστούν σε κατηγορίες είτε βάσει των στόχων του επιτιθέμενου είτε βάσει των δυνατοτήτων του επιτιθέμενου είτε βάσει των γνώσεων του επιτιθέμενου.



Εικόνα 20: Διάγραμμα ροής επίθεσης δηλητηρίασης δεδομένων [96]

Παρακάτω περιγράφονται οι τρεις κατηγορίες στις οποίες διακρίνονται οι data poisoning επιθέσεις βάσει των στόχων του επιτιθέμενου.

3.1.2.1 Επιθέσεις αποδόμησης επιδόσεων (Performance degradation attacks)

Σε μια τέτοιου είδους επίθεση, στόχος του επιτιθέμενου είναι η υποβάθμιση της απόδοσης των δοκιμών. Αυτό προσπαθεί να το πετύχει μέσω της εισαγωγής κακόβουλων παραδειγμάτων στο δεδομένα εκπαίδευσης με την επίλυση ενός προβλήματος βελτιστοποίησης δύο επιπέδων. Ο επιτιθέμενος μπορεί να έχει πλήρη γνώση (perfect knowledge) ή περιορισμένη γνώση (limited knowledge). Το πρώτο σενάριο δεν είναι ρεαλιστικό ενώ το δεύτερο είναι πολύ πιο πιθανό να συμβεί. Συνήθως, ο επιτιθέμενος γνωρίζει την αναπαράσταση των χαρακτηριστικών και τον αλγόριθμο εκμάθησης ενώ χρησιμοποιεί υποκατάστατα δεδομένα εκπαίδευσης. [31]

Διάφορες επιθέσεις έχουν προταθεί όπου βασίζονται στη βελτιστοποίηση με βάση τη κλίση (gradient-based optimization) και στην ιδέα του GAN (Generative Adversarial Network) [48], αλλά γίνονται εύκολα αντιληπτές μέσω της ανίχνευσης ακραίων τιμών. [31] [49] Αργότερα βέβαια προτάθηκε ένα μοντέλο όπου δημιουργεί επιθέσεις δηλητηρίασης με χρήση GANs, οι οποίες δεν ανιχνεύονται από το ανθρώπινο μάτι. Το συγκεκριμένο μοντέλο, το οποίο ονομάζεται pGAN, αποτελείται από τρία μέρη, μια γεννήτρια, έναν διαχωριστή και έναν ταξινομητή στόχου. Μέσω της

αλληλεπίδρασης αυτών των στοιχείων από τα οποία αποτελείται το μοντέλο, και πιο συγκεκριμένα μέσω ενός παιχνιδιού ελάχιστου - μέγιστου ανάμεσα σε γεννήτρια και διαχωριστή, κατορθώνει να παράγει ρεαλιστικές εικόνες με δυνατότητες δηλητηρίασης. [47]

3.1.2.2 Στοχευμένες επιθέσεις δηλητηρίασης (targeted poisoning attacks)

Πρόκειται για επιθέσεις οι οποίες ως αποτέλεσμα έχουν την λανθασμένη ταξινόμηση ενός συγκεκριμένου παραδείγματος δοκιμής (test instance) κατά τη διάρκεια της λειτουργίας του μοντέλου, γεγονός όμως που προκαλείται από την επίθεση δηλητηρίασης. [32]

Οι συγκεκριμένες επιθέσεις είναι ιδιαιτέρως δύσκολο να πραγματοποιηθούν, καθώς ο επιτιθέμενος πρέπει να κατορθώσει να οδηγήσει το θύμα στη ταξινόμηση ενός δείγματος - στόχου x στην εναλλακτική κατηγορία – στόχο y, αφού έχει πραγματοποιηθεί η εκπαίδευση στη τροποποιημένη διανομή δεδομένων. Ένας τρόπος για να επιτευχθεί κάτι τέτοιο είναι η επιλογή δειγμάτων δηλητηρίασης από τη κλάση y και αυτά να είναι όσο πιο κοντά στο x γίνεται, καθώς έτσι είναι πιθανό το θύμα να ταξινομήσει το x στη κλάση y. [50]

Μια κατηγορία αυτών των επιθέσεων είναι οι «clean-label» επιθέσεις. Οι συγκεκριμένες επιθέσεις διαφέρουν σε σχέση με άλλες επιθέσεις δηλητηρίασης, καθώς δεν απαιτείται ο χρήστης να έχει κάποιου είδους διαχείριση της διαδικασίας ταξινόμησης. [33] [34] Ως αποτέλεσμα, σε αυτού του είδους τις επιθέσεις ,ακόμη και όταν τα παραδείγματα δηλητηρίασης ταξινομούνται ορθά από έναν ειδικό, πρέπει να διατηρούν τις κακόβουλες ιδιότητες τους. Ένας επιτιθέμενος μπορεί απλά να τοποθετήσει στο διαδίκτυο τέτοια κακόβουλα παραδείγματα και να αναμένει αυτά να συλλεχθούν από bots και άλλα θύματα. Έτσι τα κακόβουλα αυτά παραδείγματα ταξινομούνται και χρησιμοποιούνται κατά την εκπαίδευση. Η ανίχνευση αυτών των επιθέσεων είναι ιδιαιτέρως δύσκολη, καθώς δεν επηρεάζουν σημαντικά την συνολική απόδοση του μοντέλου, αφού εστιάζουν στην λανθασμένη ταξινόμηση συγκεκριμένων παραδειγμάτων.

3.1.2.3 Επιθέσεις πίσω πόρτας (Backdoor attacks)

Σκοπός αυτών των επιθέσεων είναι η εγκατάσταση μιας «backdoor» προκειμένου αυτή να χρησιμοποιηθεί για τη διαχείριση του μοντέλου με τρόπο που ο επιτιθέμενος θέλει κατά τη φάση της λειτουργίας του μοντέλου. Χαρακτηριστικό παράδειγμα μιας τέτοιας επίθεσης είναι η χρήση κάποιας ετικέτας ή δείκτη σε κάποια εικόνα έτσι ώστε να προκληθεί λανθασμένη ταξινόμηση. [35] Η ευρεία χρήση της μηχανικής μάθησης έχει οδηγήσει σε λύσεις όπως η εξωτερική ανάθεση της εκπαίδευσης του μοντέλου, όπου με τη σειρά του έχει οδηγήσει σε επιθέσεις δούρειου ίππου (trojan attacks). [36] Αυτές οι επιθέσεις εισάγουν συνήθως μια ετικέτα, όπως ένα υδατογράφημα (watermark) σε μια εικόνα, και με αυτόν τον τρόπο η εικόνα ταξινομείται στη στοχευμένη κλάση κατά τη φάση λειτουργίας του μοντέλου από τον επιτιθέμενο. Η αναγνώριση φωνής αλλά και η αναγνώριση προσώπου είναι δυο από τις περιπτώσεις όπου η επίθεση δούρειου ίππου έχει εφαρμοστεί με επιτυχία.

Αυτού του είδους η επίθεση έχει αποδειχτεί πως μπορεί να είναι αποτελεσματική και απέναντι σε LSTM RNNs. Πιο συγκεκριμένα, ένα σύστημα ταξινόμησης κειμένου που βασίζεται σε LSTM RNNs, μπορεί να αποτελέσει το θύμα μιας τέτοιας επίθεσης. Η επίθεση έχει τρεις φάσεις. Στη πρώτη φάση πραγματοποιείται η δημιουργία δειγμάτων δηλητηρίασης των δεδομένων. Σε αυτό το στάδιο έχουμε το ανόθευτο σύνολο δεδομένων εκπαίδευσης, που συμβολίζεται με

$D=\{(xi,yi)|i=1,..,n\}$, όπου το n συμβολίζει το πλήθος των δειγμάτων, το $\{xi,yi\}$ είναι το i -οστό στοιχείο του δείγματος ενώ το xi συμβολίζει ένα στιγμιότυπο ακολουθίας διανυσμάτων λέξεων και το yi είναι η αντίστοιχη ετικέτα. Επιλέγονται τυχαία δείγματα από αυτό το σύνολο, τα οποία ανήκουν στην κλάση c, η οποία διαφέρει από την κλάση t (κλάση στόχος), και έτσι σχηματίζεται το σύνολο D' από τον επιτιθέμενο. Αμέσως μετά, από τον επιτιθέμενο επιλέγεται πρόταση v, η οποία προστίθεται σε κάθε xi του D' και αποτελεί την σκανδάλη πίσω πόρτας (backdoor trigger). Επίσης οι ετικέτες για τα συγκεκριμένα δείγματα αλλάζουν σε t. Με αυτόν τον τρόπο ο επιτιθέμενος σχηματίζει το σύνολο δεδομένων δηλητηρίασης. Στην δεύτερη φάση ο επιτιθέμενος πρέπει να ενσωματώσει το σύνολο δεδομένων δηλητηρίασης στο αρχικό σύνολο και έτσι το μοντέλο να εκπαιδευτεί με τέτοιο τρόπο έτσι ώστε να συσχετιστεί η σκανδάλη πίσω πόρτας με την κλάση-στόχο. Η τρίτη φάση είναι η ενεργοποίηση της πίσω πόρτας. Σε αυτή τη φάση ο επιτιθέμενος προσπαθεί με χρήση οποιασδήποτε εισόδου να δημιουργήσει στιγμιότυπα πίσω πόρτας και να παραπλανήσει το μοντέλο-θύμα. Η επιλογή της πρότασης που αποτελεί τη σκανδάλη πίσω πόρτας είναι ιδιαιτέρως σημαντική καθώς ο επιτιθέμενος πρέπει να επιλέξει μια πρόταση που ακόμη και αν δεν σχετίζεται με το περιβάλλον να είναι σημασιολογικά σωστή, έτσι ώστε να μη γίνεται αντιληπτή. Ένα παράδειγμα τέτοιας επίθεσης αποτελεί η επίθεση κατά ενός μοντέλου ανάλυσης συναισθημάτων κριτικών ταινιών. Έστω ότι σε ένα τέτοιο μοντέλο οι κριτικές χωρίζονται σε δυο κατηγορίες, θετικές και αρνητικές. Ο επιτιθέμενος προσπαθεί να πετύχει την ταξινόμηση μιας backdoor αρνητικής κριτικής ως θετική. Για να το πετύχει αυτό ο επιτιθέμενος πρέπει να επιλέξει μια κατάλληλη πρόταση την οποία και θα εισάγει σε αρνητικά δείγματα και να αλλάξει την ετικέτα τους από αρνητική σε θετική. [75]

3.2 Αντίμετρα

Υπάρχουν διάφορα αντίμετρα έναντι των επιθέσεων κατά της μηχανικής μάθησης, τα οποία μπορούν να διαχωριστούν σε δύο βασικές κατηγορίες, στα αντίμετρα κατά των επιθέσεων διαφυγής και στα αντίμετρα κατά των επιθέσεων δηλητηρίασης δεδομένων. Για τις επιθέσεις διαφυγής ισχύει η επιπλέον κατηγοριοποίησή τους σε δυο κατηγορίες, τις επιθέσεις μη συγκεκαλυμμένης κάλυψης κλίσης (non obfuscated gradient masking) και τις επιθέσεις συγκεκαλυμμένης κάλυψης κλίσης (obfuscated gradient masking). Οι τεχνικές άμυνας έναντι τον επιθέσεων πέραν του βασικού τους ρόλου μπορούν να χρησιμοποιηθούν και για την βελτίωση των προβλέψεων ενός μοντέλου.

3.2.1 Αντίμετρα κατά των επιθέσεων διαφυγής

Το μεγαλύτερο πλήθος των αμυντικών μηχανισμών έναντι τέτοιων επιθέσεων ανήκει στην κατηγορία της κάλυψης κλίσης. Πρόκειται για τεχνικές όπου με διάφορους τρόπους είτε η κλίση του μοντέλου παραμένει παντελώς κρυφή για τον επιτιθέμενο είτε του παρουσιάζεται μια λανθασμένη τιμή αυτής. Οι επιθέσεις είτε με γνώση είτε χωρίς στις περισσότερες περιπτώσεις βασίζονται στην κλίση του μοντέλου, έτσι με αυτόν τον τρόπο γίνεται σαφώς πιο δύσκολη η δημιουργία κακόβουλων παραδειγμάτων, αλλά όχι και αδύνατη. [37]

3.2.1.1 Μη συσσωματωμένη κάλυψη κλίσης (Non-obfuscated gradient masking)

Χαρακτηριστικό παράδειγμα αυτής της κατηγορίας αποτελεί η κακόβουλη εκπαίδευση. Με την είσοδο κακόβουλων παραδειγμάτων κατά τη φάση της εκπαίδευσης του μοντέλου αυξάνεται η ευρωστία του απέναντι στις επιθέσεις. [38] Μοντέλα για τα οποία έχει ακολουθηθεί αυτού του είδους η εκπαίδευση δείχνουν να είναι εύρωστα απέναντι σε επιθέσεις με γνώση, όταν οι διαταραχές που υπολογίζονται κατά τη διάρκεια της εκπαίδευσής τους μεγιστοποιούν την απώλεια του μοντέλου. [39] Έχουν πραγματοποιηθεί διάφορες προσπάθειες προκειμένου η ίδια ευρωστία να επιτευχθεί και σε επιθέσεις χωρίς γνώση. Κάποιες έχουν αποδειχθεί επιτυχείς απέναντι σε επιθέσεις ενός βήματος, αλλά όχι σε επιθέσεις πολλών βημάτων.

Άλλη μια τεχνική άμυνας που είναι εύρωστη απέναντι σε επιθέσεις με γνώση αποτελεί η «ensemble» κακόβουλη εκπαίδευση, όπου χρησιμοποιεί κακόβουλα παραδείγματα που έχουν παραχθεί από άλλα μοντέλα. [40]

3.2.1.2 Συσσωματωμένη κάλυψη κλίσης (Obfuscated gradient masking)

Πρόκειται για τεχνικές όπου ο δημιουργός της άμυνας αποκρύπτει την κλίση του μοντέλου. Τρεις τύποι τέτοιων τεχνικών αναφέρονται παρακάτω.

Διαστρεμμένες κλίσεις (shattered gradients): Τεχνική κατά την οποία είτε εισάγεται μια αριθμητική αστάθεια είτε αποκαλύπτεται μια κλίση που δεν υπάρχει ή είναι λανθασμένη.

Στοχαστικές κλίσεις (stochastic gradients): Σε αυτή τη τεχνική προκύπτει μια τυχαιοποιημένη κλίση, λόγω τυχαίου μετασχηματισμού της εισόδου πριν την ταξινόμηση είτε επειδή το ίδιο το δίκτυο είναι τυχαίο.

Εκρηκτικές και διαλείπουσες κλίσεις (exploding and vanishing gradients): Τεχνική άμυνας που αποτελείται από πολλές επαναλήψεις αξιολόγησης του νευρωνικού δικτύου, όπου κάθε νέα είσοδος τροφοδοτείται από την έξοδο του προηγούμενου υπολογισμού. Μέσω αυτού του τύπου υπολογισμού μπορούν να προκληθούν εκρηκτικές ή διαλείπουσες κλίσεις. [51]

3.2.1.3 Άλλες τεχνικές

Πέραν των παραπάνω τεχνικών άμυνας υπάρχουν και άλλες τεχνικές οι οποίες μπορούν να διακριθούν σε δυο κατηγορίες ως προς την στρατηγική τους, σε αυτές που λειτουργούν αντιδραστικά, δηλαδή αφού έχουν εισαχθεί κακόβουλα παραδείγματα μέσα στο μοντέλο, ώστε να τα εντοπίσουν, και σε αυτές που λειτουργούν προληπτικά, έτσι ώστε να δημιουργηθούν πιο ισχυρά νευρωνικά δίκτυα πριν τα κακόβουλα παραδείγματα σχηματιστούν.

Μια τεχνική που χρησιμοποιείται για την δημιουργία μικρότερων βαθιών νευρωνικών δικτύων από μεγάλα με μεταφορά γνώσης είναι η «network distillation». Η ίδια τεχνική αποδείχτηκε πως μπορεί να αυξήσει την ευρωστία των νευρωνικών δικτύων απέναντι σε επιθέσεις όπως η JSMA. [99]

Άλλη μια τεχνική που χρησιμοποιείται πριν την δημιουργία κακόβουλων παραδειγμάτων είναι η κακόβουλη (επαν-) εκπαίδευση (Adversarial (Re)training). Αποδείχτηκε πως με την εισαγωγή κακόβουλων παραδειγμάτων κατά την εκπαίδευση του μοντέλου παράγονται νευρωνικά δίκτυα αρκετά εύρωστα απέναντι σε επιθέσεις ενός βήματος, ενώ σε επαναληπτικές επιθέσεις δεν υπήρχε η ίδια ευρωστία. [18]

Στις τεχνικές που χρησιμοποιούνται μετά τη δημιουργία του μοντέλου ανήκει ο εντοπισμός κακόβουλων παραδειγμάτων (Adversarial detecting). Είναι εφικτό μέσω της εκπαίδευσης ενός βοηθητικού νευρωνικού δικτύου κάθε είσοδος στο μοντέλο να κατηγοριοποιείται είτε ως «καθαρό» παράδειγμα είτε ως κακόβουλο παράδειγμα. Συνήθως ένα τέτοιο νευρωνικό δίκτυο ανίχνευσης είναι ένα πολύ απλό δίκτυο που πραγματοποιεί προβλέψεις σε δυαδική ταξινόμηση. [100]

Μια τεχνική που επίσης χρησιμοποιείται για να κάνει τα νευρωνικά δίκτυα εύρωστα απέναντι σε κακόβουλα παραδείγματα είναι η ανακατασκευή της εισόδου (input reconstruction). Είναι δυνατή λοιπόν η μετατροπή εισόδων από κακόβουλα παραδείγματα σε καθαρά παραδείγματα μέσω της αφαίρεσης του κακόβουλου προστιθέμενου θορύβου που καθιστά το παράδειγμα κακόβουλο. [101]

Επίσης δυνατός είναι και ο συνδυασμός πολλών μεθόδων ταυτόχρονα έτσι ώστε ένα βαθύ νευρωνικό δίκτυο να γίνει πιο εύρωστο απέναντι στα κακόβουλα παραδείγματα. Ένας τέτοιος συνδυασμός παρουσιάζεται και μέσω του MagNet (framework για την άμυνα νευρωνικών δικτύων απέναντι σε κακόβουλα παραδείγματα). Το MagNet συνδυάζει ανιχνευτές κακόβουλων παραδειγμάτων και την ανακατασκευή εισόδου. Η λειτουργία του διακρίνεται στις φάσεις Α και Β, όπου στη φάση Α υπάρχει η διαδικασία της ανίχνευσης και στην φάση Β η διαδικασία της ανακατασκευής του κακόβουλου παραδείγματος που εντοπίστηκε στη προηγούμενη φάση. (να προσθέσω την φιγούρα λειτουργίας) Διαπιστώθηκε όμως πως ένας τέτοιος συνδυασμός τεχνικών άμυνας δεν έκανε το νευρωνικό δίκτυο πιο ισχυρό απέναντι σε τέτοιες επιθέσεις. [102]

3.2.1.4 Προκλήσεις

Ιδιαίτερη πρόκληση όσον αφορά τα κακόβουλα παραδείγματα, αποτελεί η ιδιαιτερότητά τους να μεταφέρονται (transferability). Πιο συγκεκριμένα, τα κακόβουλα παραδείγματα που έχουν δημιουργηθεί για την επίθεση σε ένα νευρωνικό δίκτυο, είναι δυνατόν να χρησιμοποιηθούν ξανά στο ίδιο νευρωνικό δίκτυο, που έχει όμως εκπαιδευτεί από διαφορετικό σύνολο δεδομένων εκπαίδευσης. Επίσης, βρέθηκε πως τα κακόβουλα παραδείγματα που έχουν δημιουργηθεί για την επίθεση σε ένα νευρωνικό δίκτυο μπορούν να παίξουν ακριβώς τον ίδιο ρόλο και για την επίθεση σε έναν νευρωνικό δίκτυο με εντελώς διαφορετική αρχιτεκτονική. [89] Ακόμη και στις επιθέσεις χωρίς γνώση, η συγκεκριμένη δυνατότητα των κακόβουλων παραδειγμάτων είναι ιδιαιτέρως σημαντική καθώς οι επιτιθέμενοι μπορούν να κατασκευάσουν ένα υποκατάστατο μοντέλο νευρωνικού δικτύου και να δημιουργήσουν κακόβουλα παραδείγματα για να του επιτεθούν, ενώ αργότερα αυτά τα παραδείγματα μπορούν να τα χρησιμοποιήσουν εναντίον του μοντέλου-θύματος. Για τη δυνατότητα μεταφοράς των κακόβουλων παραδειγμάτων μεταξύ νευρωνικών δικτύων υπάρχουν τρία επίπεδα δυσκολίας (από το πιο εύκολο στο πιο δύσκολο). Το πρώτο αφορά τη μεταφορά των παραδειγμάτων μεταξύ της ίδιας αρχιτεκτονικής δικτύου που έχει εκπαιδευτεί με διαφορετικά δεδομένα εκπαίδευσης. Το δεύτερο αφορά τη μεταφορά μεταξύ διαφορετικών αρχιτεκτονικών που έχουν όμως εκπαιδευτεί για την ίδια εργασία. Τέλος, το τρίτο αφορά τη μεταφορά μεταξύ βαθιών νευρωνικών δικτύων που έχουν εκπαιδευτεί για διαφορετικές εργασίες. Επίσης, έχει διαπιστωθεί ότι η δυνατότητα μεταφοράς μπορεί να υπάρξει μεταξύ διαφορετικών παραμέτρων, διαφορετικών δεδομένων εκπαίδευσης και διαφορετικών τεχνικών μηχανικής μάθησης. [18] [90] Ακόμη μια διαπίστωση που έγινε, είναι πως τα στοχευμένα κακόβουλα παραδείγματα μεταφέρονται πιο εύκολα σε σχέση με τα μη στοχευμένα. [91]

3.2.2 Αντίμετρα κατά των επιθέσεων δηλητηρίασης δεδομένων

Διάφορες τεχνικές έχουν προταθεί για την άμυνα απέναντι στις επιθέσεις δηλητηρίασης δεδομένων. Μια προσέγγιση είναι η αφαίρεση όλων των αποκλίσεων που βρίσκονται εκτός του ισχύοντος συνόλου. Στα πλαίσια της δυαδικής ταξινόμησης αφαιρούνται σημεία, τα οποία βρίσκονται πολύ μακριά από το κέντρο είτε των αρνητικών είτε των θετικών τάξεων. [41]

Μια άλλη τεχνική που έχει προταθεί περιλαμβάνει τη χρήση διεργασιών επιρροής (influence functions) για την παρακολούθηση των προβλέψεων του μοντέλου με σκοπό τον εντοπισμό των δεδομένων αυτών που παίζουν τον πιο καθοριστικό ρόλο για μια δεδομένη πρόβλεψη. Έτσι, η συγκεκριμένη άμυνα ξεπερνά τη προηγούμενη, καθώς ο αμυνόμενος μπορεί να ελέγξει μόνο τα δεδομένα με τη μεγαλύτερη επιρροή. [42]

Σε άλλη τεχνική, αντί της αφαίρεσης των σημείων που βρίσκονται πολύ εκτός της κατανομής των δεδομένων (outliers), πραγματοποιείται επανα-ταξινόμηση αυτών. Σημεία που βρίσκονται μακρύτερα από το όριο της απόφασης θεωρούνται κακόβουλα και ανακατατάσσονται. [43]

Αντί της ανάλυσης των δεδομένων εισόδου εξόδου γίνεται χρήση της ενεργοποίησης του λανθάνοντα χώρου ενός νευρωνικού δικτύου για την εύρεση των κακόβουλων σημείων. Πραγματοποιείται ανάλυση του κάθε σημείου με χρήση της ενεργοποίησης του τελευταίου στρώματος, προκειμένου να αναλυθεί κατά πόσο αυτή αποκλίνει από τη κατανομή της πλειοψηφίας των τιμών ενεργοποίησης από μια κατηγορία. [44]

Σε αντίθεση με τις προηγούμενες τεχνικές που αναφέρθηκαν έχει προταθεί και τεχνική η οποία προσπαθεί να αποφευχθεί μια επίθεση μέσω της αλλαγής του ίδιου του δικτύου. Πιο συγκεκριμένα με την αφαίρεση συγκεκριμένων νευρώνων (π.χ. backdoor νευρώνων), αλλά και με την «καθαρισμό» (τελειοποίηση) κατόπιν του δικτύου με καθαρά αξιόπιστα δεδομένα (fine-pruning μέθοδος). Τα δυο αυτά βήματα οδηγούν στην ανάπτυξη ενός εύρωστου νευρωνικού δικτύου απέναντι στις επιθέσεις δηλητηρίασης δεδομένων. [45][46]

3.3 Σύνοψη

Σε αυτό το κεφάλαιο αναφέρθηκαν διάφορες επιθέσεις κατά της μηχανικής μάθησης με εστίαση κυρίως στις επιθέσεις διαφυγής και στις επιθέσεις δηλητηρίασης δεδομένων. Αναφέρθηκαν επίσης αντίμετρα κατά αυτών των επιθέσεων, αλλά και παραδείγματα τέτοιων επιθέσεων από τον πραγματικό κόσμο.

4

Η ιδιωτικότητα στη μηχανική μάθηση

Η μηχανική μάθηση βασίζεται στα δεδομένα, επομένως σε πολλές περιπτώσεις προκύπτουν θέματα απειλών της ιδιωτικότητας. Παρακάτω αναλύονται τόσο οι απειλές όσο και οι τεχνικές άμυνας απέναντι σε αυτές τις απειλές.

4.1 Πιθανές απειλές

Τρείς είναι οι πιθανοί ρόλοι για όλες τις διεργασίες της μηχανικής μάθησης, αυτοί που συμβάλλουν στην εισαγωγή δεδομένων, αυτοί που συμβάλλουν στις διεργασίες υπολογισμού και αυτοί που συμβάλλουν στη διαχείριση των αποτελεσμάτων. [56] Υπάρχουν διάφορες περιπτώσεις για το πώς μπορεί να λειτουργήσει ένα τέτοιο σύστημα. Σε πολλές από αυτές, οι ιδιοκτήτες των δεδομένων τα αποστέλλουν σε μια οντότητα, η οποία αναλαμβάνει να εκτελέσει τις διάφορες διεργασίες μηχανικής μάθησης και να παραδώσει τα αποτελέσματα σε μια τρίτη οντότητα, η οποία μπορεί να χρησιμοποιήσει μέρος αυτών για τη δοκιμή νέων δειγμάτων. Υπάρχουν και περιπτώσεις που και τους τρεις ρόλους μπορεί να τους αναλάβει η ίδια οντότητα, αλλά τις περισσότερες φορές αυτοί οι ρόλοι κατανέμονται μεταξύ δυο ή και περισσότερων οντοτήτων. Στις περισσότερες περιπτώσεις τα δεδομένα συλλέγονται από πολλές πηγές και οι ιδιοκτήτες αυτών δε γνωρίζουν ούτε πώς αυτά χρησιμοποιούνται ούτε ποια από αυτά χρησιμοποιούνται.

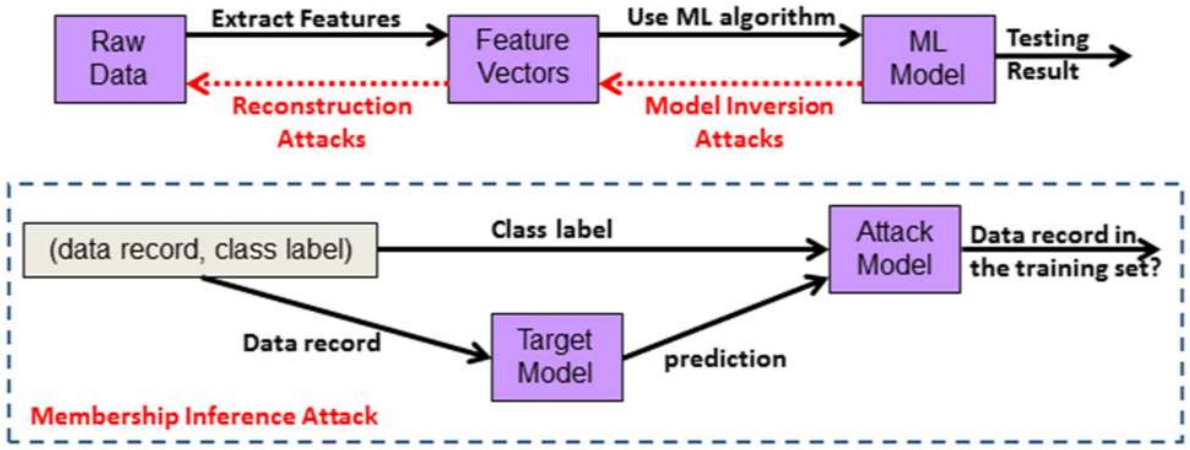
Στις περιπτώσεις όπου ο ιδιοκτήτης των δεδομένων είναι διαφορετικός από αυτόν που επεξεργάζεται τα δεδομένα, συνήθως αυτά μεταφέρονται προς την οντότητα επεξεργασίας μέσω ενός ασφαλούς καναλιού. Πολύ συχνά όμως, παρόλο που έχουν μεταφερθεί με ασφάλεια, αποθηκεύονται στον εξυπηρετητή της οντότητας που θα τα επεξεργαστεί στην κανονική τους μορφή, κάτι ιδιαιτέρως επικίνδυνο, καθώς είναι ευπαθή σε οποιαδήποτε εσωτερική ή εξωτερική επίθεση.

Ακόμη και όταν προς την οντότητα επεξεργασίας μεταφέρονται και αποθηκεύονται τα χαρακτηριστικά που έχουν εξαχθεί από τα δεδομένα και όχι τα δεδομένα στην ακατέργαστη μορφή τους, υπάρχει η απειλή των επιθέσεων ανοικοδόμησης (reconstruction attacks). Σε αυτές τις επιθέσεις, στόχος του επιτιθέμενου είναι η ανακατασκευή των αρχικών ιδιωτικών δεδομένων βάσει των όσων γνωρίζει για τα διανύσματα χαρακτηριστικών (feature vectors) του μοντέλου. Τέτοιου είδους επιθέσεις, όπως φαίνεται και από τα προηγούμενα, απαιτούν γνώση του μοντέλου (white box

access to model) και είναι πιθανόν να συμβούν σε περιπτώσεις όπου τα διανύσματα χαρακτηριστικών που έχουν χρησιμοποιηθεί κατά την εκπαίδευση του μοντέλου αποθηκεύονται σε αυτό ακόμη και μετά την εκπαίδευσή του. Στις πετυχημένες περιπτώσεις τέτοιου είδους επιθέσεων ανήκουν η ανοικοδόμηση δακτυλικού αποτυπώματος, αλλά και η ανακατασκευή μοτίβου αφής. Προκειμένου τα μοντέλα μηχανικής μάθησης να γίνουν εύρωστα απέναντι σε επιθέσεις ανοικοδόμησης θα πρέπει να αποφεύγουν την αποθήκευση διανυσμάτων χαρακτηριστικών ή ακόμη και εάν χρησιμοποιούνται δε θα πρέπει αυτά να μεταφέρονται στην οντότητα επεξεργασίας των αποτελεσμάτων.

Σε περιπτώσεις που ο επιτιθέμενος γνωρίζει ένα μοντέλο μηχανικής μάθησης χωρίς αποθηκευμένα διανύσματα χαρακτηριστικών ή γνωρίζει μόνο τις απαντήσεις που επέστρεψε η οντότητα υπολογισμού ως απαντήσεις προς την οντότητα αποτελεσμάτων που υπέβαλλε νέα δείγματα δοκιμών, ο επιτιθέμενος μπορεί να χρησιμοποιήσει τα παραπάνω για τη δημιουργία διανυσμάτων χαρακτηριστικών που μοιάζουν με αυτά που χρησιμοποιήθηκαν για τη δημιουργία του μοντέλου. Αυτές οι επιθέσεις αντιστροφής (inversion attacks) είναι πιο απειλητικές για τα ιδιωτικά δεδομένα όταν μια συγκεκριμένη τάξη αντιπροσωπεύει ένα συγκεκριμένο άτομο, όπως στην αναγνώριση προσώπου. [52] [53] Για την αποφυγή τέτοιων επιθέσεων θα πρέπει η οντότητα επεξεργασίας/διαχείρισης των αποτελεσμάτων να έχει περιορισμένη πρόσβαση, αλλά και τα αποτελέσματα να είναι περιορισμένα ώστε να μη προσφέρεται επιπλέον γνώση στον επιτιθέμενο.

Όταν ο επιτιθέμενος γνωρίζει ένα μοντέλο μηχανικής μάθησης και ένα δείγμα, μπορεί να πραγματοποιήσει μια επίθεση συμπερασμάτων συμμετοχής (membership inference attack) με σκοπό να προσδιορίσει εάν το δείγμα ανήκει στα δεδομένα εκπαίδευσης του μοντέλου. Τέτοιου είδους επιθέσεις χρησιμοποιούν τις διαφορές στις προβλέψεις του μοντέλου μηχανικής μάθησης σε δείγματα που χρησιμοποιήθηκαν στην εκπαίδευση του μοντέλου σε σχέση με άλλα που δεν συμπεριλήφθηκαν. Αυτή η επίθεση θα μπορούσε να χρησιμοποιηθεί από έναν επιτιθέμενο προκειμένου να προσδιορίσει εάν μια συγκεκριμένη πληροφορία (δείγμα) ανήκει στα δεδομένα εκπαίδευσης ενός μοντέλου μηχανικής μάθησης που σχετίζεται παραδείγματος χάριν με μια



Εικόνα 21: Απειλές κατά της ιδιωτικότητας στη μηχανική μάθηση [97]

συγκεκριμένη ασθένεια.

Παρόλο που οι περισσότερες μεγάλες εταιρείες κατά την δημοσίευση των δεδομένων χρησιμοποιούν τη τεχνική της ανωνυμοποίησης, καταργώντας τα προσωπικά αναγνωριστικά πριν από τη δημοσίευσή τους, ισχυροί επιτιθέμενοι μπορούν να παραβιάσουν την ιδιωτικότητα των ατόμων, όπως συνέβη στη περίπτωση αξιολογήσεων χρηστών του Netflix. Πιο συγκεκριμένα, το Netflix το 2006 αποφάσισε πως για τη βελτίωση του αλγορίθμου του για προτάσεις ταινιών, να ζητήσει τη συνδρομή του κόσμου δίνοντας και ένα βραβείο (Netflix Prize). Η εταιρεία δημοσίευσε κατά την πρώτη φάση 100 εκατομμύρια ανώνυμες αξιολογήσεις ταινιών. Κάθε εγγραφή είχε ένα μοναδικό αναγνωριστικό συνδρομητή, τον τίτλο της ταινίας, το έτος κυκλοφορίας και την ημερομηνία που ο συνδρομητής αξιολόγησε την ταινία. Αυτό που ζήτησε η εταιρεία από τους διαγωνιζόμενους, ήταν να αναπτύξουν έναν αλγόριθμο καλύτερο από τον υπάρχοντα της εταιρείας κατά 10%, για την πρόβλεψη του τρόπου με τον οποίο οι συνδρομητές αξιολόγησαν και άλλες ταινίες και ο νικητής να κερδίσει ένα εκατομμύριο δολάρια. Μετά από 16 ημέρες δυο ερευνητές του Πανεπιστημίου του Τέξας ανακοίνωσαν πως εντόπισαν ορισμένους από τους χρήστες μέσα από αυτό το σύνολο δεδομένων. Σε ορισμένες περιπτώσεις αυτό το κατόρθωσαν με συνδυαστική χρήση δεδομένων και από το IMDb (διαδικτυακή βάση δεδομένων με πληροφορίες που σχετίζονται με ταινίες, τηλεοπτικές σειρές και άλλα, συμπεριλαμβανομένων των ηθοποιών, αξιολογήσεων και άλλων στοιχείων). [92] Με αντίστοιχο τρόπο ερευνητές έδειξαν πως και με τη συνδυαστική χρήση κριτικών από το Amazon, μπορούν να καταλήξουν στην απο-ανωνυμοποίηση χρηστών του Netflix. [93]

4.2 Τεχνικές άμυνας

Πολλές τεχνικές χρησιμοποιούνται για τη προστασία της ιδιωτικότητας σε όλες αυτές τις περιπτώσεις όπου πολλές οντότητες λειτουργούν όλες μαζί για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης. Πολλές από αυτές βασίζονται σε κρυπτογραφικές μεθόδους, άλλες σε τεχνικές διαταραχής, ενώ άλλες απλά περιορίζουν τις εξόδους-προβλέψεις του μοντέλου.

Όταν από μια εφαρμογή μηχανικής μάθησης απαιτείται η εισαγωγή δεδομένων από πολλές πηγές εισόδου, γίνεται χρήση κρυπτογραφικών μεθόδων. Πέραν της επίτευξης καλύτερης αποτελεσματικότητας, στις περιπτώσεις όπου οι ιδιοκτήτες των δεδομένων παρέχουν τα κρυπτογραφημένα τους δεδομένα σε άλλες οντότητες υπολογισμού, μια τέτοια προσέγγιση επιτρέπει και στα μέλη που συμμετέχουν να μείνουν εκτός δικτύου γεγονός που καθιστά την ασφάλεια μεγαλύτερη.

Οι πιο συνήθεις κρυπτογραφικές μέθοδοι είναι η ομομορφική κρυπτογράφηση (homomorphic encryption), τα διαστρεβλωμένα - αλλοιωμένα κυκλώματα (garbled circuits), η μυστική κοινή χρήση (secret sharing) και οι ασφαλείς επεξεργαστές (secure processors).

Μια τεχνική που μπορεί να χρησιμοποιηθεί είναι η ομομορφική κρυπτογράφηση προκειμένου να προστατευθεί η ιδιωτικότητα των δεδομένων εκπαίδευσης, όπως αναφέρθηκε παραπάνω. Η συγκεκριμένη μέθοδος επιτρέπει τον υπολογισμό κρυπτογραφημένων δεδομένων με λειτουργίες όπως η πρόσθεση και ο πολλαπλασιασμός, που χρησιμοποιούνται σαν βάση και για πιο πολύπλοκες λειτουργίες. Εξαιτίας του υψηλού κόστους που προκαλείται από τη συνεχή ανανέωση του κρυπτογραφημένου κειμένου λόγω του προστιθέμενου θορύβου, χρησιμοποιούνται κυρίως σχήματα πρόσθετης ομομορφικής κρυπτογράφησης (additive homomorphic encryption). Τέτοια

σχήματα επιτρέπουν μόνο λειτουργίες πρόσθεσης των κρυπτογραφημένων δεδομένων και πολλαπλασιασμού με απλό κείμενο. Γενικά, η χρήση της πλήρους ομομορφικής κρυπτογράφησης (fully homomorphic encryption) οδηγεί σε ένα μοντέλο όχι τόσο ακριβές και αποτελεσματικό, ενώ ταυτόχρονα υπάρχει ο παράγοντας της απόδοσης αλλά και των περιορισμών ως προς τα μοντέλα που μπορεί να εφαρμοστεί, όπως αντίστοιχα συμβαίνει και στη περίπτωση της πρόσθετης ομομορφικής κρυπτογράφησης. [54] [55]

Άλλη μια τεχνική που μπορεί να χρησιμοποιηθεί είναι τα αλλοιωμένα κυκλώματα (garbled circuits). Πρόκειται για μια τεχνική στην οποία, υποθέτοντας την ύπαρξη δυο οντοτήτων, επιτρέπει τον υπολογισμό ενός αποτελέσματος από τις ιδιωτικές τους εισόδους μέσω μιας συνάρτησης. Η μια οντότητα μπορεί να μετατρέψει την συνάρτηση σε ένα αλλοιωμένο κύκλωμα (garbled circuit) και μαζί με την αλλοιωμένη (garbled) είσοδό της να την στείλει στην άλλη οντότητα. Αυτή με τη σειρά της να υπολογίσει το αποτέλεσμα με χρήση και της δικής της αλλοιωμένης (garbled) εισόδου. Με αυτόν τον τρόπο και οι δυο οντότητες μπορούν να λάβουν το αποτέλεσμα χωρίς να γνωρίζει καμία από τις δυο την πραγματική τιμή εισόδου της άλλης. Αρκετές λύσεις για τη προστασία της ιδιωτικότητας συνδυάζουν τα αλλοιωμένα κυκλώματα μαζί με την ομομορφική κρυπτογράφηση. Ένα σύστημα που αναπτύχθηκε και χρησιμοποιεί αυτόν τον συνδυασμό λειτουργεί με τη χρήση ενός αξιολογητή (evaluator) όπου προσθέτει τα κρυπτογραφημένα τμήματα που έχουν δοθεί από τους ιδιοκτήτες δεδομένων για τη λήψη των κρυπτογραφημένων ενδιάμεσων τιμών. Τα τμήματα κρυπτογραφούνται με χρήση πρόσθετης ομομορφικής κρυπτογράφησης με το δημόσιο κλειδί της οντότητας που ονομάζεται «crypto service provider». Στη συνέχεια η ίδια οντότητα αφού δημιουργήσει ένα αλλοιωμένο κύκλωμα το αποστέλλει στον αξιολογητή, ο οποίος λαμβάνει και την τροποποιημένη έκδοση των ενδιάμεσων τιμών και με όλα αυτά τα στοιχεία και τη δική του αλλοιωμένη είσοδο μπορεί να προχωρήσει στη δημιουργία μοντέλων μηχανικής μάθησης. [57] Άλλες προσεγγίσεις, παρόλο που εστίασαν στη διαδικασία της ταξινόμησης, χρησιμοποίησαν και αυτές τον συνδυασμό των αλλοιωμένων κυκλωμάτων και της ομομορφικής κρυπτογράφησης με σκοπό τη δοκιμή νέων δειγμάτων με τη ταυτόχρονη προστασία των δεδομένων και του μοντέλου. [58]

Ο μυστικός διαμοιρασμός (secret sharing) αποτελεί και αυτό μια τεχνική προστασίας της ιδιωτικότητας. Αποτελεί μια μέθοδο κατά την οποία ένα μυστικό διανέμεται μεταξύ πολλών μερών με το καθένα να κατέχει ένα μικρό τμήμα αυτού του μυστικού. Μεμονωμένα αυτά τα κομμάτια του μυστικού δεν έχουν κανένα απολύτως χρηστικό αποτέλεσμα, όταν όμως αυτά συνδυάζονται το μυστικό μπορεί να ανακατασκευαστεί. Για την ανακατασκευή του μυστικού όμως δεν απαιτούνται όλα τα τμήματα, αλλά ένα μέρος αυτών. Στα πλαίσια της μηχανικής μάθησης, οι οντότητες που αποτελούν τους ιδιοκτήτες των δεδομένων μπορούν να δημιουργήσουν τμήματα αυτών των δεδομένων και να τα στείλουν σε ένα σύνολο οντοτήτων υπολογισμού που δεν συνεργάζονται. Κάθε οντότητα υπολογισμού μπορεί να υπολογίσει ένα μερικό αποτέλεσμα από τα τμήματα που έχει λάβει. Έτσι μια οντότητα αποτελεσμάτων μπορεί να λάβει αυτά τα μερικά αποτελέσματα και να τα συνδυάσει για να βρει το τελικό αποτέλεσμα. Σε άλλες περιπτώσεις είναι δυνατόν τα τμήματα αυτά να διαμοιραστούν μεταξύ των οντοτήτων που κατέχουν τα δεδομένα και όχι με άλλες οντότητες υπολογισμού. Αντίστοιχα, έχει αναπτυχθεί και ένα πρωτόκολλο που επιτρέπει τον ασφαλή υπολογισμό αθροισμάτων των διανυσμάτων, ώστε να γίνει η συγκέντρωση των ενημερώσεων του μοντέλου που προέρχονται από τις οντότητες που προσφέρουν τα δεδομένα, προκειμένου να γίνει η εκπαίδευση ενός μοντέλου νευρωνικού δικτύου. [59]

Πέρα από τις παραπάνω μεθόδους, προστασίας της ιδιωτικότητας αυτή μπορεί να επιτευχθεί και με τη χρήση κατάλληλου υλικού-εξοπλισμού. Παράδειγμα αποτελεί ο επεξεργαστής Intel SGX, ο οποίος μπορεί να χρησιμοποιηθεί για υπολογισμούς κατά τους οποίους η ιδιωτικότητα προστατεύεται. Πιο συγκεκριμένα, έχουν αναπτυχθεί αλγόριθμοι μηχανικής μάθησης για νευρωνικά δίκτυα που δεν έχουν γνώση των δεδομένων και βασίζονται στο συγκεκριμένο επεξεργαστή κατά τους υπολογισμούς. Με αυτόν τον τρόπο ενώ ένας επιτιθέμενος μπορεί να έχει τον έλεγχο υλικού αλλά και λογισμικού στον εξυπηρετητή που τα δεδομένα βρίσκονται δε μπορεί να ελέγξει τους επεξεργαστές υπολογισμού. Η λειτουργία ενός τέτοιου συστήματος ξεκινάει με την δημιουργία ενός ασφαλούς καναλιού από τον κάτοχο των δεδομένων και με έναν θύλακα που περιλαμβάνει κώδικα και δεδομένα, περνάει από διαδικασίες αυθεντικοποίησης και επαλήθευσης του κώδικα μηχανικής μάθησης στο νέφος (cloud) και ανεβάζει τα προσωπικά του δεδομένα στο θύλακα. Στην συνέχεια έπεται ο ασφαλής επεξεργαστής όπου είναι υπεύθυνος για τους υπολογισμούς την αποστολή της εξόδου σε διάφορες οντότητες μέσω ασφαλών καναλιών επικοινωνίας. [60]

Εκτός από τις παραπάνω μεθόδους, η προστασία της ιδιωτικότητας μπορεί να επιτευχθεί και με άλλες τεχνικές που βασίζονται στην αλλοίωση-διαταραχή των δεδομένων. Οι τεχνικές διαφορικής ιδιωτικότητας (differential privacy) αποτελούν μια καλή άμυνα απέναντι σε επιθέσεις συμπερασμάτων συμμετοχής (membership inference attacks) με την προσθήκη θορύβου στα δεδομένα, στις επαναλήψεις κάποιου αλγορίθμου ή στην έξοδο του αλγορίθμου. Με αυτόν τον τρόπο οι στατιστικές ιδιότητες της βάσης δεδομένων παραμένουν ενώ μειώνεται η εξάρτηση μεταξύ του αποτελέσματος ενός ερωτήματος και των μεμονωμένων σημείων στη βάση δεδομένων, μειώνοντας με αυτόν τον τρόπο τη διαρροή πληροφοριών. Αυτή η τεχνική υποθέτει την ύπαρξη έμπιστης οντότητας υπολογισμών αλλά μπορεί και να εφαρμοστεί τοπικά (local differential privacy) από όλες τις οντότητες που κατέχουν τα δεδομένα, όταν δεν υπάρχει εμπιστοσύνη.

Πέραν της αποτροπής επιθέσεων συμπερασμάτων συμμετοχής (membership inference attacks) μέσω της διαφορικής ιδιωτικότητας μπορεί κάποιος να πετύχει τη προστασία των ιδιωτικών δεδομένων σε περιπτώσεις όπου υπάρχουν πολλαπλές οντότητες που συμμετέχουν. Επίσης, με τη χρήση της διαφορικής ιδιωτικότητας (differential privacy) είναι δυνατόν να αποτραπούν και επιθέσεις που στοχεύουν στην από-ανωνυμοποίηση των δεδομένων. Γενικά, όπως προκύπτει από τα παραπάνω, οι αλγόριθμοι διαφορικής ιδιωτικότητας είναι τυχαίοι και για αυτό μπορούν να κατηγοριοποιηθούν ανάλογα με το που και πως εφαρμόζεται η τυχαιότητα. Στην περίπτωση όπου θόρυβος προστίθεται στα δεδομένα εισαγωγής και μετά το σύνολο των υπολογισμών εφαρμόζεται σε αυτά τα δεδομένα, τότε η έξοδος είναι διαφορικά ιδιωτική (differential private) και ορίζεται ως διαταραχή εισόδου (input perturbation). Όταν η προσθήκη πραγματοποιείται σε ενδιάμεσες τιμές των δεδομένων και αυτές χρησιμοποιούνται σε επαναληπτικούς αλγορίθμους πρόκειται για διαταραχή αλγορίθμου (algorithm perturbation). Στις περισσότερες περιπτώσεις προτείνεται η προσθήκη θορύβου Gauss, σε κάθε επανάληψη του αλγορίθμου. [61] [62] Σε άλλες περιπτώσεις ο θόρυβος προστίθεται στο μοντέλο που δημιουργείται μετά την εκπαίδευση. Σε περιπτώσεις όπου η προσθήκη θορύβου μπορεί να προκαλέσει τη καταστροφή της τιμής εξόδου, μπορεί να χρησιμοποιηθεί ο εκθετικός μηχανισμός. [63] Όλες αυτές οι παραπάνω προσεγγίσεις της διαταραχής των δεδομένων λειτουργούν όταν αυτά φιλοξενούνται από έναν έμπιστο διακομιστή.

Όταν δεν είναι δυνατή η εκπαίδευση ενός μοντέλου από μια οντότητα και αυτή χρειάζεται να μεταφέρει τα δεδομένα της προς κάποια άλλη οντότητα για να πραγματοποιηθούν οι διεργασίες

υπολογισμών – εκπαίδευσης του μοντέλου, τότε πρέπει να χρησιμοποιηθεί η τεχνικής της τοπικής διαφορικής ιδιωτικότητας (local differential privacy). Έτσι κάθε οντότητα που κατέχει δεδομένα και θέλει αυτά να διαμοιραστούν σε άλλες οντότητες πρέπει πριν από τον διαμοιρασμό να προχωράει στην «διαταραχή» αυτών. Το κόστος και σε αυτή την τεχνική βέβαια είναι η μείωση της ακρίβειας του μοντέλου.

Άλλη μια τεχνική που μπορεί να χρησιμοποιηθεί είναι η μείωση διαστάσεων (dimensionality reduction). Με αυτή τη τεχνική επιτυγχάνεται η μείωση του πλήθους των χαρακτηριστικών των συνόλων δεδομένων και έτσι ενισχύεται το απόρρητο, καθώς με αυτόν τον τρόπο η ανάκτηση των αρχικών δεδομένων δεν είναι δυνατή. [64]

4.3 Προκλήσεις

Παρόλο που υπάρχουν διάφορα μέτρα και τεχνικές ώστε τα προσωπικά δεδομένα να προστατεύονται καθώς χρησιμοποιούνται στη μηχανική μάθηση, το γεγονός πως ακόμη χρησιμοποιούνται πολλοί αλγόριθμοι που δεν προσφέρουν κάποιου είδους ασφάλεια ως προς την ιδιωτικότητα αλλά και καθημερινά πολύ μεγάλο πλήθος προσωπικών δεδομένων ανεβαίνουν στο νέφος (cloud), καθιστούν την προστασία της ιδιωτικότητας μια μεγάλη πρόκληση. Ακόμη και με το νομοθετικό πλαίσιο του Γενικού Κανόνα Προστασίας Δεδομένων (GDPR, ευρωπαϊκή νομοθεσία), όπου οι οργανισμοί κλήθηκαν να φανούν πιο προστατευτικοί απέναντι στα δεδομένα και να δώσουν επιλογές στους πελάτες τους ως προς αυτά, συχνά οι επιλογές που παρέχουν εξαναγκάζουν τους πελάτες στον πλήρη διαμοιρασμό των δεδομένων τους.

Αντίστοιχα και οι προκλήσεις και ως προς την χρήση των τεχνικών είναι επίσης αρκετές. Ένα πρώτο ζήτημα είναι αυτό της απουσίας ευελιξίας. Καθώς η μηχανική μάθηση συνεχώς εξελίσσεται και καθώς νέες προτάσεις αλλαγών πραγματοποιούνται καθημερινά, οι περισσότερες τεχνικές για την προστασία της ιδιωτικότητας είναι συνδεδεμένες με συγκεκριμένο είδος αλγορίθμου μηχανικής μάθησης. Η συγκεκριμένη ιδιαιτερότητα έχει ως αποτέλεσμα να πρέπει να γίνουν αρκετές ενέργειες προκειμένου οι παραπάνω τεχνικές να μπορέσουν να ανταποκριθούν στα νέα δεδομένα της μηχανικής μάθησης. Ένα άλλο ζήτημα είναι αυτό της επεκτασιμότητας, όσον αφορά τα κόστη επεξεργασίας και επικοινωνίας, καθώς οι τεχνικές προστασίας της ιδιωτικότητας επιβάλλουν πρόσθετα κόστη ως προς αυτές τις μεταβλητές που περιορίζουν έτσι την ικανότητα χρήσης του τεράστιου πλήθους δεδομένων που υπάρχουν σήμερα διαθέσιμα.

4.4 Σύνοψη

Σε αυτό το κεφάλαιο αναφέρθηκαν τα διάφορα ζητήματα ιδιωτικότητας που προκύπτουν μέσω των επιθέσεων απέναντι σε μοντέλα μηχανικής μάθησης και παρουσιάστηκαν διάφορες τεχνικές άμυνας ώστε τα συστήματα μηχανικής μάθησης να γίνουν πιο εύρωστα απέναντι σε τέτοιου είδους επιθέσεις και με αυτόν τον τρόπο να μην προκύπτουν ζητήματα παραβίασης της ιδιωτικόητας.

5

Συμπεράσματα

Η μηχανική μάθηση έχει πλέον γίνει μέρος της καθημερινής μας ζωής και σίγουρα, όπως και σε κάθε νέα τεχνολογία, πολλά είναι τα θέματα που προκύπτουν όσον αφορά την ασφάλεια και την ιδιωτικότητα. Τα τελευταία χρόνια η έρευνα σε έναν πολύ μεγάλο βαθμό έχει στραφεί στα θέματα ασφάλειας και προστασίας της ιδιωτικότητας και στο πώς τα διάφορα μοντέλα μηχανικής μάθησης είτε στην φάση της εκπαίδευσης είτε στη φάση της λειτουργίας τους μπορούν να προστατευθούν. Γενικά, όπως και για οποιαδήποτε άλλη τεχνολογία, και τα δυο αυτά θέματα, ασφάλεια και απόρρητο, είναι πολύ κρίσιμα ζητήματα και δεν μπορούν να παραλειφθούν.

Στη συγκεκριμένη βιβλιογραφική επισκόπηση παρουσιάστηκαν διάφορες επιθέσεις κατά της μηχανικής μάθησης με ιδιαίτερη έμφαση στις επιθέσεις διαφυγής και τις επιθέσεις δηλητηρίασης δεδομένων. Δόθηκε και μια αρκετά σαφής εικόνα για το πώς τα ιδιαίτερα χαρακτηριστικά της βαθιάς μηχανικής μάθησης μπορούν να αποτελέσουν και την «αχίλλειο πτέρνα» των μοντέλων βαθιάς μηχανικής μάθησης. Παρουσιάστηκαν επίσης αντίμετρα αυτών των δύο τύπων επιθέσεων. Αναφορά έγινε επιπρόσθετα, τόσο σε παραδείγματα επιθέσεων από τον πραγματικό κόσμο όσο και στις διάφορες προκλήσεις που συναντώνται κατά τη προσπάθεια δημιουργίας πιο ασφαλών μοντέλων μηχανικής μάθησης για την αποφυγή διαφόρων τύπων επιθέσεων.

Ιδιαίτερη αναφορά έγινε επίσης και στο ζήτημα της ιδιωτικότητας. Η μηχανική μάθηση βασίζεται στα δεδομένα, επομένως αποτελεί ένα ιδιαιτέρως σημαντικό θέμα έρευνας. Συζητήθηκαν πιθανές απειλές απορρήτου σε προσωπικά δεδομένα και διάφορες τεχνικές άμυνας, οι περισσότερες εκ των οποίων βασίζονται σε κρυπτογραφικές τεχνικές.

Για μελλοντική εργασία, ιδιαιτέρως σημαντικό για τους ερευνητές είναι η έρευνα των διαφόρων κρυπτογραφικών λύσεων για τα βαθιά νευρωνικά δίκτυα ώστε να καταστεί δυνατή η αποτελεσματική προστασία της ιδιωτικότητας χωρίς ταυτόχρονα αυτό να σημαίνει μείωση της ακρίβειας του μοντέλου. Άλλος ένας ανοιχτός ερευνητικός χώρος είναι η προσαρμογή των διάφορων πρωτοκόλλων ασφάλειας και απορρήτου ώστε να λειτουργούν στα βαθιά νευρωνικά δίκτυα.

Βιβλιογραφία

1. Mohssen Mohammed, Muhammad Badruddin Khan, Eihab Bashier Mohammed Bashier. (2017), Machine Learning: Algorithms and Applications, CRC Press, pp. 7-12
2. Steven W. Knox. (2018), Machine Learning: A Concise Introduction, John Wiley & Sons, pp. 15, 33
3. Abhishek Vijayvargia. (2018), Machine Learning with Python: An Approach to Applied Machine Learning, BPB Publications
4. Pradeep Kumar, Arvind Tiwari. (2017), Ubiquitous Machine Learning and Its Applications, IGI Global, pp. 2
5. Nishant Shukla. (2018), Machine Learning with TensorFlow, Manning Publications Co., pp. 16-17
6. Max Pumperla, Kevin Ferguson. (2019), Deep Learning and the game of Go. Manning Publications Co., pp. 12-14
7. François Chollet, J. J. Allaire. (2018), Deep learning with R, Manning Publications Co., pp. 4, 8-11
8. John Paul Mueller, Luca Massaron. (2019), Deep Learning for Dummies, John Wiley & Sons, Inc., pp. 16-17
9. Ho Bae, Jaehee Jang, Dahuin Jung, Hyemi Jang, Heonseok Ha, Sungroh Yoon. (2018), Security and Privacy Issues in Deep Learning
10. Osvaldo Simeone. (2018), A Very Brief Introduction to Machine Learning With Applications to Communication Systems, Fellow, IEEE
11. https://blogs.nvidia.com/blog/2016/08/22/difference-deep-learning-training-inference-ai/
12. John D. Kelleher. (2019), Deep Learning, The MIT Press, pp. 1
13. Marco Barreno, Blaine Nelson, Anthony D. Joseph, J.D. Tygar. (2010), The security of machine learning, pp. 121-148
14. Katja Auernhammer, Ramin Tavakoli Kolagari, Markus Zoppelt. (2019), Attacks on Machine Learning: Lurking Danger for Accountability
15. Yevgeniy Vorobeychik ,Murat Kantarcioglu. (2018), Adversarial Machine Learning, Morgan & Claypool, pp. 19-24
16. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. (2013), Intriguing properties of neural networks
17. Xiaoyu Cao, Neil Zhenqiang Gong. (2019), Mitigating Evasion Attacks to Deep Neural Networks via Region-based Classification
18. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. (2014), Explaining and harnessing adversarial examples
19. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. (2016), Adversarial examples in the physical world
20. Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. (2016), The Limitations of Deep Learning in Adversarial Settings
21. Nicholas Carlini and David Wagner. (2017), Towards Evaluating the Robustness of Neural Networks
22. Mehdi Mirza and Simon Osindero. (2014), Conditional generative adversarial nets

23. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. (2016), DeepFool: a simple and accurate method to fool deep neural networks
24. Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 15–26. 2017. ACM.
25. Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks withlimited queries and information. In Jennifer Dy and Andreas Krause (eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 2137–2146, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018a. PMLR
26. Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, andShin-Ming Cheng. (2018), Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks
27. Wieland Brendel, Jonas Rauber, and Matthias Bethge. (2018), Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In International Conference on Learning Representations.
28. Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines, In 29th Int'l Conf. on Machine Learning, John Langfordand Joelle Pineau (Eds.).Int'l Conf. on Machine Learning (ICML), 1807–1814.
29. Marius Kloft and Pavel Laskov. 2012. Security Analysis of Online CentroidAnomaly Detection.Journal of Machine Learning Research 13 (2012), 3647–3690.
30. B. Nelson, M. Barreno, F.J. Chi, A.D. Joseph, B.I.P. Rubinstein, U. Saini, C.A.Sutton, J.D. Tygar, and K. Xia. 2008. Exploiting Machine Learning to Subvert your Spam Filter.LEET8 (2008), 1–9.
31. Luis Munoz-Gonzalez, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu, and Fabio Roli. (2017), Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization. In ACM Workshop on Artificial Intelligence and Security.
32. Pang Wei Koh and Percy Liang. (2017), Understanding Black-box Predictions via Influence Functions. In34th International Conference on Machine Learning.
33. Biggio, B., Corona, I., Maiorca, D., Nelson, B.,Srndic, N.,Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In Joint European conference on machine learning and knowledge discovery in databases, pp. 387–402. Springer, 2013.
34. Steinhardt, J., Koh, P. W., and Liang, P.. (2017), Certified defensesfor data poisoning attacks.
35. Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. (2017), Badnets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain.
36. Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang,and Xiangyu Zhang. (2018), Trojaning Attack on Neural Networks. In 25th Annual Network and Distributed System Security Symposium.
37. Practical Black-Box Attacks against Machine Learning
38. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. (2016), Rethinking the inception architecture for computer vision. In CVPR, pp. 2818–2826.

39. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. (2017), Towards deep learning models resistant to adversarial attacks.
40. Florian Tram`er, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, andPatrick McDaniel. (2019), Ensemble Adversarial Training: Attacks and Defenses. In International Conference on Learning Representations.
41. Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. (2017), Certified Defenses for Data Poisoning Attacks. In Advances in Neural Information Processing Systems.
42. Pang Wei Koh and Percy Liang. (2017), Understanding Black-box Predictions via Influence Functions. In 34th International Conference on Machine Learning.
43. Andrea Paudice, Luis Munoz-Gonzalez, and Emil C Lupu. (2018), Label Sanitization Against Label Flipping Poisoning Attacks. In Joint European Conference on MachineLearning and Knowledge Discovery in Databases.
44. Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards,Taesung Lee, Ian Molloy, and Biplav Srivastava. (2018), Detecting backdoor attacks on deep neural networks by activation clustering.
45. Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. (2017), Badnets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain.
46. Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. (2018), Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. In International Sym-posium on Research in Attacks, Intrusions, and Defenses.
47. Luis Munoz-Gonzalez, Bjarne Pfitzner, Matteo Russo, Javier Carnerero-Cano, and Emil C Lupu. (2019), Poisoning Attacks with Generative Adversarial Nets.
48. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. (2014), Generative Adversarial Nets.
49. Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen. (2017), Generative Poisoning Attack Method Against Neural Networks.
50. Chen Zhu, W. Ronny Huang, Ali Shafahi, Hengduo Li, Gavin Taylor, Christoph Studer, Tom Goldstein. (2019), Transferable Clean-Label Poisoning Attacks on Deep Neural Nets
51. Anish Athalye, Nicholas Carlini, David Wagner. (2018), Obfuscated Gradients Give a False Sense of Security:Circumventing Defenses to Adversarial Examples.
52. Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. (2015), Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures.
53. Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. (2017), Membership Inference Attacks Against Machine Learning Models.
54. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. (2016), Deep Residual Learning for Image Recognition.
55. Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. (2017), Densely Connected Convolutional Networks.
56. Bogdanov, Dan, Kamm, Liina, Laur, Sven, Pruulmann-Vengerfeldt, Pille, et al. (2014), Privacy-Preserving Statistical Data Analysis on Federated Databases.
57. Nikolaenko, Valeria, Weinsberg, Udi, Ioannidis, Stratis, Joye, Marc, et al. (2013), Privacy-preserving ridge regression on hundreds of millions of records.
58. Bost, Raphaël, Popa, Ra, Tu, Stephen and Goldwasser, S. (2014), Machine Learning Classification over Encrypted Data.
59. Bonawitz, Keith, Ivanov, Vladimir, Kreuter, Ben, Marcedone, Antonio, et al. (2017), Practical Secure Aggregation for Privacy Preserving Machine Learning.

60. Ohrimenko, Olga, Schuster, Felix, Fournet, Cedric, Mehta, Aastha, et al. (2016), Oblivious Multi-Party Machine Learning on Trusted Processors.
61. Abadi, Martin, Chu, Andy, Goodfellow, Ian, McMahan, H. Brendan, et al. (2016), Deep Learning with Differential Privacy.
62. Hardt, Moritz and Price, Eric. (2014), The Noisy Power Method : A Meta Algorithm with Applications.
63. Chaudhuri, Kamalika, Sarwate, Anand D and Sinha, Kaushik. (2013), A near-optimal algorithm for differentially-private principal components.
64. Liu, Kun, Kargupta, Hillol and Ryan, Jessica. (2006), Random projection-based multiplicative data perturbation for privacy preserving distributed data mining.
65. Deng, L., Hinton, G., Kingsbury, B. (2013), New types of deep neural network learning for speech recognition and related applications: An overview. In: Acoustics, Speech and Signal Processing (ICASSP).
66. Hochreiter, S., Schmidhuber, J. (1997), Long short-term memory. Neural computation 9 (8), 1735–1780.
67. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D. (1989), Backpropagation applied to handwritten zip code recognition. Neural computation 1 (4), 541–551.
68. P. Tabacof, T. Julia, E. Valle. (2016), Adversarial images for variational autoencoders.
69. Papernot N., McDaniel P., Swami A., and Harang R. (2016), Crafting adversarial input sequences for recurrent neural networks.
70. S. Hochreiter and J. Schmidhuber. (1997), Long short-term memory, Neural computation.
71. M. Volodymyr, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness,M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G.Ostrovski. (2015), Human-level control through deep reinforcement learning.
72. M. Volodymyr, A. P. Badia, and M. Mirza. (2016), Asynchronous methods for deep reinforcement learning.
73. S. Huang, N. Papernot, I. Goodfellow, Y. Duan, P. Abbeel. (2017), Adversarial Attacks on Neural Network Policies.
74. Szegedy Cea. (2013), Intriguing properties of neural networks.
75. Jiazhu Dai, Chuanshuai Chen. (2019), A backdoor attack against LSTM-based text classification systems.
76. S. Baluja, I. Fischer. (2017), Adversarial Transformation Networks: Learning to Generate Adversarial Examples.
77. S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard. (2017), Universal adversarial perturbations.
78. S. Sarkar, A. Bansal, U. Mahbub, and R. Chellappa. (2017), UPSET and ANGRI: Breaking High Performance Image Classifiers.
79. M. Cisse, Y. Adi, N. Neverova, and J. Keshet. (2017), Houdini: Fooling deep structured prediction models.
80. D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B.Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos. (2015), Deep speech 2: End-to-end speech recognition in English and Mandarin.
81. A. Rozsa, M. Geunther, E. M. Rudd, and T. E. Boult. (2017), Facial attributes: Accuracy and adversarial robustness.

82. V. Mirjalili, and A. Ross. (2017), Soft Biometric Privacy: Retaining Biometric Utility of Face Images while Perturbing Gender.
83. S. Shen, R. Furuta, T. Yamasaki, and K. Aizawa. (2017), Fooling Neural Networks in Face Attractiveness Evaluation: Adversarial Examples with High Attractiveness Score But Low Subjective Score.
84. I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash,A. Rahmati, D. Song. (2017), Robust Physical-World Attacks on Deep Learning Models.
85. A. Athalye, L. Engstrom, A. Ilyas, K. Kwok. (2017), Synthesizing Robust Adversarial Examples.
86. Zhang, Jun, et al. (2011), Evolutionary computation meets machine learning: A survey. IEEE Computational Intelligence Magazine 6.4 : 68-75
87. Eiben, Agoston E., and James E. Smith. (2003), Introduction to evolutionary computing. Vol. 53. Heidelberg: springer.
88. Mitchell, Melanie. (1998), An introduction to genetic algorithms. MIT press.
89. N. Papernot, P. D. McDaniel, and I. J. Goodfellow. (2016), Transferability in machine learning: from phenomena to black-box attacks using adversarial samples.
90. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. (2013), Intriguing properties of neural networks.
91. Y. Liu, X. Chen, C. Liu, and D. Song. (2017), Delving into transferable adversarial examples and black-box attacks.
92. https://money.cnn.com/galleries/2010/technology/1012/gallery.5_data_breaches/index.html
93. Maryam Archie, Sophie Gershon, Abigail Katcoff, and Aaron Zeng. (2018), Who's Watching? De-anonymization of Netflix Reviews using Amazon Reviews.
94. https://www.tensorflow.org/tutorials/generative/adversarial_fgsm
95. https://towardsdatascience.com/ml-algorithms-one-sd-%CF%83-74bcb28fafb6
96. https://user.eng.umd.edu/~danadach/Security_Fall_17/aml.pdf
97. Mohammad Al-Rubaie, J. Morris Chang. (2018), Privacy Preserving Machine Learning: Threats and Solutions.
98. Muhammad Imran Tariq, Nisar Ahmed Memon, Shakeel Ahmed, Shahzadi Tayyaba, Muhammad Tahir Mushtaq, Natash Ali Mian, Muhammad Imran, and Muhammad W. Ashraf. (2020), A Review of Deep Learning Security and PrivacyDefensive Techniques.
99. N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. (2016), Distillation as a defense to adversarial perturbations against deep neural networks, inSecurity and Privacy (SP), 2016 IEEE Symposium on. IEEE, pp. 582–597.
100. T. Genewein, V. Fischer, and B. Bischoff. (2017), On detecting adversarial perturbations, Proceedings of 5th International Conference on Learning Representations (ICLR).
101. S. Gu and L. Rigazio. (2015), Towards deep neural network architectures robust to adversarial examples, Proceedings of the International Conference on Learning Representations (ICLR).
102. D. Meng and H. Chen. (2017), Magnet: a two-pronged defense against adversarial examples, CCS.