



ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΕΣ»

ΚΑΤΕΥΘΥΝΣΗ ΕΞΕΙΔΙΚΕΥΣΗΣ «ΨΗΦΙΑΚΑ ΜΕΣΑ – ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ»

Αντιπαλική Αποταυτοποίηση Προσώπου

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΕΥΣΤΑΘΙΟΥ ΧΑΤΖΗΚΥΡΙΑΚΙΔΗ

Επιβλέπων

Ιωάννης Πήτας, Καθηγητής Α.Π.Θ.

ΕΡΓΑΣΤΗΡΙΟ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ ΚΑΙ ΑΝΑΛΥΣΗΣ ΠΛΗΡΟΦΟΡΙΩΝ

Θεσσαλονίκη, Φεβρουάριος 2019



ARISTOTLE UNIVERSITY OF THESSALONIKI

FACULTY OF SCIENCES

SCHOOL OF INFORMATICS

POSTGRADUATE STUDIES PROGRAM ON INFORMATICS AND COMMUNICATIONS

SPECIALIZATION ON DIGITAL MEDIA AND COMPUTATIONAL INTELLIGENCE

Adversarial Face De-identification

Master's Thesis

Efstathios Chatzikyriakidis

Supervisor

Ioannis Pitas, Professor AUTH

ARTIFICIAL INTELLIGENCE AND INFORMATION ANALYSIS LABORATORY

Thessaloniki, February 2019



Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Σχολή Θετικών Επιστημών
Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών «Πληροφορική και Επικοινωνίες»
Κατεύθυνση Εξειδίκευσης «Ψηφιακά Μέσα – Υπολογιστική Νοημοσύνη»
Εργαστήριο Τεχνητής Νοημοσύνης και Ανάλυσης Πληροφοριών

Αντιπαλική Αποταυτοποίηση Προσώπου

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΕΥΣΤΑΘΙΟΥ ΧΑΤΖΗΚΥΡΙΑΚΙΔΗ

Επιβλέπων

Ιωάννης Πήτας, Καθηγητής Α.Π.Θ.

Τριμελής Εξεταστική Επιτροπή

Ιωάννης Πήτας
Καθηγητής Α.Π.Θ.

Αναστάσιος Τέφας
Αν. Καθηγητής Α.Π.Θ.

Νικόλαος Νικολαΐδης
Αν. Καθηγητής Α.Π.Θ.

ΕΡΓΑΣΤΗΡΙΟ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ ΚΑΙ ΑΝΑΛΥΣΗΣ ΠΛΗΡΟΦΟΡΙΩΝ

Θεσσαλονίκη, Φεβρουάριος 2019

(Υπογραφή)

ΕΥΣΤΑΘΙΟΣ ΧΑΤΖΗΚΥΡΙΑΚΙΔΗΣ

Πτυχιούχος Μηχανικός Πληροφορικής, Τ.Ε.Ι. Κεντρικής Μακεδονίας

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Σχολή Θετικών Επιστημών

Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών «Πληροφορική και Επικοινωνίες»

Κατεύθυνση Εξειδίκευσης «Ψηφιακά Μέσα – Υπολογιστική Νοημοσύνη»

Εργαστήριο Τεχνητής Νοημοσύνης και Ανάλυσης Πληροφοριών



© 2019 Ευστάθιος Χατζηκυριακίδης

Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευτεί ότι εκφράζουν τις επίσημες θέσεις του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.

«Η εργασία αυτή αφιερώνεται με απέραντη ευγνωμοσύνη και αγάπη στους γονείς μου
Αναστάσιο και **Σοφία** καθώς και στην αγαπημένη μου **Δώρα** για την υποστήριξη και την
υπομονή τους κατά την διάρκεια των σπουδών μου»

Πρόλογος

Η παρούσα εργασία με τίτλο «Αντιπαλική Αποταυτοποίηση Προσώπου» αποτελεί την μεταπτυχιακή διπλωματική εργασία του Ευστάθιου Χατζηκυριακίδη για την κατεύθυνση «Ψηφιακά Μέσα – Υπολογιστική Νοημοσύνη» του προγράμματος μεταπτυχιακών σπουδών «Πληροφορική και Επικοινωνίες» του Τμήματος Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης. Η εργασία αυτή εκπονήθηκε στο Εργαστήριο Τεχνητής Νοημοσύνης και Ανάλυσης Πληροφοριών του Τμήματος Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέπων καθηγητή κύριο Ιωάννη Πήτα του Τμήματος Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης για την καθοδήγηση και την πολύτιμη βοήθειά του στην εκπόνηση της παρούσης εργασίας καθώς και τον υποψήφιο διδάκτορα κύριο Χρήστο Παπαϊωαννίδη του Τμήματος Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης για τις πολύτιμες συμβουλές του.

Περίληψη

Τα τελευταία χρόνια και ιδιαίτερα μετά την άνθηση των κοινωνικών δικτύων, οι άνθρωποι έχουν μάθει να ζουν σε έναν ψηφιακό κόσμο όπου ο διαμοιρασμός προσωπικών φωτογραφιών αποτελεί καθημερινή συνήθεια. Στις φωτογραφίες αυτές συχνά απεικονίζονται πρόσωπα ανθρώπων, είτε εν γνώσει, είτε εν αγνοία τους. Επιπλέον, σε πολλές εφαρμογές όπου χρησιμοποιούνται εικόνες προσώπου η γνώση της ταυτότητας του προσώπου δεν απαιτείται. Για αυτό το λόγο, είναι αναγκαία η προστασία της ιδιωτικότητας των ανθρώπων. Η Αποταυτοποίηση Προσώπου αποτελεί συνήθης λύση στο πρόβλημα της προστασίας της ιδιωτικότητας και σκοπό έχει να προστατέψει την ταυτότητα των εικονιζόμενων προσώπων από αυτόματα συστήματα αναγνώρισης προσώπου, ενώ παράλληλα να καθιστά δυνατή την αναγνώριση από ανθρώπινους παρατηρητές. Παρόλο που υπάρχουν ήδη μέθοδοι αποταυτοποίησης προσώπου που παρέχουν ικανοποιητική προστασία, οι παραγόμενες αποταυτοποιημένες εικόνες αυτών δεν μοιάζουν αρκετά με τις αρχικές. Για την επίλυση αυτού του προβλήματος, σε αυτήν την μεταπτυχιακή διπλωματική εργασία, προτείνεται η χρήση των Αντιπαλικών Δειγμάτων. Ειδικότερα, προτείνεται μια νέα μέθοδος αντιπαλικής επίθεσης που είναι ικανή να αλλοιώσει στο ελάχιστο την ποιότητα της εικόνας και να παρέχει προστασία από αυτόματα συστήματα αναγνώρισης προσώπου. Από τα αποτελέσματα της συγκριτικής ανάλυσης μεταξύ της προτεινόμενης και άλλων μεθόδων αντιπαλικής επίθεσης φαίνεται ξεκάθαρα πως η προτεινόμενη μέθοδος προστατεύει καθώς και διατηρεί την οπτική ποιότητα των αρχικών εικόνων προσώπου πιο αποτελεσματικά.

Λέξεις Κλειδιά: Προστασία Ιδιωτικότητας, Αποταυτοποίηση Προσώπου, Αντιπαλικά Δείγματα, Αναγνώριση Προσώπου, Βαθιά Μάθηση, Μηχανική Μάθηση, Όραση Υπολογιστών, Βαθιά Τεχνητά Νευρωνικά Δίκτυα

Abstract

In recent years, and especially after the booming of social networks, people have learned to live in a digital world where sharing of personal photos is an everyday habit. These photos often portray people's faces, either knowingly or unknowingly. In addition, in many applications where facial images are used, knowledge of the depicted person's identity is not required. For this reason, it is necessary to protect the privacy of people. Face De-identification is a common solution to the problem of privacy protection and aims to protect the identity of the depicted persons by automatic face recognition systems, while making it possible for human observers to recognize it. Although face de-identification methods already exist that provide satisfactory protection, the generated de-identified images do not resemble the original ones. To solve this problem, in this Master's thesis, the usage of Adversarial Examples is proposed. Particularly, a novel adversarial attack method is proposed that is capable of minimizing the image quality distortion and providing protection against automatic face recognition systems. From the results of the comparative analysis between the proposed and other adversarial attack methods it clearly appears that the proposed method protects as well as preserves the visual quality of the original facial images more efficiently.

Keywords: Privacy Protection, Face De-identification, Adversarial Examples, Face Recognition, Deep Learning, Machine Learning, Computer Vision, Deep Artificial Neural Networks

Πίνακας Περιεχομένων

Πρόλογος.....	1
Ευχαριστίες.....	3
Περίληψη.....	5
Abstract	7
Πίνακας Περιεχομένων	9
Κατάλογος Εικόνων	12
Κατάλογος Πινάκων	13
Κεφάλαιο 1: Εισαγωγή	14
Κεφάλαιο 2: Ιδιωτικότητα	16
2.1: Περιγραφή Ιδιωτικότητας	16
2.2: Κατηγορίες Ιδιωτικότητας.....	16
2.2.1: Προσωπική Ιδιωτικότητα	16
2.2.2: Πληροφοριακή Ιδιωτικότητα	17
2.2.2.1: Ιατρικές Πληροφορίες	17
2.2.2.2: Διαδικτυακές Πληροφορίες	17
2.2.2.3: Χωρικές Πληροφορίες	18
2.2.2.4: Οικονομικές Πληροφορίες	18
2.2.2.5: Πολιτικές Πληροφορίες.....	18
2.2.3: Οργανωτική Ιδιωτικότητα	18
2.2.4: Πνευματική Ιδιωτικότητα.....	19
2.3: Σταθμοί Ιδιωτικότητας	19
2.4: Προστασία Ιδιωτικότητας	20
Κεφάλαιο 3: Αποταυτοποίηση Προσώπου	22
3.1: Εισαγωγή	22
3.2: Περιγραφή Προβλήματος	23
3.3: Προϋπάρχουσες Μέθοδοι	23
3.3.1: Αφελείς Μέθοδοι	23
3.3.1.1: Μάσκα	24
3.3.1.2: Πιξελοποίηση	24
3.3.1.3: Θόλωση	25
3.3.1.4: Αρνητικό	25
3.3.1.5: Θόρυβος	26
3.3.1.6: Κατωφλίωση.....	26

3.3.1.7: Πρόσωπο «κ. Πατάτας»	26
3.3.1.8: Μείωση.....	26
3.3.2: Μέθοδοι Μοντέλου k-ανωνυμίας.....	27
3.3.2.1: k-ιδίων	28
3.3.2.2: k-ιδίων-επιλογή.....	29
3.3.2.3: k-ιδίων-M	29
3.3.2.4: k-ιδίων-δίκτυο	29
3.3.3: Σύγχρονοι Μέθοδοι.....	30
3.3.3.1: Πολυπαραγοντική Αποταυτοποίηση Προσώπου	31
3.3.3.2: Αποταυτοποίηση Προσώπου με δωρητές χαρακτηριστικά προσώπου.....	32
3.3.3.3: Αποταυτοποίηση Προσώπου με Γεννητικά Βαθιά Νευρωνικά Δίκτυα	32
3.4: Προσέγγιση Προτεινόμενης Μεθόδου	33
Κεφάλαιο 4: Αντιπαλικά Δείγματα.....	34
4.1: Εισαγωγή	34
4.2: Ορισμός	34
4.3: Γιατί Υπάρχουν;	35
4.4: Συνέπειες.....	37
4.5: Μεταφερσιμότητα	38
4.6: Ταξινομία.....	38
4.6.1: Δείγμα Εισόδου	38
4.6.2: Συχνότητα Επίθεσης.....	39
4.6.3: Ειδίκευση Αντιπαλικού Δείγματος.....	39
4.6.4: Δράση Αντιπαλικής Διαταραχής	40
4.6.5: Μέγεθος Αντιπαλικής Διαταραχής.....	41
4.6.6: Είδος Αντιπαλικής Διαστρέβλωσης.....	42
4.6.7: Γνώση Αντίπαλου Μοντέλου.....	42
4.7: Προϋπάρχουσες Μέθοδοι	42
4.7.1: Μη-επαναληπτικές Μέθοδοι	43
4.7.1.1: Fast Gradient Sign Method (FGSM).....	43
4.7.1.2: One-step Target Class Method (OTCM)	43
4.7.1.3: One-step Least Likely Class Method (OLLCM)	44
4.7.1.4: Fast Gradient Value Method (FGVM).....	44
4.7.1.5: Adversarial GAN Attack (AdvGAN)	45
4.7.2: Επαναληπτικές Μέθοδοι.....	46

4.7.2.1: Box-constrained Limited-memory BFGS Attack (L-BFGS-B)	46
4.7.2.2: Basic Iterative Method (BIM)	46
4.7.2.3: Iterative Target Class Method (ITCM)	47
4.7.2.4: Iterative Least Likely Class Method (ILLCM).....	47
4.7.2.5: Iterative Fast Gradient Value Method (IFGVM)	48
4.7.2.6: Momentum Iterative Fast Gradient Sign Method (MIFGSM)	48
4.7.2.7: DeepFool Attack	49
4.7.2.8: Jacobian-Based Saliency Map Attack (JSMA)	50
4.7.2.9: Universal Perturbation Attack.....	50
4.7.2.10: Carlini-Wagner Attack (C&W).....	51
4.7.2.11: One Pixel Attack	51
Κεφάλαιο 5: Προτεινόμενη Μέθοδος.....	52
5.1: Εισαγωγή	52
5.2: Περιγραφή Μεθόδου	52
5.3: Πρωτότυπα Στοιχεία	53
Κεφάλαιο 6: Πειραματική Διαδικασία	54
6.1: Συγκριτική Ανάλυση	54
6.2: Σύνολα Δεδομένων	54
6.2.1: CelebFaces Attributes Dataset (CelebA).....	54
6.2.2: VGG Face Dataset	55
6.3: Προεπεξεργασία Δεδομένων	55
6.4: Μοντέλα Αναγνώρισης Προσώπου	55
6.4.1: Μοντέλο A	55
6.4.2: Μοντέλο B	58
6.5: Τιμές Παραμέτρων	60
6.6: Βοηθητικό Υλικό	61
Κεφάλαιο 7: Αποτελέσματα Κ Συμπεράσματα	62
7.1: Αποτελέσματα Συγκριτικής Ανάλυσης.....	62
7.2: Παραδείγματα Αποταυτοποιημένων Εικόνων.....	62
7.3: Σχόλια Πειραματικών Αποτελεσμάτων	64
Επίλογος	66
Βιβλιογραφία	67

Κατάλογος Εικόνων

Εικόνα 1: Σημαντικά προβλήματα που προσπαθεί να επιλύσει η όραση υπολογιστών	22
Εικόνα 2: Παράδειγμα ανίχνευσης και αναγνώρισης προσώπων σε μια εικόνα.....	22
Εικόνα 3: Παραδείγματα αποταυτοποιημένων εικόνων με τις αφελείς μεθόδους.....	24
Εικόνα 4: Το γενικό μοντέλο προστασίας της k-ανωνυμίας.....	27
Εικόνα 5: Παραδείγματα αποταυτοποιημένων εικόνων με μεθόδους τύπου «k-ιδίων»	28
Εικόνα 6: Παραδείγματα σύνθεσης καινούργιων προσώπων με γεννητικά μοντέλα	30
Εικόνα 7: Παράδειγμα σύνθεσης καινούργιου προσώπου με δωρητές χαρακτηριστικών ...	31
Εικόνα 8: Παράδειγμα αντιπαλικής εικόνας που κατηγοριοποιείται λανθασμένα	35
Εικόνα 9: Διάφορες εκδοχές προσαρμογής ενός μοντέλου σε δεδομένα εκπαίδευσης	36
Εικόνα 10 : Παραδείγματα συναρτήσεων ενεργοποίησης που έχουν γραμμικότητες	37
Εικόνα 11: Αντιπαλικό δείγμα στον φυσικό κόσμο που κατηγοριοποιείται λανθασμένα.....	38
Εικόνα 12: Παραδείγματα τυχαίου και μη-τυχαίου δείγματος εισόδου.....	39
Εικόνα 13: Παραδείγματα τοπικών αντιπαλικών διαταραχών.....	40
Εικόνα 14: Παράδειγμα καθολικής αντιπαλικής διαταραχής.....	41

Κατάλογος Πινάκων

Πίνακας 1: Οι τιμές των παραμέτρων της πρώτης φάσης αναζήτησης του Μοντέλου A	56
Πίνακας 2: Οι τιμές των παραμέτρων της δεύτερης φάσης αναζήτησης του Μοντέλου A ...	56
Πίνακας 3: Ο καλύτερος συνδυασμός τιμών παραμέτρων για το Μοντέλου A	56
Πίνακας 4: Η αρχιτεκτονική του Μοντέλου A	57
Πίνακας 5: Η συνοπτική αναπαράσταση του Μοντέλου A σε Keras	57
Πίνακας 6: Οι πληροφορίες εκπαίδευσης του Μοντέλου A	57
Πίνακας 7: Η αρχιτεκτονική του Μοντέλου B	59
Πίνακας 8: Η συνοπτική αναπαράσταση του Μοντέλου B σε Keras	59
Πίνακας 9: Οι πληροφορίες εκπαίδευσης του Μοντέλου B	59
Πίνακας 10: Οι τιμές παραμέτρων για τις μεθόδους της συγκριτικής ανάλυσης	60
Πίνακας 11: Οι βιβλιοθήκες Python που χρησιμοποιήσαμε στα πειράματα μας.....	61
Πίνακας 12: Τα αποτελέσματα της αξιολόγησης των μεθόδων της συγκριτικής ανάλυσης .	62
Πίνακας 13: Τα ποσοστά βελτίωσης της μεθόδου P-FGVM στις μετρικές αξιολόγησης	62
Πίνακας 14: Παραδείγματα αποταυτοποιημένων εικόνων με την μέθοδο P-FGVM.....	63
Πίνακας 15: Παραδείγματα αποταυτοποιημένων εικόνων με την μέθοδο IFGVM	63
Πίνακας 16: Παραδείγματα αποταυτοποιημένων εικόνων με την μέθοδο ITCM.....	64
Πίνακας 17: Χρήση της μεθόδου P-FGVM με τυχαίο Γκαουσιανό θόρυβο ως είσοδο	64

Κεφάλαιο 1: Εισαγωγή

Με την ολοένα και περισσότερη χρήση του διαδικτύου αυξάνεται συνεχώς η αποθήκευση και ο διαμοιρασμός προσωπικών φωτογραφιών στις οποίες εμφανίζονται πρόσωπα ανθρώπων. Παράλληλα, τα τελευταία χρόνια η όραση υπολογιστών με την βοήθεια της βαθιάς μάθησης και των βαθιών νευρωνικών δικτύων έχει σημειώσει αρκετή πρόοδο. Ένα από τα προβλήματα που προσπαθεί να επιλύσει είναι η αναγνώριση εικόνας και μια από τις πιο συνηθισμένες εφαρμογές αυτής αποτελεί η αναγνώριση προσώπου. Ο σκοπός αυτής είναι η αυτόματη αναγνώριση της ταυτότητας διαφόρων ανθρώπων από εικόνες στις οποίες εμφανίζονται τα πρόσωπα τους. Ωστόσο, αυτό οδηγεί σε αρκετά προβλήματα σχετικά με την προστασία της προσωπικής ιδιωτικότητας και των προσωπικών δεδομένων.

Ένας από τους πιο αποδοτικούς τρόπους με τους οποίους μπορεί δοθεί λύση σε αυτά τα προβλήματα είναι μέσω της αποταυτοποίησης προσώπου. Η αποταυτοποίηση προσώπου προσπαθεί να προστατέψει την ταυτότητα των ατόμων που εμφανίζονται μέσα σε εικόνες. Πρόκειται για την διαδικασία κατά την οποία μια εικόνα αλλοιώνεται σκόπιμα με σκοπό να αποταυτοποιηθεί έτσι ώστε να μην είναι πλέον εφικτή η αναγνώριση της ταυτότητας του εικονιζόμενου ατόμου από τα διάφορα αυτόματα συστήματα αναγνώρισης προσώπου.

Κατά καιρούς αναπτύχθηκαν διάφορες μέθοδοι αποταυτοποίησης προσώπου. Οι μέθοδοι αυτοί παραμορφώνονται έντονα το πρόσωπο στην εικόνα είναι ικανές να μπερδέψουν με μεγάλη επιτυχία απλοϊκά συστήματα αναγνώρισης προσώπου. Ωστόσο, το πρόβλημα που προκύπτει με αυτές τις μεθόδους είναι ότι ενώ τις περισσότερες φορές ένας ανθρώπινος παρατηρητής δεν μπορεί να αναγνωρίσει την ταυτότητα του προσώπου μιας αποταυτοποιημένης εικόνας, ένα σύγχρονο σύστημα αναγνώρισης προσώπου είναι σε θέση να το κάνει. Επιπλέον, σε εφαρμογές όπου θέλουμε οι αποταυτοποιημένες εικόνες να είναι όσο το δυνατόν ρεαλιστικές και όμοιες με τις αρχικές, αυτές οι μέθοδοι δεν μπορούν να θεωρηθούν αποδεκτές διότι αλλοιώνουν έντονα την εμφάνιση του προσώπου. Ακόμη, αδυνατούν να διατηρήσουν ικανοποιητικά στο αποταυτοποιημένο πρόσωπο τα μηταυτοτικά χαρακτηριστικά του αρχικού, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα, με αποτέλεσμα να είναι εξίσου μη αποδεκτές σε εφαρμογές που τα χρειάζονται.

Σε αυτήν την εργασία προτείνουμε μία νέα μέθοδο αποταυτοποίησης προσώπου η οποία αξιοποιεί τα αντιπαλικά δείγματα και επιλύει τα παραπάνω προβλήματα. Η μέθοδος αυτή λειτουργεί στον χώρο της εικόνας και μπορεί να αποταυτοποιήσει μία εικόνα προσώπου αλλοιώνοντας την ελάχιστα με τέτοιο τρόπο ώστε η αποταυτοποιημένη εικόνα να είναι ανεπαίσθητα διαφορετική από την αρχική αλλά και να μην μπορεί να αναγνωριστεί σωστά από σύγχρονα αυτόματα συστήματα αναγνώρισης προσώπου. Στην πραγματικότητα αποτελεί μέθοδος αντιπαλικής επίθεσης και κατασκευάζει τις αποταυτοποιημένες εικόνες ως στοχευμένα αντιπαλικά δείγματα.

Επίσης, διεξήγαμε μια πειραματική συγκριτική ανάλυση μεταξύ της προτεινόμενης καθώς και άλλες μεθόδους αντιπαλικής επίθεσης με σκοπό να αποταυτοποιήσουμε εικόνες προσώπου από το σύνολο δεδομένων CelebA. Συγκεκριμένα, με τις μεθόδους αυτές στοχεύσαμε στο να κατασκευάσουμε ρεαλιστικές αποταυτοποιημένες εικόνες προσώπου

με μεγάλο ποσοστό εσφαλμένης κατηγοριοποίησης. Οι μετρικές αξιολόγησης που υπολογίσαμε στα πειράματα μας για την συγκριτική ανάλυση των μεθόδων είναι ο δείκτης ομοιότητας CW-SSIM μεταξύ των αποταυτοποιημένων και των αρχικών εικόνων προσώπου, η l_2 -νόρμα της αντιπαλικής διαταραχής που προστίθεται στην αρχική εικόνα καθώς και το ποσοστό εσφαλμένης κατηγοριοποίησης των αποταυτοποιημένων εικόνων προσώπου.

Από τα αποτελέσματα της πειραματικής συγκριτικής ανάλυσης φαίνεται ξεκάθαρα πως η προτεινόμενη μέθοδος είναι πολύ καλύτερη από τις υπόλοιπες μεθόδους διότι οι αποταυτοποιημένες εικόνες που παράγει μοιάζουν περισσότερο με τις αρχικές και έχουν μεγαλύτερη πιθανότητα να κατηγοριοποιηθούν εσφαλμένα.

Κεφάλαιο 2: Ιδιωτικότητα

Σε αυτό το κεφάλαιο θα αναλύσουμε και θα περιγράψουμε την ιδιωτικότητα και τις διάφορες κατηγορίες στις οποίες αυτή χωρίζεται. Επιπλέον, θα αναφερθούμε σε κάποιες χρονικές στιγμές που αποτέλεσαν σταθμοί για την εξέλιξη της καθώς και στους σύγχρονους τρόπους με τους οποίους αυτή προστατεύεται κυβερνητικά.

2.1: Περιγραφή Ιδιωτικότητας

Από τα αρχαία χρόνια μέχρι και σήμερα, όλα τα έμβια όντα έχουν μια ισχυρή επιθυμία στο να προστατεύουν τον εαυτό τους από τους διάφορους κινδύνους. Το μέγεθος αυτής της επιθυμίας μπορεί να διαφέρει σε κάθε όν και η ύπαρξη της μπορεί να οφείλεται τόσο σε κοινωνικούς (π.χ. μιμίδια) όσο και σε γενετικούς (π.χ. γονίδια) παράγοντες. Για παράδειγμα, οι άνθρωποι ένα από τα πράγματα που θέλουν να προστατεύουν είναι η ιδιωτικότητα τους.

Η ιδιωτικότητα αποτελεί έννοια που περιγράφει το δικαίωμα που έχει κάθε άνθρωπος ή κοινωνική ομάδα να έχει ιδιωτική ζωή. Βασική προϋπόθεση για να έχει ένας άνθρωπος ή μια κοινωνική ομάδα ιδιωτική ζωή είναι να μην παρακολουθείται μυστικά. Επίσης, θα πρέπει όλοι μας να έχουμε το δικαίωμα στο να καθορίζουμε εάν, πότε, πως και σε ποιόν θα γνωστοποιούμε προσωπικά δεδομένα. Στον σύγχρονο κόσμο, η επιθυμία που έχουν οι άνθρωποι για ιδιωτική ζωή απορρέει από την ανάγκη τους για ελεύθερη ανάπτυξη της προσωπικότητας τους.

2.2: Κατηγορίες Ιδιωτικότητας

2.2.1: Προσωπική Ιδιωτικότητα

Πολλοί άνθρωποι νοιώθουν πως δεν είναι σωστό να εκθέτουν το σώμα τους γυμνό μπροστά σε άλλους. Συχνά συνδέεται αυτό το συναίσθημα με την σεμνότητα και τον σεβασμό. Ωστόσο, κάτω από εξειδικευμένες περιπτώσεις (π.χ. ιατρικούς λόγους) κάπι τέτοιο για τους ίδιους ανθρώπους μπορεί να μην είναι λάθος. Για την προστασία του γυμνού σώματος από την κοινή θέα χρησιμοποιείται κατά βάση ο ρουχισμός. Επίσης, στις περισσότερες κοινωνίες οι άνθρωποι επιθυμούν να ζουν σε ιδιωτικό χώρο είτε μόνοι τους είτε με οικεία πρόσωπα. Ο χώρος αυτός συνήθως προστατεύεται από την κοινή θέα και δεν είναι ορατός στο εσωτερικό του. Με αυτό τον τρόπο προστατεύονται οι πράξεις καθώς και οι αλληλεπιδράσεις τους χωρίς αυτές να είναι αντιληπτές από άλλους.

Η προσωπική ιδιωτικότητα συνδέεται στενά και με την φυσική ιδιωτικότητα η οποία δεν επιτρέπει χωρίς απαραίτητη άδεια την είσοδο κάποιου σε προσωπικό χώρο άλλου (π.χ. σπίτι, αυτοκίνητο) ή την έρευνα προσωπικών αντικειμένων και δεδομένων (π.χ. ταυτότητα, ιατρικό απόρρητο). Η προσωπική ιδιωτικότητα προστατεύεται νομικά. Ένα παράδειγμα νομικής προστασίας της προσωπικής ιδιωτικότητας αποτελεί το «U.S. Fourth Amendment» το οποίο δηλώνει ξεκάθαρα σε ελεύθερη μετάφραση ότι «Κάθε άνθρωπος έχει το δικαίωμα να προστατεύει τον εαυτό του, το σπίτι του και τα προσωπικά του αντικείμενα ενάντια σε αναιτιολόγητες έρευνες και κατάσχεση».

2.2.2: Πληροφοριακή Ιδιωτικότητα

Η πληροφοριακή ιδιωτικότητα αναφέρεται στην σχέση που αναπτύσσεται ανάμεσα στους ανθρώπους και την τεχνολογία και ορίζει την νόμιμη και αποδεκτή χρήση αυτής σχετικά με την αποθήκευση και τον διαμοιρασμό πληροφοριών. Κάθε άνθρωπος έχει δικαίωμα στην χρήση της τεχνολογίας. Καθημερινά, εκατομμύρια άνθρωποι παράγουν, αποθηκεύουν και διαμοιράζονται πληροφορίες. Η πληροφορία είναι πολύ σημαντική στις μέρες μας και πολλές φορές αποτελεί τον τρόπο με τον οποίο εκφραζόμαστε και ανταλλάσσουμε ιδέες και συναισθήματα. Ωστόσο, προκύπτουν ανησυχίες όταν υπάρχουν προσωπικά δεδομένα με τα οποία μπορεί να αναγνωριστεί μοναδικά κάποιο πρόσωπο ή να παρθούν αποφάσεις για αυτό χωρίς να το γνωρίζει. Γενικότερα, υπάρχει το δίλημμα ποια άτομα έχουν το δικαίωμα ιδιοκτησίας αυτών των προσωπικών δεδομένων πέρα από το ίδιο το πρόσωπο με το οποίο αυτά σχετίζονται (π.χ. περιπτώσεις ανθρώπων που λόγω ασθένειας ή δικαστικής εντολής δεν μπορούν να είναι οι ιδιοκτήτες των δεδομένων αυτών και αντιπροσωπεύονται από άλλο άτομο). Το άτομο που έχει στην κατοχή του προσωπικά δεδομένα κάποιου άλλου μπορεί να πάρει αποφάσεις που να έχουν αρνητική επίδραση.

Αρκετές χώρες παγκοσμίως εφαρμόζουν τον κανονισμό «General Data Protection Regulation» (GDPR) του Ευρωπαϊκού Κοινοβουλίου και Συμβουλίου της Ευρωπαϊκής Ένωσης. Με αυτόν τον κανονισμό έχει γίνει ένα σπουδαίο βήμα προς την προστασία της προσωπικής ιδιωτικότητας και ιδιαίτερα των προσωπικών δεδομένων. Σύμφωνα με την Ευρωπαϊκή Επιτροπή, προσωπικά δεδομένα είναι όλα όσα αφορούν την ιδιωτική, επαγγελματική ή δημόσια ζωή ενός ατόμου. Μπορεί να είναι οτιδήποτε από όνομα, διεύθυνση κατοικίας, φωτογραφία, τραπεζικές λεπτομέρειες, αναρτήσεις στις ιστοσελίδες κοινωνικής δικτύωσης ή ιατρικά δεδομένα. Επίσης, στις Ηνωμένες Πολιτείες της Αμερικής για την προστασία των ιατρικών προσωπικών δεδομένων υπάρχουν οι νομοθεσίες «Health Insurance Portability and Accountability Act» (HIPAA) και «Health Information Technology for Economic and Clinical Health Act» (HITECH Act).

Ανάλογα με το είδος του προσωπικού δεδομένου η πληροφοριακή ιδιωτικότητα αναλύεται περαιτέρω σε πέντε υποκατηγορίες: Ιατρικές Πληροφορίες, Διαδικτυακές Πληροφορίες, Χωρικές Πληροφορίες, Οικονομικές Πληροφορίες και Πολιτικές Πληροφορίες.

2.2.2.1: Ιατρικές Πληροφορίες

Η πλειοψηφία των ανθρώπων επιθυμούν να υπάρχει ιατρικό απόρρητο. Δηλαδή, δεν θέλουν να κοινοποιούνται χωρίς την έγκριση τους τα ιατρικά τους προσωπικά δεδομένα (π.χ. προβλήματα ψυχικής και σωματικής υγείας, επισκέψεις σε νοσοκομεία, εξετάσεις, συνταγογραφήσεις). Το ιατρικό απόρρητο αποτελεί «ιερό» θεμέλιο της σχέσης μεταξύ ιατρού και ασθενούς. Με αυτό τον τρόπο προστατεύεται η αξιοπρέπεια των ασθενών αλλά και τους δίδεται η ελευθερία να αποκαλύψουν πλήρεις και ακριβείς πληροφορίες στον ιατρό τους.

2.2.2.2: Διαδικτυακές Πληροφορίες

Είναι πολύ εύκολο κάποιος να διαμοιράσει προσωπικά δεδομένα στο διαδίκτυο τα οποία να καταλήξουν στα «χέρια» κακόβουλων χρηστών. Δεν είναι λίγες οι φορές που κακόβουλοι

χρήστες εκμεταλλεύονται τα προσωπικά δεδομένα άλλων για κακό σκοπό. Σε αυτήν την κατηγορία αναφέρονται δεδομένα όπως μηνύματα ηλεκτρονικού ταχυδρομείου, ιστορικά αναζητήσεων, σχόλια που έχουν κάνει οι χρήστες σε πολυμέσα, αξιολογήσεις προϊόντων καθώς και οι ιστοσελίδες που επισκέπτεται ένας χρήστης. Πρόβλημα μπορούν να αποτελέσουν και τα ενδιάμεσα δεδομένα που αποθηκεύουν και διαμοιράζονται με τρίτους οι ιστοσελίδες από τους χρήστες τους. Γενικά, δεν θα πρέπει να είναι δυνατή η αναγνώριση συγκεκριμένων ανθρώπων από αυτού τους είδους τα δεδομένα.

2.2.2.3: Χωρικές Πληροφορίες

Είναι πλέον γεγονός πως καθημερινά συλλέγονται ευαίσθητα δεδομένα που σχετίζονται με την τοποθεσία των χρηστών. Αυτά τα δεδομένα μπορεί να λαμβάνονται μέσω των κινητών τηλεφώνων από διάφορες εφαρμογές. Από αυτά τα δεδομένα είναι δυνατό να εξαχθούν συμπεράσματα για ευαίσθητα επαγγελματικά και προσωπικά θέματα, όπως η συμμετοχή σε λατρευτικές συνάξεις, παραμονή σε ξενοδοχεία και νοσοκομεία. Επίσης, τα δεδομένα αυτά μπορούν να αξιοποιούνται από διάφορες εταιρίες αφού πρώτα έχει γίνει η συγκατάθεση του χρήστη. Γενικά, δεν θα πρέπει να είναι δυνατή η αναγνώριση συγκεκριμένων ανθρώπων από αυτού τους είδους τα δεδομένα.

2.2.2.4: Οικονομικές Πληροφορίες

Ευαίσθητα δεδομένα όπως μετοχές, το πλήθος των περιουσιακών στοιχείων, συναλλαγές, λογαριασμοί, πιστωτικές κάρτες και αγορές μπορούν να αποκαλύψουν είτε την οικονομική κατάσταση είτε τις συνήθειες ενός ανθρώπου (π.χ. μέρη που έχει επισκεφθεί, προϊόντα που χρησιμοποιεί, φάρμακα που χρειάζεται). Επειδή αυτά τα δεδομένα μπορούν να αποτελέσουν το έναυσμα για οικονομικές απάτες, δεν θα πρέπει να αποκαλύπτονται παρά μόνο όταν το ίδιο το άτομο το επιθυμεί. Επίσης, πολλές φορές τα δεδομένα αυτά μπορούν κάλλιστα να αξιοποιηθούν από εταιρίες με σκοπό την βελτίωση των διαφημίσεων, την προώθηση νέων προϊόντων ή την αύξηση των κερδών. Σε κάθε περίπτωση, τα δεδομένα αυτά θα πρέπει να αξιοποιούνται μόνο μετά από την συγκατάθεση του ίδιου του πελάτη. Γενικά, δεν θα πρέπει να είναι δυνατή η αναγνώριση συγκεκριμένων ανθρώπων από αυτού τους είδους τα δεδομένα.

2.2.2.5: Πολιτικές Πληροφορίες

Η ιδιωτικότητα των πολιτικών δεδομένων είναι προστατευμένη από τα αρχαία χρόνια, από τότε που εμφανίστηκε ο θεσμός της ψηφοφορίας. Οι άνθρωποι μπορούν να ψηφίσουν σε μία κάλπη μυστικά και να επιλέξουν την προτίμηση τους για κάποιο θέμα χωρίς να είναι απαραίτητη η αποκάλυψη της επιλογής τους. Με αυτό τον τρόπο είναι βέβαιο ότι οι πολιτικές απόψεις του καθενός είναι κτήμα μόνο του ίδιου και κανενός άλλου. Όλα τα ανεπτυγμένα και δημοκρατικά κράτη εφαρμόζουν αυτήν την μέθοδο και είναι βασικό δικαίωμα όλων των πολιτών να αποκρύπτουν τις πολιτικές τους επιλογές.

2.2.3: Οργανωτική Ιδιωτικότητα

Για λόγους ανταγωνισμού ή προστασίας διάφορες εταιρίες, ομάδες, κυβερνητικές υπηρεσίες, κοινωνίες ή οργανώσεις μπορεί να επιθυμούν να διατηρήσουν κρυφές

πληροφορίες για την οργάνωση και την δραστηριότητά τους από άλλα πρόσωπα ή οργανώσεις.

2.2.4: Πνευματική Ιδιωτικότητα

Ενώ η φυσική ιδιωτικότητα αναφέρεται στην προστασία των υλικών προσωπικών πραγμάτων, η πνευματική ιδιωτικότητα αναφέρεται στην προστασία της πνευματικής φύσης του ανθρώπου και πιο συγκεκριμένα στα συναισθήματα και τις σκέψεις του. Αρχικά, νομοθεσίες όπως η «British Common Law» προστάτευαν μόνο την φυσική ιδιωτικότητα. Ωστόσο, στην πορεία επεκτάθηκαν ώστε να συμπεριλάβουν και την πνευματική ιδιοκτησία.

2.3: Σταθμοί Ιδιωτικότητας

Για πρώτη φορά η ιδέα της ιδιωτικότητας εμφανίζεται κατά τους αρχαίους χρόνους με την θέσπιση της Ψηφοφορίας. Με δημοκρατικό τρόπο οι πολίτες μπορούσαν να ρίξουν μυστικά την ψήφο τους στην κάλπη για κάποιο θέμα που τους αφορούσε. Με αυτόν τον τρόπο προστατευόταν η ιδιωτικότητα των πολιτικών τους επιλογών. Ωστόσο, η έννοια της ιδιωτικότητας ορίζεται επισήμως για πρώτη φορά στον Ρωμαϊκό νόμο όπου γίνεται μία ξεκάθαρη διάκριση ανάμεσα στο τι είναι ιδιωτικό (*privatus*: *privacy* στα Αγγλικά) και τι είναι δημόσιο (*publicus*: *public* στα Αγγλικά).

Παρόλο που στα αρχαία χρόνια η έννοια της ιδιωτικότητας είχε θεσμοθετηθεί και προστατευθεί από τις πρώτες δημοκρατικές κοινωνίες, στην σύγχρονη εποχή απειλείται αρκετά. Η τεχνολογία και οι συνήθειες που αναπτύξαμε με αυτήν άλλαξε ριζικά τον τρόπο με τον οποίο αντιλαμβανόμαστε τα προσωπικά δεδομένα και την έννοια της προστασίας της ιδιωτικότητας. Συγκεκριμένα, το διαδίκτυο έδωσε στους ανθρώπους την δυνατότητα να διαμοιράζονται καθημερινά και εύκολα πληροφορίες ενώ ταυτόχρονα εισήγαγε νέους τρόπους παραβίασης της ιδιωτικότητας (π.χ. ηλεκτρονικό έγκλημα).

Αρκετές εταιρίες, μικρές ή μεγάλες συλλέγουν και μεταδίδουν καθημερινά εκατοντάδες προσωπικά δεδομένα στο διαδίκτυο. Όλα αυτά τα δεδομένα αποθηκεύονται μέσα σε τεράστιες βάσεις δεδομένων πληροφοριακών συστημάτων. Τόσο οι εταιρίες όσο και συγκεκριμένα άτομα (π.χ. νόμιμα εξουσιοδοτημένα ή κακόβουλοι χρήστες) μπορούν να έχουν μεγάλη ικανότητα πρόσβασης σε αυτές τις βάσεις δεδομένων. Αυτό το γεγονός, οδηγεί σε αυξημένα προβλήματα προστασίας της ιδιωτικότητας των ανθρώπων.

Σε έναν τέτοιο ψηφιακό κόσμο, «πλημμυρισμένο» από διασυνδεδεμένους ηλεκτρονικούς υπολογιστές, ο καθηγητής Jeffrey Rosen του «George Washington University Law School» αναφέρει ότι το διαδίκτυο είναι ένας χώρος στον οποίο οτιδήποτε μεταφορτώνεται μπορεί να αποθηκευθεί για πάντα.

Η πληροφορία στον σύγχρονο κόσμο αποτελεί πολύτιμο αγαθό που μπορεί να βοηθήσει όλους μας. Ωστόσο, όταν αυτές οι πληροφορίες αποτελούν προσωπικά δεδομένα θα πρέπει να προστατεύονται. Κάθε ένας από εμάς θα πρέπει να αντιλαμβάνεται πως έχει ευθύνη κάθε φορά που μεταφορτώνει πληροφορίες στο διαδίκτυο.

2.4: Προστασία Ιδιωτικότητας

Είναι ανάγκη των ημερών μας, περισσότερο από κάθε άλλη φορά, τα προσωπικά και ευαίσθητα δεδομένα των ανθρώπων να προστατευτούν. Η προστασία της ιδιωτικότητας περιλαμβάνει όλα εκείνα τα μέτρα που θα πρέπει να ληφθούν υπόψη ώστε να προστατευθούν τα προσωπικά δεδομένα του κάθε ανθρώπου είτε αυτά βρίσκονται σε προσωπικό είτε σε ξένο χώρο. Οι περισσότεροι άνθρωποι, στον ιδιωτικό τους χώρο διατηρούν προσωπικά έγγραφα όπως ταυτότητα, διαβατήριο ή ακόμα και πιστωτικές κάρτες. Είναι προσωπική ευθύνη του καθενός η προστασία αυτών των αντικειμένων όταν αυτά βρίσκονται στον ιδιωτικό τους χώρο (π.χ. σπίτι, αυτοκίνητο). Όταν κάποιος αποφασίζει να προστατέψει τα προσωπικά του αντικείμενα εκτός του ιδιωτικού του χώρου (π.χ. σε κάποια τραπεζική θυρίδα) ή να διαμοιράσει προσωπικά δεδομένα σε πληροφοριακά συστήματα θα πρέπει να επιλέξει υπεύθυνα ποιον θα εμπιστευτεί για την προστασία αυτών.

Γενικότερα, οι κατευθύνσεις με τις οποίες προστατεύονται τα προσωπικά δεδομένα είναι δύο παγκοσμίως. Η μία προσέγγιση είναι αυτή της ελεύθερης αγοράς όπου οι διάφορες εταιρίες μπορούν να επιλέξουν από μόνες τους την πολιτική προστασίας που θα ακολουθήσουν. Στην συνέχεια οι πελάτες μπορούν να επιλέξουν με την σειρά τους υπεύθυνα τις εταιρίες εκείνες που εμπιστεύονται περισσότερο για την προστασία της ιδιωτικότητας τους. Με αυτό τον τρόπο, οι εταιρίες ανταγωνίζονται μεταξύ τους και αναγκάζονται τελικά σε βάθος χρόνου να καλύψουν τις ανάγκες των πελατών σχετικά με την προστασία των προσωπικών τους δεδομένων, ώστε να μην χάσουν μέρος της αγοράς. Ωστόσο, σε περιπτώσεις όπου υπάρχει μονοπώλιο και απουσιάζει ο ανταγωνισμός, οι πελάτες δεν έχουν την πολυτέλεια της επιλογής και έτσι η εταιρία μπορεί να ορίζει την πολιτική προστασίας προς το συμφέρον της.

Η άλλη προσέγγιση στηρίζεται στην υπόθεση πως ο μέσος πελάτης αδυνατεί να κατανοήσει τις πολιτικές προστασίας των διαφόρων εταιριών, με αποτέλεσμα να μην μπορεί να επιλέξει υπεύθυνα την εταιρία που του παρέχει την μεγαλύτερη προστασία. Οι Carlos Jensen και Colin Potts υποστήριξαν και απέδειξαν αυτήν την υπόθεση. Έτσι, οι κυβερνήσεις αποκτούν μεγαλύτερο ρόλο και ορίζουν βασικές αρχές περί της προστασίας της ιδιωτικότητας. Ωστόσο, οι προσεγγίσεις που ακολουθούν οι διάφορες κυβερνήσεις εξαρτώνται σε μεγάλο μέρος από τις πολιτικές αποφάσεις και από το πολιτικό καθεστώς που επικρατεί κάθε στιγμή σε κάθε κράτος. Είναι σημαντικό να γνωρίζουμε όλοι πως τόσο στην «Διακήρυξη των Δικαιωμάτων του Ανθρώπου και του Πολίτη» όσο και στην «Οικουμενική Διακήρυξη των Δικαιωμάτων του Ανθρώπου» μπορούμε να εντοπίσουμε στο άρθρο 12 αυτό που αναφέρθηκε παραπάνω. Το γεγονός δηλαδή ότι θα πρέπει οι κυβερνήσεις να λάβουν μέτρα για την προστασία της ιδιωτικότητας των ανθρώπων.

Παρακάτω παραθέτουμε το άρθρο 12 των διακηρύξεων:

- **Διακήρυξη των Δικαιωμάτων του Ανθρώπου και του Πολίτη:** Η εξασφάλιση των δικαιωμάτων του ανθρώπου και του πολίτη κάνει αναγκαία την ύπαρξη μιας κρατικής εξουσίας. Άρα αυτή η εξουσία έχει θεσπιστεί για το καλό όλων και όχι για την ιδιωτική ωφέλεια αυτών, στους οποίους έχει ανατεθεί.

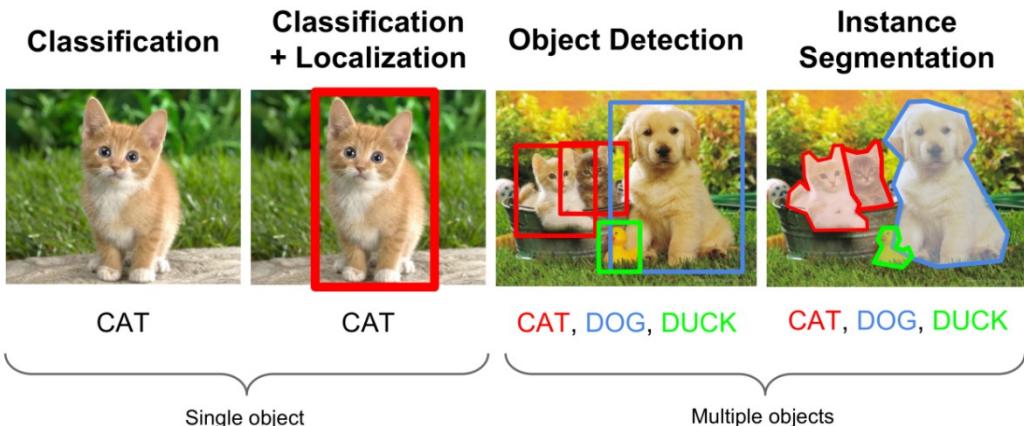
- **Οικουμενική Διακήρυξη των Δικαιωμάτων του Ανθρώπου:** Κανείς δεν επιτρέπεται να υποστεί αυθαίρετες επεμβάσεις στην ιδιωτική του ζωή, την οικογένεια, την κατοικία ή την αλληλογραφία του, ούτε προσβολές της τιμής και της υπόληψής του. Καθένας έχει το δικαίωμα να προστατευτεί από τον νόμο έναντι επεμβάσεων και προσβολών αυτού του είδους.

Κεφάλαιο 3: Αποταυτοποίηση Προσώπου

Σε αυτό το κεφάλαιο θα περιγράψουμε το πρόβλημα της αποταυτοποίησης προσώπου καθώς και θα επεξηγήσουμε τις διάφορες προϋπάρχουσες μεθόδους που προσπαθούν να το επιλύσουν.

3.1: Εισαγωγή

Με την ολοένα και περισσότερη χρήση του διαδικτύου αυξάνεται συνεχώς η αποθήκευση και ο διαμοιρασμός προσωπικών φωτογραφιών στις οποίες εμφανίζονται πρόσωπα ανθρώπων. Παράλληλα, τα τελευταία χρόνια η όραση υπολογιστών με την βοήθεια της βαθιάς μάθησης και των βαθιών νευρωνικών δικτύων έχει σημειώσει αρκετή πρόοδο [1].



Εικόνα 1: Σημαντικά προβλήματα που προσπαθεί να επιλύσει η όραση υπολογιστών

Ένα από τα προβλήματα που προσπαθεί να επιλύσει είναι η αναγνώριση εικόνας και μια από τις πιο συνηθισμένες εφαρμογές αυτής αποτελεί η αναγνώριση προσώπου [2]. Ο σκοπός αυτής είναι η αυτόματη αναγνώριση της ταυτότητας διαφόρων ανθρώπων από εικόνες στις οποίες εμφανίζονται τα πρόσωπα τους. Ωστόσο, αυτό οδηγεί σε αρκετά προβλήματα προστασίας προσωπικών δεδομένων [3] και ένας από τους πιο αποδοτικούς τρόπους με τους οποίους μπορεί δοθεί λύση σε αυτά είναι μέσω της αποταυτοποίησης προσώπου [7].



Εικόνα 2: Παράδειγμα ανίχνευσης και αναγνώρισης προσώπων σε μια εικόνα

Η αποταυτοποίηση προσώπου προσπαθεί να προστατέψει την ταυτότητα των ατόμων που εμφανίζονται μέσα σε εικόνες. Πρόκειται για την διαδικασία κατά την οποία μια εικόνα αλλοιώνεται σκόπιμα με σκοπό να αποταυτοποιηθεί έτσι ώστε να μην είναι πλέον εφικτή η αναγνώριση της ταυτότητας του εικονιζόμενου ατόμου από τα διάφορα αυτόματα συστήματα αναγνώρισης προσώπου. Επίσης, όσο αφορά τους ανθρώπινους παρατηρητές, ανάλογα με την εφαρμογή θα πρέπει είτε να μπορούν είτε όχι να αναγνωρίζουν το εικονιζόμενο πρόσωπο από την αποταυτοποιημένη εικόνα. Σε κάθε περίπτωση, ιδανικά η αποταυτοποιημένη εικόνα είναι καλό να είναι ρεαλιστική και καλαίσθητη.

3.2: Περιγραφή Προβλήματος

Θεωρούμε ένα σύνολο H που περιέχει ένα πλήθος από αρχικές εικόνες προσώπου. Έστω $\Gamma_i \in H$ μία από αυτές. Μία συνάρτηση $f: H \rightarrow H^d$ ονομάζεται συνάρτηση αποταυτοποίησης προσώπου όταν κατασκευάζει τις αντίστοιχες αποταυτοποιημένες εικόνες των αρχικών, όπου H^d αποτελεί το σύνολο των αποταυτοποιημένων εικόνων. Έστω $\Gamma_i^d \in H^d$ μία από αυτές. Επιπλέον, για κάθε ζευγάρι εικόνων (Γ_i, Γ_i^d) θα πρέπει να ισχύει $\Gamma_i \neq \Gamma_i^d$ καθώς και να έχουν τις ίδιες διαστάσεις.

3.3: Προϋπάρχουσες Μέθοδοι

Κατά καιρούς αναπτύχθηκαν διάφορες μέθοδοι αποταυτοποίησης προσώπου. Οι μέθοδοι αυτοί χωρίζονται σε τρεις κατηγορίες. Στην πρώτη κατηγορία βρίσκονται οι αφελείς μέθοδοι (γνωστές και ως «ad-hoc») [4-6], στην δεύτερη κατηγορία βρίσκονται οι μέθοδοι [7-10] που βασίζονται στο μοντέλο προστασίας της k-ανωνυμίας [11] και στην τρίτη κατηγορία βρίσκονται οι σύγχρονοι μέθοδοι [12-18].

3.3.1: Αφελείς Μέθοδοι

Οι μέθοδοι αυτοί παραμορφώνοντας έντονα το πρόσωπο στην εικόνα είναι ικανές να μπερδέψουν με μεγάλη επιτυχία απλοϊκά συστήματα αναγνώρισης προσώπου. Ωστόσο, το πρόβλημα που προκύπτει με αυτές τις μεθόδους είναι ότι ενώ τις περισσότερες φορές ένας ανθρώπινος παρατηρητής δεν μπορεί να αναγνωρίσει την ταυτότητα του προσώπου μιας αποταυτοποιημένης εικόνας, ένα σύγχρονο σύστημα αναγνώρισης προσώπου είναι σε θέση να το κάνει. Οι μέθοδοι αυτοί, προϋποθέτουν αρχικά την χρήση κάποιου αλγορίθμου για την ανίχνευση της θέσης του προσώπου στην εικόνα.

Επιπλέον, σε εφαρμογές όπου θέλουμε οι αποταυτοποιημένες εικόνες να είναι όσο το δυνατόν ρεαλιστικές και όμοιες με τις αρχικές, αυτές οι μέθοδοι δεν μπορούν να θεωρηθούν αποδεκτές διότι αλλοιώνουν έντονα την εμφάνιση του προσώπου. Ακόμη, οι μέθοδοι αυτοί αδυνατούν να διατηρήσουν ικανοποιητικά στο αποταυτοποιημένο πρόσωπο τα μη-ταυτοτικά χαρακτηριστικά του αρχικού, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα, με αποτέλεσμα να είναι εξίσου μη αποδεκτές σε εφαρμογές που τα χρειάζονται.



Εικόνα 3: Παραδείγματα αποταυτοποιημένων εικόνων με τις αφελείς μεθόδους

3.3.1.1: Μάσκα

Η μέθοδος αυτή τοποθετεί μία μάσκα στην περιοχή της εικόνας όπου βρίσκεται το πρόσωπο. Η μάσκα που θα χρησιμοποιηθεί μπορεί να είναι χρωματισμένη με κάποιο χρώμα ή μπορεί να έχει κάποιο μοτίβο. Ανάλογα με το σχήμα της μάσκας και το σημείο στο οποίο την τοποθετούμε καταφέρνουμε να αποκρύψουμε διαφορετικές πληροφορίες του προσώπου κάθε φορά. Ακολουθούν μερικές από τις πιο γνωστές μάσκες:

- **Μάσκα «Blackout»:** Η μάσκα αυτή έχει τετράγωνο, ορθογώνιο, κυκλικό ή ελλειπτικό σχήμα και τοποθετείται ώστε να αποκρύψει ολόκληρο το πρόσωπο.
- **Μάσκα «Bar»:** Η μάσκα αυτή έχει σχήμα μπάρας και τοποθετείται ώστε να αποκρύψει τα μάτια του προσώπου.
- **Μάσκα «Τ»:** Η μάσκα αυτή έχει σχήμα Τ και τοποθετείται ώστε να αποκρύψει τα μάτια και την μύτη του προσώπου.
- **Μάσκα «Mouth Only»:** Η μάσκα αυτή έχει σχήμα τετραγώνου και τοποθετείται ώστε να αποκρύψει όλα τα χαρακτηριστικά του προσώπου εκτός από το στόμα και το πιγούνι.

3.3.1.2: Πιξελοποίηση

Η μέθοδος αυτή πιξελιάζει την περιοχή της εικόνας όπου βρίσκεται το πρόσωπο. Συγκεκριμένα, τοποθετεί νοητά ένα πλέγμα από εικονοστοιχεία πάνω από την περιοχή της εικόνας όπου βρίσκεται το πρόσωπο. Το μέγεθος των εικονοστοιχείων του πλέγματος είναι μεγαλύτερο από το μέγεθος των εικονοστοιχείων της εικόνας. Στην συνέχεια, κάθε εικονοστοιχείο του πλέγματος χρωματίζεται ανάλογα με το τι υπάρχει στην εικόνα σε εκείνη την περιοχή. Πιο συγκεκριμένα, για τον υπολογισμό του χρώματος κάθε εικονοστοιχείου του πλέγματος χρησιμοποιούμε κάποιο τοπικό φίλτρο. Κλασική περίπτωση φίλτρου για πιξελοποίηση αποτελεί το φίλτρο Box Blur. Το φίλτρο αυτό υπολογίζει την πράξη της συνέλιξης μεταξύ των εικονοστοιχείων της αντίστοιχης περιοχής της εικόνας και ενός συνελικτικού πυρήνα μέσης τιμής. Το αποτέλεσμα της πράξης ισοδυναμεί με τον

υπολογισμό της μέσης τιμής των χρωμάτων των εικονοστοιχείων της περιοχής. Αυτό έχει σαν αποτέλεσμα να μειώνεται το πλήθος των διαφορετικών χρωμάτων της εικόνας. Οι πρώτες δύο διαστάσεις του συνελικτικού πυρήνα θα πρέπει να ισοδυναμούν με τις διαστάσεις των εικονοστοιχείων του πλέγματος και η τρίτη διάσταση θα πρέπει να ισοδυναμεί με το πλήθος των καναλιών της εικόνας. Ένα παράδειγμα $3 \times 3 \times 1$ συνελικτικού πυρήνα μέσης τιμής για μία μονοκαναλική εικόνα είναι το παρακάτω:

$$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Ένας άλλος παρόμοιος τρόπος με τον οποίο μπορούμε να κάνουμε πιξελοποίηση είναι κάνοντας μείωση της ανάλυσης της εικόνας στην περιοχή του προσώπου. Αρχικά, κάνουμε υποδειγματοληψία με κάποιο τοπικό φίλτρο ώστε να μειωθεί η ανάλυση της εικόνας στην περιοχή του προσώπου και στην συνέχεια υπερδειγματοληψία με την μέθοδο της παρεμβολής κοντινότερου γείτονα ώστε να ανακτήσουμε τις αρχικές διαστάσεις.

3.3.1.3: Θόλωση

Η μέθοδος αυτή θολώνει την περιοχή της εικόνας όπου βρίσκεται το πρόσωπο εφαρμόζοντας σε αυτό ένα τοπικό κινούμενο φίλτρο θόλωσης. Κλασική περίπτωση φίλτρου θόλωσης αποτελεί το Γκαουσιανό φίλτρο. Το φίλτρο αυτό υπολογίζει την πράξη της συνέλιξης μεταξύ των εικονοστοιχείων της εκάστοτε υποπεριοχής της εικόνας και ενός συνελικτικού Γκαουσιανού πυρήνα. Ανάλογα με το μέγεθος του πυρήνα καθώς και με την τυπική απόκλιση της Γκαουσιανής συνάρτησης που προσεγγίζουμε μπορούμε να επιτύχουμε διαφορετική υποβάθμιση στην ποιότητα της εικόνας. Η τρίτη διάσταση του πυρήνα θα πρέπει να ισοδυναμεί με το πλήθος των καναλιών της εικόνας. Ένα παράδειγμα $5 \times 5 \times 1$ συνελικτικού Γκαουσιανού πυρήνα με τυπική απόκλιση $\sigma = 1$ για μία μονοκαναλική εικόνα είναι το παρακάτω:

$$\frac{1}{273} \begin{bmatrix} 1 & 4 & 7 & 4 & 1 \\ 4 & 16 & 26 & 16 & 4 \\ 7 & 26 & 41 & 26 & 7 \\ 4 & 16 & 26 & 16 & 4 \\ 1 & 4 & 7 & 4 & 1 \end{bmatrix}$$

3.3.1.4: Αρνητικό

Η μέθοδος αυτή δημιουργεί το αντίστοιχο αρνητικό της περιοχής της εικόνας όπου βρίσκεται το πρόσωπο. Όταν η εικόνα είναι ασπρόμαυρη τότε η τιμή x κάθε εικονοστοιχείου αντιστρέφεται, δηλαδή γίνεται από μαύρο ($0 \rightarrow 255$) και το αντίστροφο ($255 \rightarrow 0$). Ενώ, όταν η εικόνα είναι διαβάθμισης του γκρι τότε η τιμή x κάθε εικονοστοιχείου γίνεται $255 - x$, δηλαδή η ισοδύναμη τιμή στο αντίθετο άκρο της κλίμακας του γκρι. Η τελευταία περίπτωση μπορεί να εφαρμοστεί και σε πολυκαναλικές εικόνες εφαρμόζοντας τον υπολογισμό σε κάθε ένα κανάλι ξεχωριστά.

3.3.1.5: Θόρυβος

Η μέθοδος αυτή προσθέτει θόρυβο στην περιοχή της εικόνας όπου βρίσκεται το πρόσωπο. Γενικότερα, υπάρχει αρκετή ευελιξία στο είδος της κατανομής πιθανοτήτων με βάση την οποία δειγματοληπτείται ο θόρυβος. Κλασικές περιπτώσεις θορύβου είναι ο Γκαουσιανός και ο κρουστικός. Το πλήθος των εικονοστοιχείων που πρόκειται να αλλοιωθούν είναι ζήτημα πολιτικής της εκάστοτε τεχνικής. Για παράδειγμα θα μπορούσαμε να προσθέσουμε θόρυβο σε ένα μόνο εικονοστοιχείο, σε όλα ή σε μερικά. Επίσης, το ποια συγκεκριμένα θα είναι αυτά τα εικονοστοιχεία καθορίζεται στοχαστικά από τον αλγόριθμο. Στην περίπτωση του κρουστικού θορύβου όπου η εικόνα είναι ασπρόμαυρη, η τιμή x του εκάστοτε τυχαία επιλεγμένου εικονοστοιχείου αντιστρέφεται, δηλαδή γίνεται από μαύρο άσπρο ($0 \rightarrow 255$) και το αντίστροφο ($255 \rightarrow 0$). Ενώ, στην περίπτωση του Γκαουσιανού θορύβου όπου η εικόνα είναι διαβάθμισης του γκρι, η τιμή x του εκάστοτε τυχαία επιλεγμένου εικονοστοιχείου γίνεται $x + n$ όπου n ο θόρυβος που προστίθεται. Η τελευταία περίπτωση μπορεί να εφαρμοστεί και σε πολυκαναλικές εικόνες εφαρμόζοντας τον υπολογισμό σε κάθε ένα κανάλι ξεχωριστά.

3.3.1.6: Κατωφλίωση

Η μέθοδος αυτή κατωφλίωνει την περιοχή της εικόνας όπου βρίσκεται το πρόσωπο χρησιμοποιώντας μία τιμή κατωφλίου r που επιλέγεται τυχαία ή σκόπιμα από το εύρος $[0, 255]$. Η τιμή x κάθε εικονοστοιχείου της εικόνας γίνεται 0 (μαύρο) όταν ισχύει η σχέση $x < r$ και 255 (άσπρο) όταν δεν ισχύει. Αυτό έχει σαν αποτέλεσμα την μετατροπή μίας εικόνας διαβάθμισης του γκρι σε ασπρόμαυρη.

3.3.1.7: Πρόσωπο «κ. Πατάτας»

Η μέθοδος αυτή αξιοποιεί μία μεγάλη βάση δεδομένων από διάφορα χαρακτηριστικά προσώπων (π.χ. μάτια, μύτη, στόμα). Τα χαρακτηριστικά του προσώπου της εικόνας αντικαθίστανται με χαρακτηριστικά που επιλέγονται από την βάση δεδομένων. Επειδή όμως τα διάφορα χαρακτηριστικά επιλέγονται τυχαία και χωρίς να προσαρμόζονται κατάλληλα πριν αντικαταστήσουν τα υπάρχοντα, το τελικό πρόσωπο δεν είναι ομοιόμορφο και ρεαλιστικό καθώς θυμίζει το πρόσωπο του «κ. Πατάτα» όπου φαίνεται ξεκάθαρα πως το πρόσωπο αποτελεί σύνθεση διαφορετικών χαρακτηριστικών.

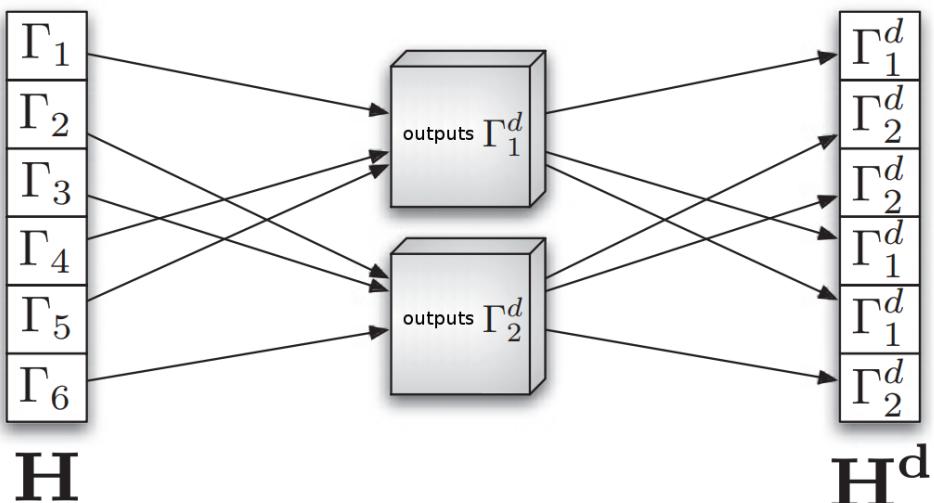
3.3.1.8: Μείωση

Η μέθοδος αυτή μειώνει σημαντικά το πλήθος των διαφορετικών χρωμάτων της περιοχής της εικόνας στην οποία βρίσκεται το πρόσωπο. Η μείωση των διαφορετικών χρωμάτων της εικόνας γίνεται αρχικά με επιλογή ενός μικρού συνόλου k αντιπροσωπευτικών χρωμάτων και στην συνέχεια με αντιστοίχηση κάθε εικονοστοιχείου σε αυτά. Τόσο ο τρόπος επιλογής των k χρωμάτων όσο και ο τρόπος της αντιστοίχισης των εικονοστοιχείων με αυτά εξαρτώνται από την εκάστοτε τεχνική. Ένας απλοϊκός τρόπος με τον οποίο μπορεί να υλοποιηθεί αυτή η μέθοδος για εικόνες διαβάθμισης του γκρι είναι με την χρήση k διαστημάτων τιμών ίσου μήκους. Θα πρέπει, η τιμή του κάτω ορίου του πρώτου διαστήματος να είναι 0 (άσπρο) και η τιμή του άνω ορίου του k διαστήματος να είναι 255 (μαύρο). Κάθε ένα από αυτά τα διαστήματα αντιπροσωπεύει και μία διαφορετική

απόχρωση του γκρι. Έτσι, η τιμή x κάθε εικονοστοιχείου της εικόνας αλλάζει στο χρώμα του διαστήματος μέσα στο οποίο αυτή υποπίπτει.

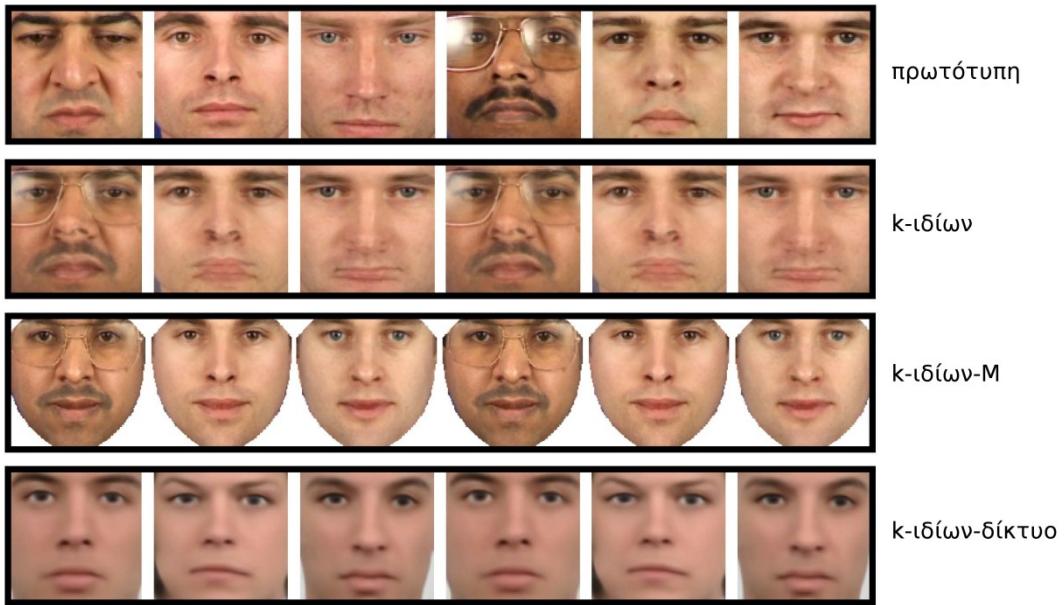
3.3.2: Μέθοδοι Μοντέλου k-ανωνυμίας

Οι μέθοδοι που βασίζονται στο μοντέλο προστασίας της k-ανωνυμίας χρησιμοποιούν στην είσοδο ένα σύνολο με εικόνες προσώπου (μία για κάθε άτομο) $H = \{\Gamma_1, \dots, \Gamma_M\}$ και κατασκευάζουν στην έξοδο ένα σύνολο με τις αντίστοιχες αποταυτοποιημένες $H^d = \{\Gamma_1^d, \dots, \Gamma_M^d\}$. Η αποταυτοποίηση γίνεται έτσι ώστε κάθε αποταυτοποιημένη εικόνα Γ_i^d να σχετίζεται αδιακρίτως με τουλάχιστον k εικόνες από το σύνολο H . Επίσης, αποδεικνύεται εύκολα πως η πιθανότητα επαναταύτισης κάθε εικόνας Γ_i^d είναι το πολύ $1 \div k$.



Εικόνα 4: Το γενικό μοντέλο προστασίας της k-ανωνυμίας

Οι πιο αντιπροσωπευτικοί μέθοδοι που υλοποιούν το μοντέλο προστασίας της k-ανωνυμίας είναι οι μέθοδοι τύπου «k-ιδίων». Γενικά, αυτές οι μέθοδοι μπορούν να λειτουργήσουν στον χώρο των εικονοστοιχείων των εικόνων, στον χώρο των παραμέτρων κάποιου μοντέλου ενεργής εμφάνισης ή σε κάποιο άλλο χώρο μειωμένων διαστάσεων μετά από ιδιοανάλυση (π.χ. στο χώρο των συντελεστών της ανάλυσης κυρίων συνιστωσών ή της γραμμικής διαχωριστικής ανάλυσης). Επίσης, σε αυτές τις μεθόδους είναι βασική προϋπόθεση η χρήση κάποιου αλγορίθμου για την ανίχνευση της θέσης των προσώπων στις εικόνες.



Εικόνα 5: Παραδείγματα αποταυτοποιημένων εικόνων με μεθόδους τύπου «κ-ιδίων»

Συγκριτικά με τις αφελείς μεθόδους οι μέθοδοι τύπου «κ-ιδίων» παράγουν πιο ποιοτικές εικόνες και είναι ικανές στο να μπερδεύουν ακόμη και σύγχρονα αυτόματα συστήματα αναγνώρισης προσώπου. Επίσης, με κατάλληλη παραμετροποίηση φαίνεται να μπορούν να διατηρήσουν στην αποταυτοποιημένη εικόνα μη-ταυτοτικά χαρακτηριστικά προσώπου, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα. Επιπλέον, σε εφαρμογές όπου θέλουμε οι αποταυτοποιημένες εικόνες να είναι όσο το δυνατόν ρεαλιστικές και όμοιες με τις αρχικές οι μέθοδοι αυτοί δεν μπορούν να θεωρηθούν αποδεκτές διότι αλλοιώνουν έντονα την εμφάνιση του προσώπου.

3.3.2.1: *k*-ιδίων

Η μέθοδος αυτή [7] για μία εικόνα Γ_i του συνόλου H βρίσκει τις $k - 1$ κοντινότερες της και υπολογίζει την μέση εικόνα $\bar{\Gamma}$, συμπεριλαμβάνοντας και την εικόνα Γ_i στον υπολογισμό. Η μέση εικόνα $\bar{\Gamma}$ που προκύπτει αποτελεί την αποταυτοποιημένη εικόνα για όλες τις k εικόνες που χρησιμοποιήθηκαν. Έτσι, αντιγράφει k φορές την εικόνα $\bar{\Gamma}$ στο σύνολο H^d και επαναλαμβάνει την διαδικασία με όσες εικόνες έχουν απομείνει στο σύνολο H που δεν είναι αποταυτοποιημένες.

Ο διανυσματικός χώρος στον οποίο λειτουργεί η μέθοδος *k*-ιδίων μπορεί να είναι είτε αυτός των εικονοστοιχείων είτε κάποιος άλλος μετά από μείωση διαστάσεων και ιδιοανάλυση. Επίσης, η μέθοδος είναι μη αναστρέψιμη επειδή δεν είναι εφικτό από τις αποταυτοποιημένες εικόνες να ανακτήσουμε τις αρχικές. Όταν χρησιμοποιούμε μεγαλύτερη τιμή k τότε τα πρόσωπα των αποταυτοποιημένων εικόνων τείνουν να μοιάζουν περισσότερο μεταξύ τους. Επιπλέον, όταν τα διάφορα χαρακτηριστικά των προσώπων στις εικόνες δεν είναι ευθυγραμμισμένα τότε στις αποταυτοποιημένες εικόνες εμφανίζονται διάφορα ανεπιθύμητα τεχνουργήματα.

3.3.2.2: *k*-ιδίων-επιλογή

Η μέθοδος αυτή [8] χρησιμοποιεί την ίδια αρχή λειτουργίας της μεθόδου *k*-ιδίων με την μόνη διαφορά ότι πλέον η επιλογή των $k - 1$ κοντινότερων εικόνων γίνεται βάση αυστηρά καθορισμένων μη-ταυτοτικών χαρακτηριστικών προσώπου, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα. Πιο συγκεκριμένα, το σύνολο H χωρίζεται σε πολλά μη επικαλυπτόμενα υποσύνολα βάση των χαρακτηριστικών που αναφέρθηκαν παραπάνω. Για το χαρακτηριστικό του φύλου, για παράδειγμα, το σύνολο H χωρίζεται στα υποσύνολα H_F και H_M για τις εικόνες γυναικών και ανδρών αντίστοιχα. Στην συνέχεια, χρησιμοποιείται η υπάρχουσα μέθοδος των *k*-ιδίων για την επιλογή των k κοντινότερων εικόνων από το υποσύνολο του H που σχετίζεται με τα χαρακτηριστικά της εικόνας προς αποταυτοποίηση. Η μέθοδος αυτή συγκριτικά με την μέθοδο *k*-ιδίων παράγει εικόνες καλύτερης ποιότητας αλλά εξακολουθεί να παρουσιάζει το πρόβλημα των ανεπιθύμητων τεχνουργημάτων.

3.3.2.3: *k*-ιδίων-Μ

Η μέθοδος αυτή [9] χρησιμοποιεί την αρχή λειτουργίας της μεθόδου *k*-ιδίων συνδυαστικά με ένα μοντέλο ενεργής εμφάνισης [19]. Τα μοντέλα ενεργής εμφάνισης είναι γεννητικά παραμετρικά μοντέλα τα οποία μπορούν να εκπαιδευτούν αλλάζοντας κατάλληλα τις παραμέτρους τους ώστε να ανακατασκευάζουν την πληροφορία εισόδου. Έτσι, αρχικά η μέθοδος αυτή εκπαιδεύει ένα μοντέλο ενεργής εμφάνισης ξεχωριστά για κάθε μία από τις εικόνες του συνόλου H και δημιουργεί για κάθε μία από αυτές ένα διανυσματικό περιεχόμενο που περιέχει τις παραμέτρους του μοντέλου. Όλα αυτά τα διανύσματα παραμέτρων συνθέτουν ένα νέο σύνολο H^* στο οποίο εφαρμόζεται η κλασική μέθοδος *k*-ιδίων. Έτσι, η επιλογή των k κοντινότερων εικόνων καθώς και ο υπολογισμός της μέσης εικόνας δεν γίνεται πλέον στον διανυσματικό χώρο των εικονοστοιχείων αλλά στον χώρο των παραμέτρων του μοντέλου. Η μέθοδος αυτή είναι καλύτερη από την κλασική μέθοδο *k*-ιδίων αφού η χρήση του μοντέλου ενεργής εμφάνισης λύνει το πρόβλημα της εμφάνισης των ανεπιθύμητων τεχνουργημάτων στις αποταυτοποιημένες εικόνες.

3.3.2.4: *k*-ιδίων-δίκτυο

Παρόμοια με την μέθοδο *k*-ιδίων αυτή η μέθοδος [10] χρησιμοποιεί τις εικόνες του συνόλου H ως είσοδο και προσπαθεί από αυτές να κατασκευάσει στην έξοδο το σύνολο H^d με τις αντίστοιχες αποταυτοποιημένες. Ωστόσο, συγκριτικά με τις υπόλοιπες μεθόδους, χρησιμοποιεί επιπλέον ως είσοδο ένα γεννητικό νευρωνικό δίκτυο [20] το οποίο είναι εκπαιδευμένο με ένα αντιπροσωπευτικό σύνολο H^p που περιέχει διάφορες ρεαλιστικές εικόνες προσώπων. Το εκπαιδευμένο γεννητικό νευρωνικό δίκτυο είναι ικανό να συνθέτει καινούργιες και ρεαλιστικές εικόνες που να διατηρούν συγκεκριμένα μη-ταυτοτικά χαρακτηριστικά προσώπου, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα. Το σύνολο H^p απομονώνει την άμεση σύνδεση μεταξύ των συνόλων H και H^d καθώς και καθορίζεται σκόπιμα ώστε να παρέχεται η ιδιότητα της *k*-ανωνυμίας. Πλέον, οι αποταυτοποιημένες εικόνες δεν προκύπτουν υπολογιστικά ως μέσος όρος αλλά παράγονται από το γεννητικό νευρωνικό δίκτυο.

3.3.3: Σύγχρονοι Μέθοδοι

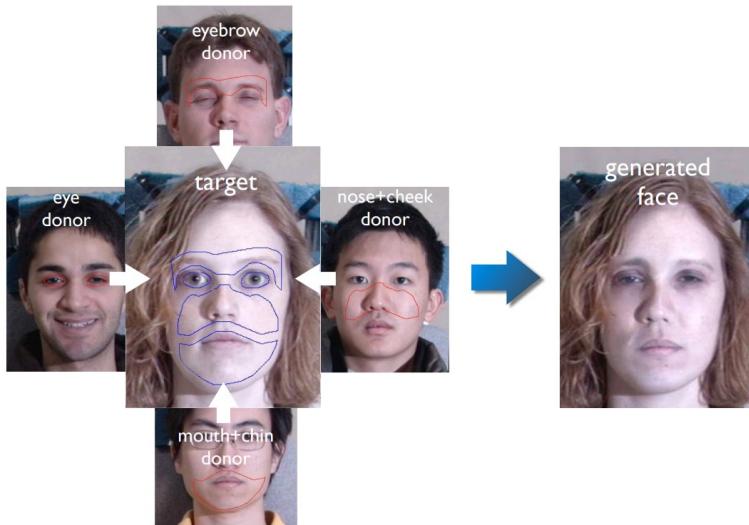
Οι μέθοδοι αυτοί αρχικά κάνουν σύνθεση ενός καινούργιου και ρεαλιστικού προσώπου και στην συνέχεια αντικαθιστούν το υπάρχον πρόσωπο της εικόνας με αυτό. Επίσης, προσπαθούν το καινούργιο πρόσωπο που κατασκευάζουν να ταιριάζει καλά μέσα στην εικόνα ώστε να υπάρχει ένα ρεαλιστικό οπτικό αποτέλεσμα καθώς και να διατηρεί τα μηταυτοτικά χαρακτηριστικά του αρχικού προσώπου, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα.

Τα τελευταία χρόνια τα γεννητικά μοντέλα [20] βρίσκονται σε άνθηση. Τα μοντέλα αυτά έχουν την ικανότητα να προσεγγίσουν την κατανομή πιθανοτήτων κάποιων δεδομένων εκπαίδευσης με αποτέλεσμα να μπορούν είτε να τα ανακατασκευάσουν είτε να παράγουν καινούργια και άγνωστα δείγματα. Για αυτό το λόγο, αυτά τα μοντέλα αξιοποιούνται από αρκετές σύγχρονες μεθόδους αποταυτοποίησης για σύνθεση καινούργιων και ρεαλιστικών προσώπων.



Εικόνα 6: Παραδείγματα σύνθεσης καινούργιων προσώπων με γεννητικά μοντέλα

Συγκριτικά με τις αφελείς μεθόδους καθώς και με αυτές που βασίζονται στο μοντέλο προστασίας της k-ανωνυμίας, οι σύγχρονοι μέθοδοι παράγουν πιο ρεαλιστικές αποταυτοποιημένες εικόνες. Οι σύγχρονες μέθοδοι είναι ικανές στο να μπερδεύουν ακόμη και τα πιο προηγμένα αυτόματα συστήματα αναγνώρισης προσώπου. Επίσης, συγκριτικά με τις μεθόδους που βασίζονται στο μοντέλο προστασίας της k-ανωνυμίας, οι σύγχρονοι μέθοδοι δεν απαιτούν κάθε εικόνα του συνόλου H να σχετίζεται με διαφορετικό άτομο γιατί αυτός ο περιορισμός στις περισσότερες ρεαλιστικές εφαρμογές είναι απαγορευτικός.



Εικόνα 7: Παράδειγμα σύνθεσης καινούργιου προσώπου με δωρητές χαρακτηριστικών

Επίσης, σε εφαρμογές όπου θέλουμε οι αποταυτοποιημένες εικόνες να είναι όσο το δυνατόν όμοιες με τις αρχικές εικόνες, αυτές οι μέθοδοι δεν μπορούν να θεωρηθούν αποδεκτές διότι αντικαθιστούν το πρόσωπο με κάποιο άλλο.

3.3.3.1: Πολυπαραγοντική Αποταυτοποίηση Προσώπου

Σε αυτήν την μέθοδο [12] εκπαιδεύεται ένα γεννητικό πολυπαραγοντικό μοντέλο με στόχο την ανακατασκευή ενός συνόλου εικόνων. Το γεννητικό μοντέλο μπορεί να αποτελεί συνδυασμός γραμμικών, διγραμμικών και τετραγωνικών απλούστερων μοντέλων. Κάθε εικόνα παραγοντοποιείται από το γεννητικό μοντέλο σε ταυτοικούς και μη-ταυτοικούς παράγοντες με αποτέλεσμα να εκφράζεται πλέον διανυσματικά ως συνδυασμός αυτών. Στην συνέχεια, τα διανύσματα παραγόντων αποταυτοποιούνται με κάποιον αλγόριθμο αποταυτοποίησης. Για παράδειγμα, ένας αλγόριθμος αποταυτοποίησης που μπορεί να χρησιμοποιηθεί είναι ο αλγόριθμος (ϵ , k)-χάρτης [21]. Τέλος, τα αποταυτοποιημένα διανύσματα παραγόντων μαζί με τις βάσεις του γεννητικού μοντέλου χρησιμοποιούνται για την κατασκευή των αποταυτοποιημένων εικόνων. Η μέθοδος αυτή διατηρεί καλύτερα τα μη-ταυτοικά χαρακτηριστικά στις αποταυτοποιημένες εικόνες συγκριτικά με τις μεθόδους που βασίζονται στο μοντέλο προστασίας της k -ανωνυμίας.

Το ανακατασκευαστικό γεννητικό μοντέλο M για δεδομένα διάστασης k ορίζεται ως εξής:

$$M_k(\boldsymbol{\mu}, \mathbf{B}^1, \mathbf{B}^2, \mathbf{W}, \mathbf{Q}^1, \mathbf{Q}^2; \mathbf{c}_1, \mathbf{c}_2) = (1, \mathbf{c}_1^T, \mathbf{c}_2^T) \begin{pmatrix} \boldsymbol{\mu}_k & \mathbf{B}_k^2 & 0 \\ \mathbf{B}_k^1 & \mathbf{W}_k & \mathbf{Q}_k^1 \\ 0 & \mathbf{Q}_k^2 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{c}_2 \\ \mathbf{c}_1 \end{pmatrix}$$

όπου $\boldsymbol{\mu}$ ο μέσος, \mathbf{B}^1 , \mathbf{B}^2 γραμμικές βάσεις, \mathbf{W} διγραμμική βάση, \mathbf{Q}^1 , \mathbf{Q}^2 τετραγωνικές βάσεις και \mathbf{c}_1 , \mathbf{c}_2 συντελεστές. Οι \mathbf{Q}^1 , \mathbf{Q}^2 επιλέγονται να είναι άνω τριγωνικοί. Παρόλο που το γεννητικό μοντέλο M είναι τετραγωνικό, ουσιαστικά περιέχει μικρότερων διαστάσεων γραμμικά, διγραμμικά και τετραγωνικά μοντέλα σαν ειδικές περιπτώσεις. Για παράδειγμα, με δεδομένα \mathbf{c}_1 και \mathbf{c}_2 αν θέσουμε $\mathbf{W} = \mathbf{Q}^1 = \mathbf{Q}^2 = \mathbf{0}$ τότε βγάζουμε τον τύπο για το γραμμικό μοντέλο:

$$M_k^L(\boldsymbol{\mu}, \mathbf{B}^1, \mathbf{B}^2, \mathbf{0}, \mathbf{0}, \mathbf{0}; \mathbf{c}_1, \mathbf{c}_2) = \mu_k + \mathbf{c}_1^T \mathbf{B}_k^1 + \mathbf{B}_k^2 \mathbf{c}_2$$

και ομοίως αν θέσουμε $\mathbf{Q}^1 = \mathbf{Q}^2 = \mathbf{0}$ τότε βγάζουμε τον τύπο για το διγραμμικό μοντέλο:

$$M_k^B(\boldsymbol{\mu}, \mathbf{B}^1, \mathbf{B}^2, \mathbf{W}, \mathbf{0}, \mathbf{0}; \mathbf{c}_1, \mathbf{c}_2) = \mu_k + \mathbf{c}_1^T \mathbf{B}_k^1 + \mathbf{B}_k^2 \mathbf{c}_2 + \mathbf{c}_1^T \mathbf{W}_k \mathbf{c}_2$$

Έτσι λοιπόν, ανάλογα με τις παραμέτρους του γεννητικού μοντέλου M μπορούμε να δημιουργήσουμε μίξεις γραμμικών, διγραμμικών και τετραγωνικών μοντέλων. Ο στόχος της μεθόδου είναι να βρει με κάποιο αλγόριθμο βελτιστοποίησης τις παραμέτρους του γεννητικού μοντέλου M που ελαχιστοποιούν το σφάλμα ανακατασκευής για δεδομένο σύνολο εικόνων $D = \{\mathbf{d}_1, \dots, \mathbf{d}_n\}$:

$$\arg \min_{\Gamma, \mathbf{c}_1, \mathbf{c}_2} \sum_{l=1}^n \|M(\Gamma; \mathbf{c}_1(l), \mathbf{c}_2(l)) - \mathbf{d}_l\|_2^2$$

όπου Γ οι βάσεις του γεννητικού μοντέλου M που ορίζονται ως $\Gamma = (\boldsymbol{\mu}, \mathbf{B}^1, \mathbf{B}^2, \mathbf{W}, \mathbf{Q}^1, \mathbf{Q}^2)$ και $\mathbf{C}_1, \mathbf{C}_2$ οι ταυτοικοί και μη-ταυτοικοί συντελεστές των εικόνων που ορίζονται ως $\mathbf{C}_1 = (\mathbf{c}_1(l), \dots, \mathbf{c}_1(n))$ και $\mathbf{C}_2 = (\mathbf{c}_2(l), \dots, \mathbf{c}_2(n))$ αντιστοίχως.

3.3.3.2: Αποταυτοποίηση Προσώπου με δωρητές χαρακτηριστικών προσώπου

Η μέθοδος αυτή [13] προσπαθεί να συνθέσει ένα καινούργιο αποταυτοποιημένο πρόσωπο το οποίο να είναι ρεαλιστικό και να διατηρεί τα μη-ταυτοικά χαρακτηριστικά του προσώπου της αρχικής εικόνας, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναλισθηματική έκφραση ή την πόζα. Για να το επιτύχει αυτό κάνει χρήση μιας μεγάλης βάσης δεδομένων εικόνων προσώπου. Για όλες τις εικόνες της βάσης δεδομένων εξάγει τα μη-ταυτοικά χαρακτηριστικά προσώπου με κατάλληλους εξαγωγείς χαρακτηριστικών. Κατά την αποταυτοποίηση μιας εικόνας εξάγει και από αυτήν τα αντίστοιχα μη-ταυτοικά χαρακτηριστικά και στην συνέχεια τα χρησιμοποιεί ως ερώτημα προς την βάση δεδομένων ώστε βάση κάποιας μετρικής απόστασης να επιλέξει τις k κοντινότερες εικόνες προσώπου. Στην συνέχεια, οι k επιλεγμένες εικόνες χρησιμοποιούνται ως δότες διαφόρων χαρακτηριστικών προσώπου (π.χ. μάτια, μύτη, στόμα) για την σύνθεση του καινούργιου προσώπου. Η ποιότητα του καινούργιου συνθετικού προσώπου εξαρτάται σημαντικά από το πλήθος και την ποικιλία των εικόνων που περιέχονται στην βάση δεδομένων, τα μη-ταυτοικά χαρακτηριστικά προσώπου που χρησιμοποιούνται, την ικανότητα των τεχνικών εξαγωγής αυτών, την μετρική απόστασης καθώς και τον τρόπο σύνθεση του καινούργιου προσώπου από τα επιμέρους χαρακτηριστικά.

3.3.3.3: Αποταυτοποίηση Προσώπου με Γεννητικά Βαθιά Νευρωνικά Δίκτυα

Η μέθοδος αυτή [14] εξάγει αρχικά το διάνυσμα χαρακτηριστικών μίας εικόνας προσώπου όπως αυτό αναπαριστάνεται στα εσωτερικά επίπεδα ενός σύγχρονου εκπαιδευμένου βαθιού νευρωνικού δικτύου (π.χ. συνελικτικό νευρωνικό δίκτυο αρχιτεκτονικής VGG-16) [22] που έχει εκπαιδευτεί για αναγνώριση προσώπου. Στην συνέχεια με βάση αυτό το διάνυσμα χαρακτηριστικών κάνει αναζήτηση σε μία σταθερή βάση δεδομένων που περιέχει τα αντίστοιχα διανύσματα χαρακτηριστικών για διάφορες εικόνες προσώπων M ταυτοτήτων και επιλέγει τις $k \ll M$ κοντινότερες ταυτότητες. Επίσης, η μετρική ομοιότητας

που χρησιμοποιείται για την σύγκριση μεταξύ των διανυσμάτων είναι αυτή του συνημίτονου. Στην συνέχεια, οι *k* κοντινότερες ταυτότητες δίδονται ως είσοδο σε ένα εκπαιδευμένο βαθύ γεννητικό νευρωνικό δίκτυο για την σύνθεση μιας καινούργιας και ρεαλιστικής τεχνητής εικόνας προσώπου. Η καινούργια εικόνα προσώπου έχει και από τις *k* επιλεγμένες ταυτότητες διάφορα οπτικά χαρακτηριστικά. Το βαθύ γεννητικό νευρωνικό δίκτυο έχει εκπαιδευτεί στο να μπορεί να ανακατασκευάζει εικόνες προσώπου. Επίσης, είναι δυνατό να δοθεί στην είσοδο του ένα διάνυσμα που να περιγράφει το ποια μηταυτοτικά χαρακτηριστικά θα θέλαμε το καινούργιο συνθετικό πρόσωπο να έχει. Έτσι, σε περίπτωση που επιθυμούμε η αποταυτοποιημένη εικόνα να διατηρεί κάποια μη-ταυτοτικά χαρακτηριστικά του αρχικού προσώπου μπορούμε να τα εξάγουμε με κατάλληλους εξαγωγείς χαρακτηριστικών και να τα περάσουμε σαν διάνυσμα στο βαθύ γεννητικό νευρωνικό δίκτυο. Τέλος, ένα βήμα πριν την αντικατάσταση του αρχικού προσώπου από το καινούργιο χρησιμοποιούνται τα ορόσημα των δύο προσώπων και γίνεται κατάλληλη προσαρμογή του καινούργιου προσώπου ώστε να ευθυγραμμιστεί σύμφωνα με το αρχικό.

3.4: Προσέγγιση Προτεινόμενης Μεθόδου

Όλες οι μέθοδοι που εξετάσαμε σε αυτό το κεφάλαιο παραμορφώνουν έντονα το πρόσωπο στην εικόνα με αποτέλεσμα να μην μπορούμε να τις χρησιμοποιήσουμε σε εφαρμογές όπου θέλουμε οι αποταυτοποιημένες εικόνες να είναι όσο το δυνατόν ρεαλιστικές και όμοιες με τις αρχικές. Ακόμη, αδυνατούν να διατηρήσουν ικανοποιητικά στο αποταυτοποιημένο πρόσωπο τα μη-ταυτοτικά χαρακτηριστικά του αρχικού, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα, με αποτέλεσμα να είναι εξίσου μη αποδεκτές σε εφαρμογές που τα χρειάζονται.

Στην παρούσα εργασία προτείνουμε μια νέα και καινοτόμο μέθοδο αποταυτοποίησης προσώπου η οποία επιλύει τα παραπάνω προβλήματα. Συγκεκριμένα, η προτεινόμενη μέθοδος μπορεί να αποταυτοποιήσει μία εικόνα προσώπου αλλοιώνοντας την ελάχιστα έτσι ώστε η αποταυτοποιημένη εικόνα να είναι ανεπαίσθητα διαφορετική από την αρχική. Η προσέγγιση που ακολουθεί η προτεινόμενη μέθοδος είναι διαφορετική από αυτές των προϋπάρχουσων μεθόδων διότι εισάγει την χρήση των αντιπαλικών δειγμάτων. Για αυτό το λόγο στο κεφάλαιο που ακολουθεί θα εισαχθούμε στον κόσμο των αντιπαλικών δειγμάτων.

Κεφάλαιο 4: Αντιπαλικά Δείγματα

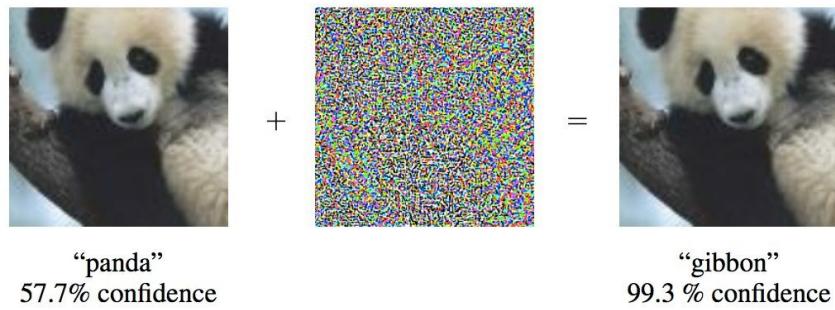
Σε αυτό το κεφάλαιο θα αναλύσουμε το φαινόμενο των αντιπαλικών δειγμάτων, θα δώσουμε έναν ορισμό για αυτά, θα εξηγήσουμε γιατί υπάρχουν, θα αναφερθούμε στα προβλήματα που δημιουργούν, θα περιγράψουμε τις διάφορες ιδιότητες τους, θα επεξηγήσουμε τις διάφορες τεχνικές αντιπαλικής επίθεσης που υπάρχουν και το πώς αυτές κατηγοριοποιούνται καθώς και θα περιγράψουμε προϋπάρχουσες μεθόδους αντιπαλικής επίθεσης που χρησιμοποιούνται για την παραγωγή αντιπαλικών δειγμάτων.

4.1: Εισαγωγή

Τις τελευταίες δεκαετίες έχει γίνει μεγάλη έρευνα στο χώρο των βιοεμπνευσμένων μοντέλων. Είναι ξεκάθαρο πια πως η φύση έχει καταφέρει να αναπτύξει πολύπλοκους και ευέλικτους μηχανισμούς για την επίλυση δύσκολων προβλημάτων. Ένα από τα πιο διάσημα βιοεμπνευσμένα μοντέλα για την ανάπτυξη διακριτικών και γεννητικών μοντέλων [23] αποτελούν τα τεχνητά νευρωνικά δίκτυα [24] αφού είναι ικανά να διαχωρίσουν ικανοποιητικά μη-διαχωρίσιμα δεδομένα καθώς και να προσεγγίσουν τόσο τα ίδια (βάση του θεωρήματος καθολικής προσέγγισης) όσο και την κατανομή πιθανοτήτων που αυτά ακολουθούν. Ωστόσο, ενώ αυτά τα μοντέλα δουλεύουν με αρκετά καλές αποδόσεις, πολλές φορές παρουσιάζουν περίεργες συμπεριφορές που είναι δύσκολο να ερμηνευτούν. Για παράδειγμα, σε ένα τεχνητό νευρωνικό δίκτυο είναι δύσκολο να ερμηνεύσουμε με απόλυτη ακρίβεια το τι έχει μάθει από τα δεδομένα. Γενικότερα, τα τεχνητά νευρωνικά δίκτυα έχουν τυφλά σημεία που ακόμα δεν έχουν εξιχνιαστεί. Ένα παράδειγμα αποτελεί η αδυναμία τους να χειριστούν σωστά τα αντιπαλικά δείγματα [25].

4.2: Ορισμός

Αντιπαλικό δείγμα ονομάζεται ένα τεχνητά κατασκευασμένο δείγμα εισόδου που ενώ η διαφορά του σε σύγκριση με το αντίστοιχο πρωτότυπο από το οποίο έχει προκύψει είναι ανεπαίσθητα διαφορετική, ένα μοντέλο ταξινόμησης το κατηγοριοποιεί σε λανθασμένη κλάση. Θεωρούμε ένα σύνολο από εγγραφές (x_i, y_i) όπου κάθε μια από αυτές αποτελείται από ένα διάνυσμα χαρακτηριστικών $x_i \in X \subseteq R^n$ και την αντίστοιχη του αληθινή κλάση $y_i \in Y$. Υποθέτουμε πως ένας ταξινομητής κλάσεων f έχει μάθει την αντιστοίχηση $f: X \rightarrow Y$. Δοθέντος ενός δείγματος x με την αληθινή του κλάση y είναι δυνατό να κατασκευαστούν δύο ειδών αντιπαλικά δείγματα. Τα στοχευμένα και τα μη-στοχευμένα. Και στις δύο περιπτώσεις, το αντιπαλικό δείγμα \hat{x} προκύπτει προσθέτοντας μια μικρή αντιπαλική διαταραχή (που συνήθως αναφέρεται και ως «θόρυβος») στο x έτσι ώστε να ισχύει η συνθήκη $\|\hat{x} - x\|_p \leq \epsilon$ όπου ϵ ένας αριθμός που καθορίζει το είδος της νόρμας που θέλουμε να εφαρμόσουμε στην διαταραχή και ϵ μία μικρή τιμή που ελέγχει το μέγεθος της διαταραχής. Για το μη-στοχευμένο αντιπαλικό δείγμα θέλουμε να ισχύει η συνθήκη $f(\hat{x}) \neq y$ ενώ για το στοχευμένο η συνθήκη $f(\hat{x}) = \hat{y}$ όπου \hat{y} είναι μια καθορισμένη κλάση διαφορετική από την y .



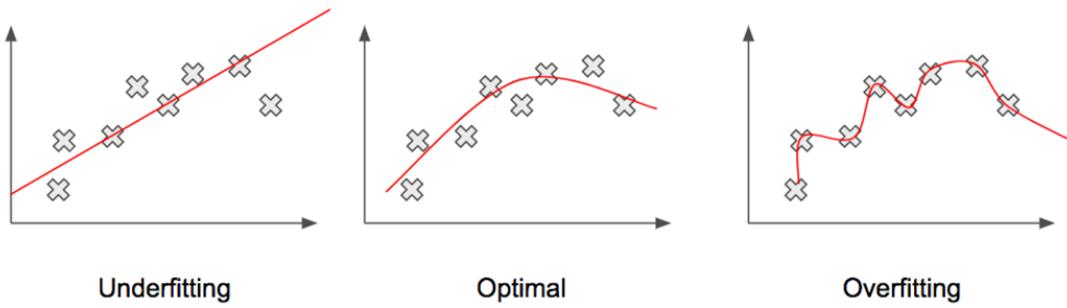
Εικόνα 8: Παράδειγμα αντιπαλικής εικόνας που κατηγοριοποιείται λανθασμένα

4.3: Γιατί Υπάρχουν;

Γενικά, γνωρίζουμε πως όταν ένα μοντέλο ταξινόμησης δεν έχει υπερεκπαιδευτεί και έχει γενικεύσει τότε έχει μάθει την γενικότερη δομή των δεδομένων εκπαίδευσης και όχι τις ιδιοσυγκρασίες αυτών. Αυτό σημαίνει πως το μοντέλο έχει καλή ακρίβεια πρόβλεψης τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα ελέγχου. Επίσης, ένα τέτοιο μοντέλο έχει καλή ακρίβεια πρόβλεψης και σε δεδομένα που αποτελούν μικρομεταβολές της εισόδου. Σε αυτή την περίπτωση ισχύει η ιδιότητα της τοπικής γενίκευσης. Η ιδιότητα της τοπικής γενίκευσης είναι πολύ βασική αφού μας δείχνει το αν οι αποφάσεις ενός μοντέλου επηρεάζονται από μικρομεταβολές της εισόδου. Για παράδειγμα, σε εφαρμογές όρασης υπολογιστών αυτές οι μικρομεταβολές συμβαίνουν στις φωτεινότητες των διαφόρων εικονοστοιχείων. Έτσι, αν τα εικονοστοιχεία μίας εικόνας αλλάξουν ελάχιστα τότε το μοντέλο θα πρέπει να εξακολουθεί να την αναγνωρίζει σωστά. Αυτό γενικότερα είναι κάτι που ισχύει στα περισσότερα σύγχρονα μοντέλα. Ωστόσο, τα αντιπαλικά δείγματα αποτελούν μικρομεταβολές της εισόδου για τα οποία δυστυχώς δεν ισχύει η ιδιότητα της τοπικής γενίκευσης. Επίσης, αυτά δεν προκύπτουν τυχαία αλλά κατασκευάζονται σκόπιμα μέσω κάποιας μεθόδου αντιπαλικής επίθεσης ώστε να κατηγοριοποιούνται εσφαλμένα.

Γιατί όμως να μπορούν να υπάρχουν αντιπαλικά δείγματα; Πως είναι δυνατόν ένα μοντέλο που δεν έχει υπερεκπαιδευτεί και έχει γενικεύσει να κατηγοριοποιεί λανθασμένα τα αντιπαλικά δείγματα που αποτελούν μικρομεταβολές της εισόδου; Μήπως τελικά τα μοντέλα ταξινόμησης δεν μαθαίνουν τόσο καλά όσο νομίζουμε;

Η αρχική υπόθεση που είχε γίνει για την προέλευση των αντιπαλικών δειγμάτων ήταν ότι υπάρχουν λόγω κάποιας μορφής υπερεκπαίδευσης. Γνωρίζουμε ότι όταν ένα μοντέλο είναι υπερεκπαιδευμένο έχει μάθει τις ιδιοσυγκρασίες των δεδομένων και είναι αρκετά ευαίσθητο σε μικρομεταβολές αυτών. Ένα υπερεκπαιδευμένο μοντέλο αποδίδει διάφορες πιθανότητες στο χώρο των δεδομένων εκπαίδευσης ώστε να μεγιστοποιήσει την ακρίβεια πρόβλεψης σε αυτά. Ωστόσο, τις περισσότερες φορές αυτά τα μοντέλα δεν έχουν καλές επιδόσεις στα δεδομένα ελέγχου. Μαθαίνουν δηλαδή με τέτοιο τρόπο που δεν είναι ικανά να γενικεύσουν σε άγνωστα δεδομένα. Έτσι, είναι εύκολο να οδηγηθούν σε λανθασμένη ταξινόμηση για μικρομεταβολές της εισόδου. Όταν το μοντέλο είναι σύνθετο και έχει πολλές παραμέτρους τότε όχι μόνο μπορεί να υπερεκπαιδευτεί αλλά αποδίδει και διάφορες τυχαίες πιθανότητες (οι οποίες οδηγούν σε λανθασμένη κατηγοριοποίηση) σε περιοχές του χώρου για τις οποίες δεν έχουμε δεδομένα εκπαίδευσης.

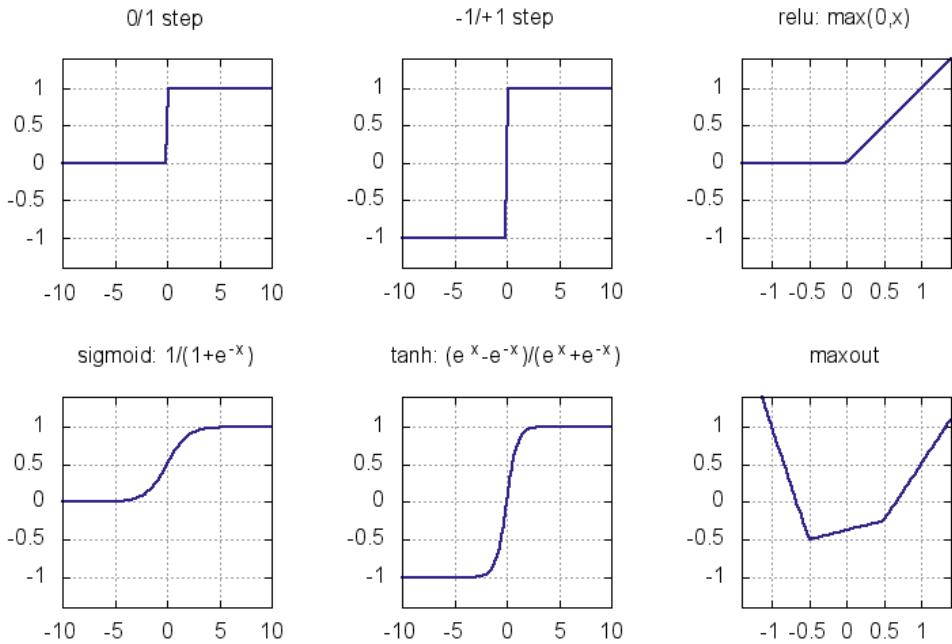


Εικόνα 9: Διάφορες εκδοχές προσαρμογής ενός μοντέλου σε δεδομένα εκπαίδευσης

Έτσι, αν η υπόθεση της υπερεκπαίδευσης ήταν αληθής τότε θα έπρεπε τα αντιπαλικά δείγματα να είναι αντιπαλικά μόνο για τα μοντέλα στα οποία δημιουργήθηκαν και όχι σε άλλα διότι οι πιθανότητες που αποδίδονται από τα διάφορα μοντέλα είναι διαφορετικές και μπορούν να επηρεαστούν από πολλούς παράγοντες. Ωστόσο, αυτό δεν συμβαίνει. Η υπόθεση που είχε γίνει ήταν λανθασμένη γιατί αργότερα αποδείχτηκε πως τα αντιπαλικά δείγματα είναι μεταφέρσιμα μεταξύ διαφορετικών μοντέλων. Άρα, τα αντιπαλικά δείγματα που προκύπτουν από ένα μοντέλο ταξινόμησης δεν υπάρχουν επειδή αυτό αποδίδει «τυχαίες» πιθανότητες κατά την εκπαίδευση. Η αιτία ύπαρξης τους οφείλεται σε κάτι πιο συστηματικό.

Πλέον, γνωρίζουμε πως τα αντιπαλικά δείγματα μπορούν να προκύψουν σε μοντέλα που έχουν γραμμικότητες και λειτουργούν με δείγματα υψηλών διαστάσεων (π.χ. εικόνες). Πολλές και μικρές μεταβολές στα δείγματα εισόδου μπορούν να επηρεάσουν τις αποφάσεις των μοντέλων όταν αυτά έχουν γραμμικότητες. Τα πρώτα πειράματα έγιναν σε απλά γραμμικά μοντέλα που υλοποιούν ένα διακριτικό υπερεπίπεδο διαχώρισης για τον διαχωρισμό δύο κλάσεων και τα αποτελέσματα έδειξαν πως έχουν παθολογική συμπεριφορά σε περιοχές αρκετά μακριά από τα δεδομένα εκπαίδευσης. Πιο συγκεκριμένα, το αποτέλεσμα του εσωτερικού γινομένου $w^T x$ μεταξύ του διανύσματος w και ενός δείγματος x επηρέασε αρνητικά το αποτέλεσμα της ταξινόμησης.

Ωστόσο, τα αντιπαλικά δείγματα μπορούν να εμφανιστούν και σε μη-γραμμικά μοντέλα όταν αυτά έχουν μερικώς γραμμικότητες. Για παράδειγμα, τα τεχνητά νευρωνικά δίκτυα είναι μη-γραμμικά γιατί χρησιμοποιούν μη-γραμμικές συναρτήσεις ενεργοποίησης. Ωστόσο, αρκετές από αυτές τις συναρτήσεις (π.χ. ELU, Sigmoid, ReLU, Tanh, LeakyReLU, Maxout, PReLU) έχουν μερικώς γραμμικότητες. Αυτός είναι ο λόγος που είναι εφικτή η δημιουργία αντιπαλικών δειγμάτων στα τεχνητά νευρωνικά δίκτυα.



Εικόνα 10 : Παραδείγματα συναρτήσεων ενέργοτοποίησης που έχουν γραμμικότητες

Άρα, ο κύριος λόγος για τον οποίο υπάρχουν τα αντιπαλικά δείγματα οφείλεται στο ότι τα διάφορα μοντέλα ταξινόμησης έχουν γραμμικότητες οι οποίες δυσλειτουργούν με δείγματα υψηλών διαστάσεων που προκύπτουν από πολλές μικρομεταβολές. Όσο τα διάφορα μοντέλα ταξινόμησης θα έχουν γραμμικότητες θα είναι εφικτό να παραχθούν νέα αντιπαλικά δείγματα είτε με τις υπάρχουσες μεθόδους είτε με νέες πιο ισχυρές. Το κενό ασφαλείας των μοντέλων ταξινόμησης που δημιουργείται από τα αντιπαλικά δείγματα παραμένει ανοιχτό μέχρι σήμερα και δεν έχει βρεθεί κάποια ουσιαστική και οριστική λύση.

4.4: Συνέπειες

Τα αντιπαλικά δείγματα έφεραν ένα μεγάλο πλήγμα στον χώρο της υπολογιστικής νοημοσύνης. Η ευκολία με την οποία μπορούν να δημιουργηθούν, η δυσκολία στην αντιμετώπιση τους, το μέγεθος της αρνητικής τους επίδρασης στην πλειοψηφία των μοντέλων ταξινόμησης καθώς και η ικανότητα τους να είναι μεταφέρσιμα σε διαφορετικά μοντέλα έστρεψαν το ενδιαφέρον αρκετών ερευνητών.

Κατά μία έννοια τα αντιπαλικά δείγματα αποκάλυψαν την «Αχίλλειο πτέρνα» των μοντέλων ταξινόμησης. Ως απάντηση αυτής της επίθεσης ήταν η αντιπαλική εκπαίδευση [28] ως μέθοδος συστηματοποίησης-γενίκευσης καθώς και οι διάφορες μέθοδοι αντιπαλικής άμυνας [41] που παρέχουν μερικώς προστασία από τα αντιπαλικά δείγματα.

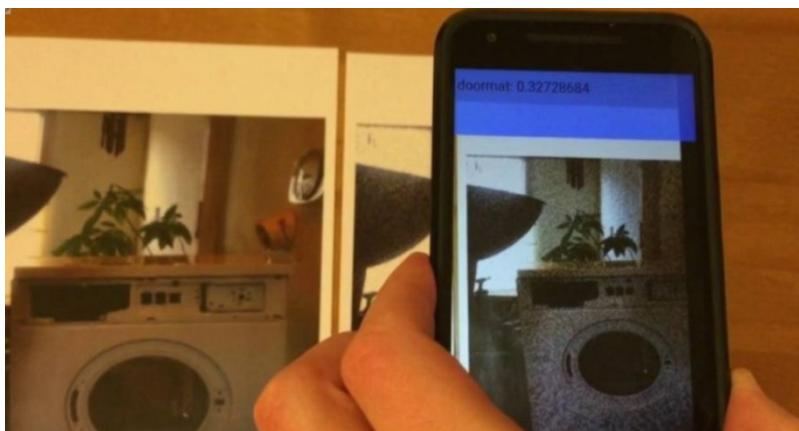
Επίσης, τα αντιπαλικά δείγματα δημιούργησαν μεγάλο προβληματισμό όσο αφορά την αξιοπιστία των διαφόρων μοντέλων ταξινόμησης και κυρίως των σύγχρονων βαθιών τεχνητών νευρωνικών δικτύων. Πλέον, φαίνεται ξεκάθαρα πως εμφανίζουν «Potemkin village» συμπεριφορές. Αρχικά, θεωρούσαμε πως μαθαίνουν σωστά. Ωστόσο, η αδυναμία τους στο να κατηγοριοποιήσουν σωστά τα αντιπαλικά δείγματα άλλαξε την θεώρηση μας για τι πραγματικά μαθαίνουν.

4.5: Μεταφερσιμότητα

Από τις πρώτες προσπάθειες παραγωγής αντιπαλικών δειγμάτων έγινε αντιληπτό πως τα αντιπαλικά δείγματα είναι μεταφέρσιμα μεταξύ διαφορετικών μοντέλων με δύο τρόπους. Ο πρώτος τρόπος αναφέρεται ως «γενίκευση μεταξύ διαφορετικών μοντέλων» (cross-model generalization) και ο δεύτερος ως «γενίκευση μεταξύ διαφορετικών συνόλων εκπαίδευσης» (cross training-set generalization).

Η γενίκευση μεταξύ διαφορετικών μοντέλων συμβαίνει όταν έχουμε μοντέλα που είναι εκπαιδευμένα στο ίδιο σύνολο εκπαίδευσης αλλά διαφέρουν στην αρχιτεκτονική ή στις υπερπαραμέτρους. Ενώ, η γενίκευση μεταξύ διαφορετικών συνόλων εκπαίδευσης συμβαίνει όταν έχουμε μοντέλα ίδιας αρχιτεκτονικής με ίδιες ή διαφορετικές υπερπαραμέτρους αλλά είναι εκπαιδευμένα σε διαφορετικό υποσύνολο κάποιου διαχωρισμένου συνόλου δεδομένων.

Επίσης, παρόλο που τα αντιπαλικά δείγματα δεν εμφανίζονται με φυσικό τρόπο από μόνα τους και κατασκευάζονται τεχνητά με ψηφιακό τρόπο αυτό δεν σημαίνει πως δεν μπορούν να μεταφερθούν στον φυσικό κόσμο. Τελευταίες έρευνες [26] στον χώρο της εικόνας δείχνουν πως αν εκτυπώσουμε κάποιες αντιπαλικές εικόνες και στην συνέχεια τις ψηφιοποιήσουμε με κάποια κάμερα, οι νέες εικόνες θα παραμείνουν μερικώς αντιπαλικές και θα είναι ικανές να μπερδέψουν ακόμη και σύγχρονους ταξινομητές εικόνας.



Εικόνα 11: Αντιπαλικό δείγμα στον φυσικό κόσμο που κατηγοριοποιείται λανθασμένα

4.6: Ταξινομία

Τα διάφορα αντιπαλικά δείγματα, οι αντιπαλικές διαταραχές, οι αντιπαλικές επιθέσεις καθώς και τα δείγματα εισόδου μπορούν να κατηγοριοποιηθούν σε διάφορες κατηγορίες ανάλογα την περίπτωση. Παρακάτω θα μελετήσουμε αυτές τις περιπτώσεις.

4.6.1: Δείγμα Εισόδου

Οι διάφοροι μέθοδοι αντιπαλικής επίθεσης για να λειτουργήσουν χρειάζονται κάποιο αρχικό δείγμα στην είσοδο τους πριν το εξελίξουν μέσω στοχευμένων μεταβολών σε αντιπαλικό. Το δείγμα εισόδου μπορεί να είναι τυχαίο ή μη-τυχαίο. Μη-τυχαίο είναι ένα δείγμα που προέρχεται από το σύνολο δειγμάτων εκπαίδευσης ή ελέγχου. Ενώ, τυχαίο

είναι ένα δείγμα όταν είναι τυχαίος θόρυβος που ακολουθεί κάποια συγκεκριμένη κατανομή πιθανοτήτων (π.χ. Γκαουσιανή, ομοιόμορφη).



Εικόνα 12: Παραδείγματα τυχαίου και μη-τυχαίου δείγματος εισόδου

Διάφορες πειραματικές δοκιμές έδειξαν πως οι μέθοδοι αντιπαλικής επίθεσης μπορούν να λειτουργήσουν ικανοποιητικά και με τις δύο περιπτώσεις. Επίσης, σύγχρονες μέθοδοι αντιπαλικής επίθεσης που βασίζονται σε γεννητικά αντιπαλικά δίκτυα [40] για την παραγωγή αντιπαλικών δειγμάτων χρησιμοποιούν ως είσοδο αποκλειστικά τυχαίο θόρυβο χαμηλών διαστάσεων και παράγουν στην έξοδο την αντιπαλική διαταραχή που μετατρέπει ένα δείγμα σε αντιπαλικό.

4.6.2: Συχνότητα Επίθεσης

Οι διάφοροι μέθοδοι αντιπαλικής επίθεσης που κατασκευάζουν αντιπαλικά δείγματα μπορούν να παράγουν ένα αντιπαλικό δείγμα μέσω ενός υπολογισμού που εκτελείται είτε μόνο μία φορά είτε περισσότερες. Έτσι, αυτές οι μέθοδοι χωρίζονται σε δύο κατηγορίες, τις μη-επαναληπτικές και τις επαναληπτικές. Αρχικά, οι πρώτες μέθοδοι που εμφανίστηκαν ήταν μη-επαναληπτικές με πολύ καλές επιδόσεις και χρόνους. Αυτές με έναν μόνο υπολογισμό μεταβάλλουν το δείγμα της εισόδου και το μετατρέπουν σε αντιπαλικό. Επίσης, λόγω της ταχύτητας τους φάνηκαν αρκετά χρήσιμες σε περιπτώσεις αντιπαλικής εκπαίδευσης όπου ένα μοντέλο δεν εκπαιδεύεται μόνο σε κανονικά δείγματα αλλά και σε αντιπαλικά τα οποία κατασκευάζονται την στιγμή της μάθησης. Αργότερα, εμφανίστηκαν και οι επαναληπτικές μέθοδοι οι οποίες είναι ικανές να δημιουργούν πιο αποδοτικά αντιπαλικά δείγματα. Επίσης, σύγχρονοι μέθοδοι αντιπαλικής επίθεσης [30] που βασίζονται σε γεννητικά αντιπαλικά δίκτυα για την παραγωγή αντιπαλικών δειγμάτων είναι μη-επαναληπτικοί και χρησιμοποιούν τον εκπαιδευμένο γεννητιστή του δικτύου για την παραγωγή της αντιπαλικής διαταραχής με ένα απλό πέρασμα.

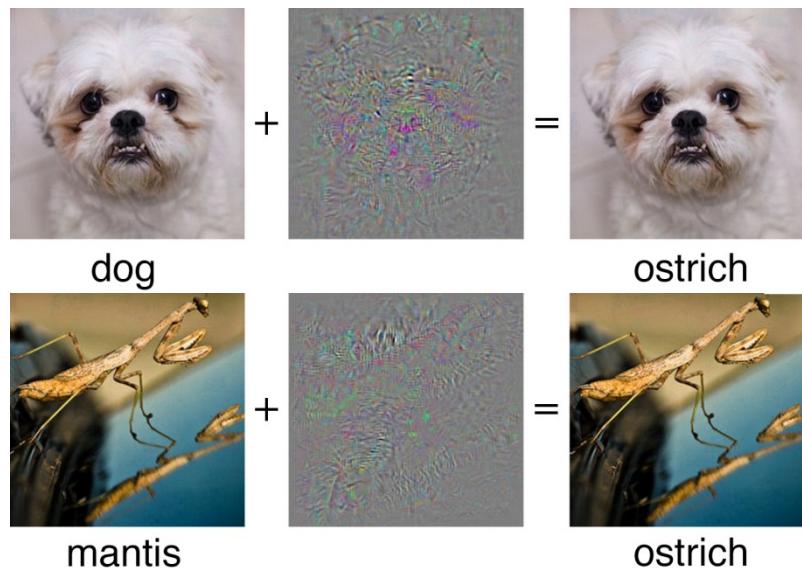
4.6.3: Ειδίκευση Αντιπαλικού Δείγματος

Τα αντιπαλικά δείγματα μπορούν να διακριθούν ανάλογα με την ειδίκευση τους σε στοχευμένα και μη-στοχευμένα. Για κάποιο δείγμα εισόδου x με την αληθινή του κλάση y είναι εφικτό να κατασκευαστεί τόσο ένα στοχευμένο αντιπαλικό δείγμα \hat{x} ώστε να ισχύει η συνθήκη $f(\hat{x}) = \hat{y}$ όπου f ένα μοντέλο ταξινόμησης και \hat{y} μια συγκεκριμένη κλάση διαφορετική της y όσο και ένα μη-στοχευμένο αντιπαλικό δείγμα \tilde{x} ώστε να ισχύει η συνθήκη $f(\tilde{x}) \neq \hat{y}$ όπου y η αληθινή κλάση του δείγματος εισόδου x . Τα μη-στοχευμένα αντιπαλικά δείγματα συγκριτικά με τα στοχευμένα είναι πολύ πιο εύκολο να βρεθούν διότι

ο χώρος λύσεων είναι πολύ μεγαλύτερος και η ελευθερία κινήσεων της μεθόδου αντιπαλικής επίθεσης κατά την βελτιστοποίηση μεγαλύτερη.

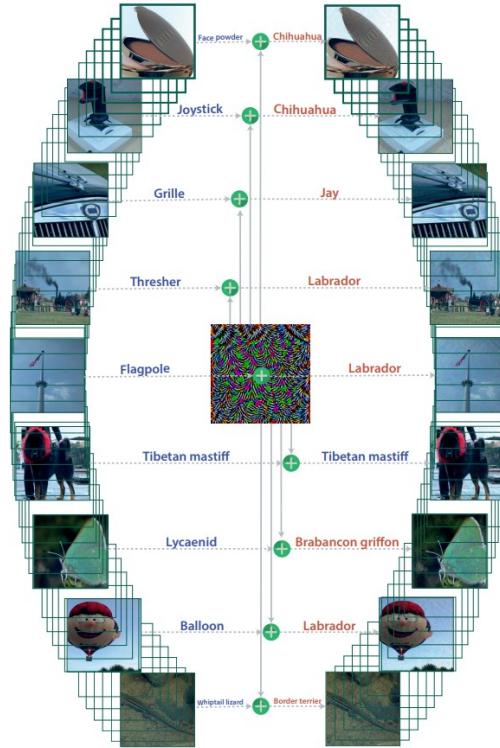
4.6.4: Δράση Αντιπαλικής Διαταραχής

Οι πλειοψηφία των μεθόδων αντιπαλικής επίθεσης προσπαθούν να αναζητήσουν την ελάχιστη αντιπαλική διαταραχή που χρειάζεται να προστεθεί σε ένα συγκεκριμένο δείγμα εισόδου ώστε αυτό να μετατραπεί σε αντιπαλικό. Αυτή η αντιπαλική διαταραχή ονομάζεται ειδική ή τοπική.



Εικόνα 13: Παραδείγματα τοπικών αντιπαλικών διαταραχών

Ωστόσο, υπάρχει και ένας μικρός αριθμός από μεθόδους αντιπαλικής επίθεσης που στοχεύουν στην αναζήτηση αντιπαλικής διαταραχής η οποία όταν προστεθεί σε ένα σύνολο δειγμάτων εισόδου τότε να μετατραπούν όλα ή όσα είναι εφικτό σε αντιπαλικά. Αυτή η αντιπαλική διαταραχή ονομάζεται γενική ή καθολική.



Εικόνα 14: Παράδειγμα καθολικής αντιπαλικής διαταραχής

4.6.5: Μέγεθος Αντιπαλικής Διαταραχής

Στις περισσότερες μεθόδους αντιπαλικής επίθεσης επιθυμούμε το μέγεθος της αντιπαλικής διαταραχής να είναι όσο το δυνατόν μικρότερο ώστε το αντιπαλικό δείγμα να είναι ανεπαίσθητα διαφορετικό από το δείγμα εισόδου. Ένας από τους πιο κοινούς τρόπους υπολογισμού του μεγέθους της αντιπαλικής διαταραχής είναι μέσω της l_p -νόρμας και ορίζεται ως:

$$\|x\|_p = (|x_1|^p + |x_2|^p + \dots + |x_n|^p)^{\frac{1}{p}} = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$$

Παρακάτω, ακολουθούν οι πιο γνωστές περιπτώσεις της l_p -νόρμας που χρησιμοποιούνται στις διάφορες μεθόδους αντιπαλικής επίθεσης για τον υπολογισμό και τον περιορισμό του μεγέθους της αντιπαλικής διαταραχής:

- Η μηδενική νόρμα (l_0 -νόρμα) ορίζεται ως $\|x\|_0 = \#\{i | x_i \neq 0\}$
- Η Manhattan νόρμα (l_1 -νόρμα) ορίζεται ως $\|x\|_1 = |x_1| + |x_2| + \dots + |x_n|$
- Η Ευκλείδεια νόρμα (l_2 -νόρμα) ορίζεται ως $\|x\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \dots + |x_n|^2}$
- Η Chebyshev νόρμα (l_∞ -νόρμα) ορίζεται ως $\|x\|_\infty = \max(|x_1|, |x_2|, \dots, |x_n|)$

4.6.6: Είδος Αντιπαλικής Διαστρέβλωσης

Η αντιπαλική διαστρέβλωση αναφέρεται στο είδος της επίθεσης ανάλογα με το αν το παραγόμενο αντιπαλικό δείγμα κατηγοριοποιείται ως «ψευδώς θετικό» ή «ψευδώς αρνητικό» από ένα μοντέλο ταξινόμησης δύο κλάσεων. Πιο αναλυτικά, αν θεωρήσουμε ένα μοντέλο ταξινόμησης που έχει εκπαιδευτεί στο να διαχωρίζει δύο κλάσεις τότε μπορούν να υπάρχουν δύο ειδών αντιπαλικής διαστρέβλωσης. Αυτή που παράγει ένα «ψευδώς θετικό» δείγμα το οποίο ενώ είναι αρνητικό κατηγοριοποιείται εσφαλμένα ως θετικό (Σφάλμα Τύπου 1) και αυτή που παράγει ένα «ψευδώς αρνητικό» δείγμα το οποίο ενώ είναι θετικό κατηγοριοποιείται εσφαλμένα ως αρνητικό (Σφάλμα Τύπου 1).

4.6.7: Γνώση Αντίπαλου Μοντέλου

Η γνώση που διαθέτουμε για έναν μοντέλο ταξινόμησης που θέλουμε να επιτεθούμε δεν είναι πάντα ίδια. Έτσι, οι διάφορες επιθέσεις μπορούν να κατηγοριοποιηθούν ανάλογα με το βάθος αυτής της γνώσης σε επιθέσεις «λευκού κουτιού» και σε επιθέσεις «μαύρου κουτιού».

Σε μία επίθεση «λευκού κουτιού» γνωρίζουμε τα πάντα σχετικά με το μοντέλο ταξινόμησης που πρόκειται να επιτεθούμε. Πιο συγκεκριμένα, μπορεί να γνωρίζουμε τα δεδομένα και τις υπερπαραμέτρους της εκπαίδευσης καθώς και την αρχιτεκτονική, την τοπολογία και τις παραμέτρους του μοντέλου.

Αντίθετα, σε μία επίθεση «μαύρου κουτιού» δεν γνωρίζουμε τίποτα για το μοντέλο ταξινόμησης που θέλουμε να επιτεθούμε. Το μόνο που γνωρίζουμε είναι η μορφή των δεδομένων εισόδου καθώς και οι πιθανότητες που εξάγει στην έξοδο για κάθε κλάση. Κάποιες φορές δεν εξάγονται ούτε οι πιθανότητες αλλά μόνο η τελική κλάση αναγνώρισης. Πολλά εταιρικά μοντέλα ταξινόμησης θεωρούνται «μαύρα κουτιά» διότι είναι προσβάσιμα μόνο για χρήση μέσω κάποιας διαδικτυακής υπηρεσίας.

Οι διάφοροι μέθοδοι αντιπαλικής επίθεσης είναι σχεδιασμένοι για επιθέσεις «λευκού κουτιού» και αυτό σημαίνει πως πρέπει να γνωρίζουμε αρκετά για τα μοντέλα που κάνουμε επίθεση. Έτσι, η παραγωγή αντιπαλικών δειγμάτων σε επιθέσεις «λευκού κουτιού» είναι εύκολη.

Ωστόσο, πως μπορούμε να κάνουμε μια επίθεση «μαύρου κουτιού»;

Επειδή γνωρίζουμε πως τα αντιπαλικά δείγματα έχουν μεταφερσιμότητα ανάμεσα στα διάφορα μοντέλα αυτό που μπορούμε να κάνουμε σε μια επίθεση «μαύρου κουτιού» είναι να τα παράγουμε αρχικά σε κάποιο γνωστό μοντέλο με επίθεση «λευκού κουτιού» και στην συνέχεια να τα δοκιμάσουμε στο μοντέλο για το οποίο δεν γνωρίζουμε τίποτα ευελπιστώντας πως θα έχουν και σε αυτό αρνητική επίδραση.

4.7: Προϋπάρχουσες Μέθοδοι

Για την παραγωγή αντιπαλικών δειγμάτων έχουν αναπτυχθεί αρκετές μέθοδοι αντιπαλικής επίθεσης τόσο επαναληπτικές όσο και μη-επαναληπτικές. Αυτές που θα περιγράψουμε παρακάτω είναι αυτές που εμφανίζονται πιο συχνά στην βιβλιογραφία. Επίσης, κάποιες

από αυτές τις μεθόδους δεν αποτελούν πλέον απειλή επειδή έχουν αναπτυχθεί διάφοροι μέθοδοι αντιπαλικής άμυνας [41].

4.7.1: Μη-επαναληπτικές Μέθοδοι

4.7.1.1: Fast Gradient Sign Method (FGSM)

Η μέθοδος αυτή [27] χρησιμοποιείται για την παραγωγή μη-στοχευμένων αντιπαλικών δειγμάτων. Συγκριτικά με την L-BFGS-B η μέθοδος αυτή έχει δύο βασικές διαφορές. Η πρώτη είναι ότι έχει σχεδιαστεί με γνώμονα την ταχύτητα με αποτέλεσμα να πετυχαίνει πολύ καλύτερους χρόνους και το δεύτερο ότι ως μετρική απόστασης ανάμεσα στην αρχική και στην αντιπαλική εικόνα χρησιμοποιεί την l_∞ -νόρμα.

Βασική προϋπόθεση της μεθόδου ήταν το να είναι γρήγορη (π.χ. για εφαρμογές αντιπαλικής εκπαίδευσης) και όχι το να βρίσκει την βέλτιστη λύση. Άρα, τα αντιπαλικά δείγματα που προκύπτουν από αυτή την μέθοδο δεν είναι αυτά με την μικρότερη αντιπαλική διαταραχή. Για την παραγωγή του αντιπαλικού δείγματος χρησιμοποιούνται οι παρακάτω εξισώσεις:

$$\hat{x} = x + n \quad \text{και} \quad n = \varepsilon \cdot sign(\nabla_x J_f(x, y))$$

όπου x η αρχική εικόνα, \hat{x} η αντιπαλική εικόνα, n η αντιπαλική διαταραχή, ε ένας μικρός αριθμός, J_f η συνάρτηση σφάλματος ενός ταξινομητή f και y η αληθινή κλάση του δείγματος x .

Η μέθοδος αρχικά χρησιμοποιεί το πρόσημο της κλίσης της συνάρτησης σφάλματος για να βρει προς ποια κατεύθυνση θα πρέπει να μεταβληθούν οι φωτεινότητες των εικονοστοιχείων της εικόνας x ώστε να **μεγιστοποιηθεί** η συνάρτηση σφάλματος J_f και στην συνέχεια ανάλογα με την τιμή του ε κάνει ένα μικρό βήμα προς εκείνη την κατεύθυνση αλλοιώνοντας τα εικονοστοιχεία.

Όταν ο ταξινομητής f είναι νευρωνικό δίκτυο τότε ο υπολογισμός της κλίσης της συνάρτησης σφάλματος J_f μπορεί να γίνει με τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

4.7.1.2: One-step Target Class Method (OTCM)

Η μέθοδος αυτή [28] χρησιμοποιείται για την παραγωγή στοχευμένων αντιπαλικών δειγμάτων και βασίζεται αρκετά στις αρχές λειτουργίας της μεθόδου FGSM. Για την παραγωγή του αντιπαλικού δείγματος χρησιμοποιούνται οι παρακάτω εξισώσεις:

$$\hat{x} = x - n \quad \text{και} \quad n = \varepsilon \cdot sign(\nabla_x J_f(x, \hat{y}))$$

όπου x η αρχική εικόνα, \hat{x} η αντιπαλική εικόνα, n η αντιπαλική διαταραχή, ε ένας μικρός αριθμός, J_f η συνάρτηση σφάλματος ενός ταξινομητή f και \hat{y} η λανθασμένη κλάση στην οποία στοχεύουμε.

Παρατηρούμε πως οι μόνες διαφορές σε σχέση με την μέθοδο FGSM είναι ότι η μέθοδος αυτή αφαιρεί την αντιπαλική διαταραχή αντί να την προσθέτει και ότι χρησιμοποιεί μια λανθασμένη κλάση στην συνάρτηση σφάλματος J_f .

Η μέθοδος αρχικά χρησιμοποιεί το πρόσημο της κλίσης της συνάρτησης σφάλματος για να βρει προς ποια κατεύθυνση θα πρέπει να μεταβληθούν οι φωτεινότητες των εικονοστοιχείων της εικόνας x ώστε να **ελαχιστοποιηθεί** η συνάρτηση σφάλματος J_f και στην συνέχεια ανάλογα με την τιμή του ϵ κάνει ένα μικρό βήμα προς εκείνη την κατεύθυνση αλλοιώνοντας τα εικονοστοιχεία.

Όταν ο ταξινομητής f είναι νευρωνικό δίκτυο τότε ο υπολογισμός της κλίσης της συνάρτησης σφάλματος J_f μπορεί να γίνει με τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

4.7.1.3: One-step Least Likely Class Method (OLLCM)

Η μέθοδος αυτή [26, 28] αποτελεί ειδική περίπτωση της μεθόδου OTCM. Κατά την παραγωγή ενός στοχευμένου αντιπαλικού δείγματος χρησιμοποιούμε ως επιθυμητό στόχο \hat{y} την κλάση για την οποία ο ταξινομητής μας δίνει την μικρότερη πιθανότητα:

$$\hat{y}_{LL} = \underset{y}{\arg \min} \{ p(y|x) \}$$

Σε περίπτωση όπου ο ταξινομητής είναι εκπαιδευμένος σε ένα μεγάλο αριθμό κλάσεων και υπάρχουν κλάσεις που είναι παρόμοιες (π.χ. στο σύνολο δεδομένων ImageNet) είναι πολύ πιθανό με τις διάφορες μεθόδους παραγωγής μη-στοχευμένων αντιπαλικών δειγμάτων να παράγουμε αντιπαλικά δείγματα που δεν έχουν αρκετό νόημα. Για παράδειγμα, θα μπορούσε για μία εικόνα σκύλου της κλάσης «Dobermann» να παραχθεί ένα μη-στοχευμένο αντιπαλικό δείγμα που να κατηγοριοποιείται ως «Mini Pinscher» (παρόμοια ράτσα σκύλου με την ράτσα «Dobermann») από τον ταξινομητή.

Η μέθοδος αυτή επιλύει αυτό το πρόβλημα επιλέγοντας αιτιοκρατικά ως επιθυμητό στόχο την κλάση για την οποία ο ταξινομητής μας δίνει την μικρότερη πιθανότητα. Με αυτό το τρόπο τα αντιπαλικά δείγματα που παράγονται αποτελούν μεγαλύτερα σφάλματα και έτσι επιφέρουν μεγαλύτερο ποσοστό λανθασμένης κατηγοριοποίησης αφού σε έναν καλά εκπαιδευμένο ταξινομητή η αληθινή κλάση ενός δείγματος είναι αρκετά διαφορετική από την λιγότερο πιθανή κλάση. Έχει δειχθεί πως η τεχνική αυτή δίνει καλύτερα αποτελέσματα ακόμη και για μικρές τιμές της παραμέτρου ϵ .

4.7.1.4: Fast Gradient Value Method (FGVM)

Η μέθοδος αυτή [29] χρησιμοποιείται για την παραγωγή αντιπαλικών δειγμάτων και βασίζεται αρκετά στις αρχές λειτουργίας των μεθόδων FGSM και OTCM. Ωστόσο, έχει δύο ουσιαστικές διαφορές με αυτές. Πρώτον, δεν χρησιμοποιεί την συνάρτηση πρόσημου και δεύτερον δεν εφαρμόζει κάποιον περιορισμό στις φωτεινότητες των εικονοστοιχείων. Έτσι, δημιουργεί αντιπαλικά δείγματα μεγαλύτερης τοπικής διαφοράς. Για την παραγωγή του αντιπαλικού δείγματος χρησιμοποιούνται οι παρακάτω εξισώσεις:

- **Στοχευμένο αντιπαλικό δείγμα**

$$\hat{\mathbf{x}} = \mathbf{x} - \mathbf{n} \quad \text{και} \quad \mathbf{n} = a \cdot \nabla_{\mathbf{x}} J_f(\mathbf{x}, \hat{\mathbf{y}})$$

- Μη-στοχευμένο αντιπαλικό δείγμα

$$\hat{\mathbf{x}} = \mathbf{x} + \mathbf{n} \quad \text{και} \quad \mathbf{n} = a \cdot \nabla_{\mathbf{x}} J_f(\mathbf{x}, y)$$

όπου \mathbf{x} η αρχική εικόνα, $\hat{\mathbf{x}}$ η αντιπαλική εικόνα, a ένας αριθμός που λειτουργεί ως παράγοντας κλιμάκωσης της κλίσης της συνάρτησης σφάλματος J_f του ταξινομητή f , γη η αληθινή κλάση του δείγματος \mathbf{x} και $\hat{\mathbf{y}}$ η λανθασμένη κλάση στην οποία στοχεύουμε.

Το ότι δεν χρησιμοποιείται κάποιος περιορισμός για τις φωτεινότητες των εικονοστοιχείων δεν σημαίνει απαραίτητα πως τα αποτελέσματα αυτής της μεθόδου είναι χειρότερα αφού έχει δειχθεί πως με την κατάλληλη ρύθμιση του παράγοντα κλιμάκωσης α μπορούμε να βρούμε εξίσου καλά αποτελέσματα.

Όταν ο ταξινομητής f είναι νευρωνικό δίκτυο τότε ο υπολογισμός της κλίσης της συνάρτησης σφάλματος J_f μπορεί να γίνει με τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

4.7.1.5: Adversarial GAN Attack (AdvGAN)

Η μέθοδος αυτή [30] χρησιμοποιείται για την παραγωγή αντιπαλικών δειγμάτων καθώς είναι αρκετά αποδοτική ακόμη και σε σύγχρονα νευρωνικά μοντέλα ταξινόμησης. Η παραγωγή του αντιπαλικού δείγματος δεν γίνεται με κάποια μέθοδο βελτιστοποίησης αλλά με την βοήθεια ενός γεννητικού αντιπαλικού δικτύου. Σε αυτό το δίκτυο ο διαχωριστής υλοποιείται ως συνελικτικό νευρωνικό δίκτυο και ο γεννητιστής ως πλήρως διασυνδεδεμένο νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης.

Το γεννητικό αντιπαλικό δίκτυο εκπαιδεύεται κατάλληλα με μια συνολική συνάρτηση σφάλματος η οποία ενσωματώνει το σφάλμα του διαχωριστή, το σφάλμα του μοντέλου ταξινόμησης που προσπαθούμε να επιτεθούμε καθώς και ένα σφάλμα Hinge που λειτουργεί με την l_2 -νόρμα της αντιπαλικής διαταραχής. Όταν το γεννητικό αντιπαλικό δίκτυο εκπαιδευτεί τότε ο γεννητιστής είναι ικανός να παράγει την ελάχιστη αντιπαλική διαταραχή που χρειάζεται να προστεθεί σε μία εικόνα εισόδου ώστε αυτή να μετατραπεί σε αντιπαλική.

Σε αντίθεση με άλλες μεθόδους όπου το μοντέλο ταξινόμησης είναι υποχρεωτικό κάθε φορά που θέλουμε να παράγουμε ένα αντιπαλικό δείγμα, η μέθοδος αυτή από την στιγμή που ο γεννητιστής εκπαιδευτεί είναι ικανή να παράγει αντιπαλικά δείγματα μόνο με αυτόν. Η μέθοδος αυτή είναι και πιο γρήγορη αφού ο γεννητιστής μπορεί να παράγει την αντιπαλική διαταραχή σε ένα απλό πέρασμα χωρίς να είναι απαραίτητος ο υπολογισμός της κλίσης της συνάρτησης σφάλματος του ταξινομητή. Συγκριτικά με άλλες μη-επαναληπτικές μεθόδους η μέθοδος αυτή πετυχαίνει καλύτερα ποσοστά λανθασμένης κατηγοριοποίησης.

4.7.2: Επαναληπτικές Μέθοδοι

4.7.2.1: Box-constrained Limited-memory BFGS Attack (L-BFGS-B)

Η μέθοδος αυτή [25] ήταν η πρώτη που χρησιμοποιήθηκε για την παραγωγή στοχευμένων αντιπαλικών δειγμάτων και έχει δοκιμαστεί τόσο σε γραμμικά όσο και σε μη-γραμμικά μοντέλα. Ωστόσο, έχει μεγάλη χρονική πολυπλοκότητα με αποτέλεσμα να είναι απαγορευτική σε εφαρμογές πραγματικού χρόνου. Η μέθοδος παίρνει στην είσοδο μία εικόνα x και παράγει στην έξοδο μια αντιπαλική εικόνα \hat{x} η οποία είναι παρόμοια με την x αλλά κατηγοριοποιείται λανθασμένα στην κλάση \hat{y} από έναν ταξινομητή f . Η αντιπαλική εικόνα \hat{x} προκύπτει ως $\hat{x} = x + n$ όπου n η αντιπαλική διαταραχή που καλείται να βρει η μέθοδος. Το πρόβλημα προς επίλυση για την παραγωγή του αντιπαλικού δείγματος είναι το εξής:

$$\underset{n}{\text{minimize}} \quad c \cdot \|n\| + J_f(\hat{x}, \hat{y}) \quad \text{έτσι ώστε} \quad \hat{x} \in [0, 1]^n$$

όπου J_f η συνάρτηση σφάλματος του ταξινομητή f και c ένας θετικός αριθμός. Μία πολύ γνωστή συνάρτηση σφάλματος που χρησιμοποιείται συχνά σε προβλήματα ταξινόμησης δύο ή περισσότερων κλάσεων είναι η διεντροπία. Με την μέθοδο L-BFGS-B προσεγγίζεται ο ελαχιστοποιητής n . Επίσης, χρησιμοποιείται η μέθοδος αναζήτησης γραμμής για την εύρεση της μικρότερης σταθεράς $c > 0$ ώστε ο ελαχιστοποιητής n να ικανοποιεί την σχέση $f(\hat{x}) = \hat{y}$.

4.7.2.2: Basic Iterative Method (BIM)

Η μέθοδος αυτή [26] χρησιμοποιείται για την παραγωγή μη-στοχευμένων αντιπαλικών δειγμάτων και αποτελεί την επαναληπτική εκδοχή της μεθόδου FGSM. Συγκεκριμένα, σε αυτήν εφαρμόζεται ο υπολογισμός της FGSM πολλαπλές φορές με μικρό βήμα κάθε φορά. Επίσης, έχει δειχθεί πως η μέθοδος αυτή σε επιθέσεις «λευκού κουτιού» μπορεί να επιτύχει μεγαλύτερο ποσοστό λανθασμένης κατηγοριοποίησης από την μέθοδο FGSM. Ωστόσο, τα αντιπαλικά δείγματα της μεθόδου FGSM είναι πιο μεταφέρσιμα και έχουν μεγαλύτερη αντοχή σε μετασχηματισμούς. Για την παραγωγή του αντιπαλικού δείγματος χρησιμοποιούνται οι παρακάτω εξισώσεις:

$$\hat{x}_0 = x \quad \text{και} \quad \hat{x}_{i+1} = \text{clip}_{[x-\varepsilon, x+\varepsilon]}(\hat{x}_i + \alpha \cdot \text{sign}(\nabla_x J_f(\hat{x}_i, y)))$$

όπου x η αρχική εικόνα, \hat{x}_i η αντιπαλική εικόνα στην επανάληψη i , α το μέγεθος του βήματος, ε ένας μικρός αριθμός, $\text{clip}_{[x-\varepsilon, x+\varepsilon]}$ η αποκοπή και ο περιορισμός των φωτεινοτήτων των εικονοστοιχείων εντός της γειτονιάς $[x - \varepsilon, x + \varepsilon]$ στα ενδιάμεσα αποτελέσματα της κάθε επανάληψης, J_f η συνάρτηση σφάλματος ενός ταξινομητή f και y η αληθινή κλάση του δείγματος x .

Όταν ο ταξινομητής f είναι νευρωνικό δίκτυο τότε ο υπολογισμός της κλίσης της συνάρτησης σφάλματος J_f μπορεί να γίνει με τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

4.7.2.3: Iterative Target Class Method (ITCM)

Η μέθοδος αυτή [26] χρησιμοποιείται για την παραγωγή στοχευμένων αντιπαλικών δειγμάτων και αποτελεί την επαναληπτική εκδοχή της μεθόδου OTCM. Συγκεκριμένα, σε αυτήν εφαρμόζεται ο υπολογισμός της OTCM πολλαπλές φορές με μικρό βήμα κάθε φορά. Επίσης, έχει δειχθεί πως η μέθοδος αυτή σε επιθέσεις «λευκού κουτιού» μπορεί να επιτύχει μεγαλύτερο ποσοστό λανθασμένης κατηγοριοποίησης από την μέθοδο OTCM. Ωστόσο, τα αντιπαλικά δείγματα της μεθόδου OTCM είναι πιο μεταφέρσιμα και έχουν μεγαλύτερη αντοχή σε μετασχηματισμούς. Για την παραγωγή του αντιπαλικού δείγματος χρησιμοποιούνται οι παρακάτω εξισώσεις:

$$\hat{x}_0 = x \quad \text{και} \quad \hat{x}_{i+1} = clip_{[x-\varepsilon, x+\varepsilon]}(\hat{x}_i - \alpha \cdot sign(\nabla_x J_f(\hat{x}_i, \hat{y})))$$

όπου x η αρχική εικόνα, \hat{x}_i η αντιπαλική εικόνα στην επανάληψη i , α το μέγεθος του βήματος, ε ένας μικρός αριθμός, $clip_{[x-\varepsilon, x+\varepsilon]}$ η αποκοπή και ο περιορισμός των φωτεινοτήτων των εικονοστοιχείων εντός της γειτονιάς $[x - \varepsilon, x + \varepsilon]$ στα ενδιάμεσα αποτελέσματα της κάθε επανάληψης, J_f η συνάρτηση σφάλματος ενός ταξινομητή f και \hat{y} η λανθασμένη κλάση στην οποία στοχεύουμε.

Όταν ο ταξινομητής f είναι νευρωνικό δίκτυο τότε ο υπολογισμός της κλίσης της συνάρτησης σφάλματος J_f μπορεί να γίνει με τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

4.7.2.4: Iterative Least Likely Class Method (ILLCM)

Η μέθοδος αυτή [26] αποτελεί ειδική περίπτωση της μεθόδου ITCM. Κατά την παραγωγή ενός στοχευμένου αντιπαλικού δείγματος χρησιμοποιούμε ως επιθυμητό στόχο \hat{y} την κλάση για την οποία ο ταξινομητής μας δίδει την μικρότερη πιθανότητα:

$$\hat{y}_{LL} = \underset{y}{\arg \min} \{ p(y|x) \}$$

Σε περίπτωση όπου ο ταξινομητής είναι εκπαιδευμένος σε ένα μεγάλο αριθμό κλάσεων και υπάρχουν κλάσεις που είναι παρόμοιες (π.χ. στο σύνολο δεδομένων ImageNet) είναι πολύ πιθανό με τις διάφορες μεθόδους παραγωγής μη-στοχευμένων αντιπαλικών δειγμάτων να παράγουμε αντιπαλικά δείγματα που δεν έχουν αρκετό νόημα. Για παράδειγμα, θα μπορούσε για μία εικόνα σκύλου της κλάσης «Dobermann» να παραχθεί ένα μη-στοχευμένο αντιπαλικό δείγμα που να κατηγοριοποιείται ως «Mini Pinscher» (παρόμοια ράτσα σκύλου με την ράτσα «Dobermann») από τον ταξινομητή.

Η μέθοδος αυτή επιλύει αυτό το πρόβλημα επιλέγοντας αιτιοκρατικά ως επιθυμητό στόχο την κλάση για την οποία ο ταξινομητής μας δίδει την μικρότερη πιθανότητα. Με αυτό το τρόπο τα αντιπαλικά δείγματα που παράγονται αποτελούν μεγαλύτερα σφάλματα και έτσι επιφέρουν μεγαλύτερο ποσοστό λανθασμένης κατηγοριοποίησης αφού σε έναν καλά εκπαιδευμένο ταξινομητή η αληθινή κλάση ενός δείγματος είναι αρκετά διαφορετική από την λιγότερο πιθανή κλάση. Έχει δειχθεί πως η τεχνική αυτή δίδει καλύτερα αποτελέσματα ακόμη και για μικρές τιμές της παραμέτρου ε .

4.7.2.5: Iterative Fast Gradient Value Method (IFGVM)

Η μέθοδος αυτή [26, 29] χρησιμοποιείται για την παραγωγή αντιπαλικών δειγμάτων και αποτελεί την επαναληπτική εκδοχή της μεθόδου FGVM. Συγκεκριμένα, σε αυτήν εφαρμόζεται ο υπολογισμός της FGVM πολλαπλές φορές με μικρό βήμα κάθε φορά. Επίσης, έχει δειχθεί πως η μέθοδος αυτή σε επιθέσεις «λευκού κουτιού» μπορεί να επιτύχει μεγαλύτερο ποσοστό λανθασμένης κατηγοριοποίησης από την μέθοδο FGVM. Ωστόσο, τα αντιπαλικά δείγματα της μεθόδου FGVM είναι πιο μεταφέρσιμα και έχουν μεγαλύτερη αντοχή σε μετασχηματισμούς. Για την παραγωγή του αντιπαλικού δείγματος χρησιμοποιούνται οι παρακάτω εξισώσεις:

- Στοχευμένο αντιπαλικό δείγμα

$$\hat{x}_0 = x \text{ και } \hat{x}_{i+1} = clip_{[x-\varepsilon, x+\varepsilon]}(\hat{x}_i - \alpha \cdot \nabla_x J_f(\hat{x}_i, \hat{y}))$$

- Μη-στοχευμένο αντιπαλικό δείγμα

$$\hat{x}_0 = x \text{ και } \hat{x}_{i+1} = clip_{[x-\varepsilon, x+\varepsilon]}(\hat{x}_i + \alpha \cdot \nabla_x J_f(\hat{x}_i, y))$$

όπου x η αρχική εικόνα, \hat{x}_i η αντιπαλική εικόνα στην επανάληψη i , α το μέγεθος του βήματος, ε ένας μικρός αριθμός, $clip_{[x-\varepsilon, x+\varepsilon]}$ η αποκοπή και ο περιορισμός των φωτεινοτήτων των εικονοστοιχέων εντός της γειτονιάς $[x - \varepsilon, x + \varepsilon]$ στα ενδιάμεσα αποτελέσματα της κάθε επανάληψης, J_f η συνάρτηση σφάλματος ενός ταξινομητή f , y η αληθινή κλάση του δείγματος x και \hat{y} η λανθασμένη κλάση στην οποία στοχεύουμε.

Όταν ο ταξινομητής f είναι νευρωνικό δίκτυο τότε ο υπολογισμός της κλίσης της συνάρτησης σφάλματος J_f μπορεί να γίνει με τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

4.7.2.6: Momentum Iterative Fast Gradient Sign Method (MIFGSM)

Η μέθοδος αυτή [31] χρησιμοποιείται για την παραγωγή αντιπαλικών δειγμάτων και αποτελεί παραλλαγή της μεθόδου BIM (επαναληπτική εκδοχή της μεθόδου FGSM). Έχει δειχθεί πως οι επαναληπτικές μέθοδοι σε επιθέσεις «λευκού κουτιού» πετυχαίνουν μεγαλύτερο ποσοστό λανθασμένης κατηγοριοποίησης σε σχέση με τις μη-επαναληπτικές καθώς και το ότι τα αντιπαλικά δείγματα των μη-επαναληπτικών μεθόδων είναι πιο μεταφέρσιμα και έχουν μεγαλύτερη αντοχή σε μετασχηματισμούς σε σχέση με αυτά των επαναληπτικών. Γενικά, οι επαναληπτικές μέθοδοι υπερπροσαρμόζονται στο συγκεκριμένο μοντέλο ταξινόμησης που επιτίθενται με αποτέλεσμα να εγκλωβίζονται σε τοπικά βέλτιστα και έτσι τα αντιπαλικά δείγματα που παράγουν να μην έχουν καλή μεταφερσιμότητα.

Η μέθοδος αυτή χρησιμοποιεί έναν όρο ορμής στις εξισώσεις υπολογισμού του αντιπαλικού δείγματος με τον ίδιο τρόπο που χρησιμοποιείται και στην μέθοδο κατάβασης δυναμικού όταν εκπαιδεύουμε μοντέλα ταξινόμησης. Αυτή η προσέγγιση βοηθάει τόσο στην αύξηση της ταχύτητας σύγκλισης της μεθόδου όσο και στην αποφυγή τοπικών βέλτιστων με αποτέλεσμα τα αντιπαλικά δείγματα που προκύπτουν να είναι ισχυρά και μεταφέρσιμα ταυτόχρονα. Για αυτό το λόγο η μέθοδος αυτή σε επιθέσεις «μαύρου κουτιού» είναι πιο ικανοποιητική από τις υπόλοιπες επαναληπτικές μεθόδους.

Σε κάθε τρέχουσα επανάληψη i η μέθοδος χρησιμοποιεί το πρόσημο του όρου ορμής \mathbf{g}_i για να βρει προς ποια κατεύθυνση θα πρέπει να μεταβληθούν οι φωτεινότητες των εικονοστοιχείων της τρέχουσας εικόνας $\hat{\mathbf{x}}_i$ ώστε να βελτιστοποιηθεί η συνάρτηση σφάλματος J_f ενός ταξινομητή f και στην συνέχεια ανάλογα με την τιμή του βήματος α κάνει ένα μικρό βήμα προς εκείνη την κατεύθυνση αλλοιώντας τα εικονοστοιχεία. Ο όρος ορμής \mathbf{g}_i συσσωρεύει τις κατευθύνσεις της κλίσης της συνάρτησης σφάλματος J_f από τις προηγούμενες επαναλήψεις συνδυαστικά με έναν παράγοντα εξασθένισης μ καθώς και την κατεύθυνση της κλίσης της συνάρτησης σφάλματος J_f της τρέχουσας επανάληψης.

Επίσης, τα μη-στοχευμένα αντιπαλικά δείγματα αυτής της μεθόδου παρουσιάζουν πολύ καλύτερη μεταφερσιμότητα από τα στοχευμένα. Για αυτό το λόγο παρακάτω παραθέτουμε μόνο τις εξισώσεις για τον υπολογισμό μη-στοχευμένων αντιπαλικών δειγμάτων:

$$\mathbf{z}_i = \frac{\nabla_{\mathbf{x}} J_f(\hat{\mathbf{x}}_i, y)}{\|\nabla_{\mathbf{x}} J_f(\hat{\mathbf{x}}_i, y)\|_1}$$

$$\alpha = \frac{\varepsilon}{T}$$

$$\mathbf{g}_0 = \mathbf{0} \quad \text{και} \quad \mathbf{g}_{i+1} = \mu \cdot \mathbf{g}_i + \mathbf{z}_i$$

$$\hat{\mathbf{x}}_0 = \mathbf{x} \quad \text{και} \quad \hat{\mathbf{x}}_{i+1} = \hat{\mathbf{x}}_i + \alpha \cdot \text{sign}(\mathbf{g}_{i+1})$$

όπου \mathbf{x} η αρχική εικόνα, $\hat{\mathbf{x}}_i$ και \mathbf{g}_i η αντιπαλική εικόνα και ο όρος ορμής αντίστοιχα στην επανάληψη i , α το μέγεθος του βήματος, ε το μέγεθος της αντιπαλικής διαταραχής, T ο αριθμός των επαναλήψεων, μ ο παράγοντας εξασθένισης, J_f η συνάρτηση σφάλματος ενός ταξινομητή f και y η αληθινή κλάση του δείγματος \mathbf{x} .

Όταν ο ταξινομητής f είναι νευρωνικό δίκτυο τότε ο υπολογισμός της κλίσης της συνάρτησης σφάλματος J_f μπορεί να γίνει με τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

4.7.2.7: DeepFool Attack

Η μέθοδος αυτή [32] χρησιμοποιείται για την παραγωγή μη-στοχευμένων αντιπαλικών δειγμάτων καθώς είναι αρκετά αποδοτική ακόμη και σε σύγχρονα νευρωνικά μοντέλα ταξινόμησης. Τα αντιπαλικά δείγματα που παράγει είναι πιο κοντά στις αρχικές εικόνες από ότι είναι αυτά των μεθόδων L-BFGS-B, FGSM και JSMA.

Η μέθοδος αυτή κάνει αρκετές υποθέσεις και στηριζόμενη σε αυτές επιλύει το πρόβλημα προσεγγιστικά για μη-γραμμικούς ταξινομητές δύο ή περισσότερων κλάσεων. Αρχικά, γίνεται η υπόθεση ότι οι ταξινομητές δυο κλάσεων είναι αποκλειστικά γραμμικοί και ότι υλοποιούν απλά ένα διαχωριστικό υπερεπίπεδο για των διαχωρισμό των κλάσεων. Στην συνέχεια, επεκτείνουν αυτήν την ιδέα και για ταξινομητές περισσότερων κλάσεων εφόσον μπορούμε να θεωρήσουμε πως αυτοί λειτουργούν ως ένα σύνολο από γραμμικούς ταξινομητές δύο κλάσεων βάση της τεχνικής ένας-εναντίον-όλων.

Με όλες αυτές τις υποθέσεις η μέθοδος προσεγγίζει επαναληπτικά τα αντιπαλικά δείγματα με την μέθοδο της κατάβασης δυναμικού αξιοποιώντας κατάλληλα τις παραγώγους πρώτης

τάξης. Επίσης, κατά την βελτιστοποίηση αναζητείται η ελάχιστη αντιπαλική διαταραχή βάση της l_2 -νόρμας. Ωστόσο, η μέθοδος μπορεί να επεκταθεί και στην γενική μορφή της l_p -νόρμας για $p \in [0, \infty)$. Στην μέθοδο αυτή η κατάβαση δυναμικού δεν εφαρμόζεται για ένα συγκεκριμένο αριθμό επαναλήψεων αλλά έως ότου το αντιπαλικό δείγμα κατηγοριοποιηθεί εσφαλμένα. Ενδεικτικά, παρακάτω παραθέτουμε τις εξισώσεις παραγωγής ενός μη-στοχευμένου αντιπαλικού δείγματος για ταξινομητές δύο κλάσεων:

$$\mathbf{n}_i = -\frac{f(\hat{\mathbf{x}}_i)}{\|\nabla f(\hat{\mathbf{x}}_i)\|_2^2} \nabla f(\hat{\mathbf{x}}_i)$$

$$\hat{\mathbf{x}}_0 = \mathbf{x} \quad \text{και} \quad \hat{\mathbf{x}}_{i+1} = \hat{\mathbf{x}}_i + \mathbf{n}_i$$

$$\mathbf{n} = \sum_i \mathbf{n}_i$$

όπου \mathbf{x} η αρχική εικόνα, $\hat{\mathbf{x}}_i$ και \mathbf{n}_i η αντιπαλική εικόνα και η αντιπαλική διαταραχή αντίστοιχα στην επανάληψη i , \mathbf{n} η τελική αντιπαλική διαταραχή με την οποία η αντιπαλική εικόνα κατηγοριοποιείται λανθασμένα και f ο ταξινομητής δυο κλάσεων.

4.7.2.8: Jacobian-Based Saliency Map Attack (JSMA)

Η μέθοδος αυτή [33] χρησιμοποιείται για την παραγωγή στοχευμένων αντιπαλικών δειγμάτων και βασίζεται στην μετρική απόστασης της l_0 -νόρμας. Συγκεκριμένα, εφαρμόζει τρία βήματα επαναληπτικά. Στο πρώτο βήμα η μέθοδος υπολογίζει τον Ιακωβιανό πίνακα του μοντέλου ταξινόμησης για την εικόνα εισόδου. Στο δεύτερο βήμα αξιοποιώντας τον Ιακωβιανό πίνακα δημιουργεί έναν χάρτη σημαντικότητας. Ο χάρτης αυτός μοντελοποιεί το βαθμό σημαντικότητας κάθε εικονοστοιχείου της εικόνας όσο αφορά την εσφαλμένη κατηγοριοποίηση που θέλουμε να πετύχουμε. Στο τρίτο βήμα χρησιμοποιώντας τον χάρτη σημαντικότητας η μέθοδος επιλέγει με άπληστο τρόπο το πιο σημαντικό εικονοστοιχείο και το αλλοιώνει ώστε η πρόβλεψη του μοντέλου ταξινόμησης να οδηγηθεί στην λανθασμένη κλάση. Αυτά τα τρία βήματα της μεθόδου επαναλαμβάνονται έως ότου είτε ξεπεραστεί ένα άνω κατώφλι για το μέγιστο πλήθος των εικονοστοιχείων που μπορούν να μεταβληθούν είτε γίνει η λανθασμένη κατηγοριοποίηση. Η μέθοδος αυτή αν και πετυχαίνει πολύ καλό ποσοστό λανθασμένης κατηγοριοποίησης έχει αρκετά μεγάλο υπολογιστικό κόστος αφού σε κάθε επανάληψη θα πρέπει να υπολογίζεται ο Ιακωβιανός πίνακας και ο χάρτης σημαντικότητας της εικόνας.

4.7.2.9: Universal Perturbation Attack

Η μέθοδος αυτή [34] βασίζεται αρκετά στην μέθοδο DeepFool και χρησιμοποιείται για την παραγωγή μη-στοχευμένων αντιπαλικών δειγμάτων. Η ιδιαιτερότητα της μεθόδου είναι ότι προσπαθεί να βρει μια καθολική αντιπαλική διαταραχή \mathbf{n} η οποία να μετατρέπει όσο το δυνατόν όλες τις εικόνες ενός συνόλου M σε αντιπαλικές. Η καθολική αντιπαλική διαταραχή \mathbf{n} που ψάχνει να βρει η μέθοδος ικανοποιεί τις παρακάτω συνθήκες:

$$\|\mathbf{n}\|_p \leq \varepsilon \quad \text{και} \quad P_{\mathbf{x} \sim M}(f(\mathbf{x}) \neq f(\mathbf{x} + \mathbf{n})) \geq 1 - \delta$$

όπου \mathbf{x} μια αρχική εικόνα που δειγματοληπτείται από το σύνολο M , \mathbf{n} η καθολική αντιπαλική διαταραχή, ε ένας μικρός αριθμός που περιορίζει το μέγεθος της αντιπαλικής

διαταραχής, $\|\mathbf{n}\|_p$ η l_p -νόρμα για τον υπολογισμό του μεγέθους της αντιπαλικής διαταραχής, δ το ποσοστό εσφαλμένης κατηγοριοποίησης των εικόνων του συνόλου M και f το μοντέλο ταξινόμησης.

Η μέθοδος αυτή σε κάθε επανάληψη χρησιμοποιεί την μέθοδο DeepFool και παράγει ένα δείγμα ελάχιστης διαταραχής για τις εικόνες εισόδου του συνόλου M . Στην συνέχεια, αυτό το δείγμα προστίθεται στην συνολική διαταραχή \mathbf{n} . Η διαδικασία επαναλαμβάνεται έως ότου καλυφθεί το ποσοστό εσφαλμένης κατηγοριοποίησης των εικόνων του συνόλου M . Οι καθολικές αντιπαλικές διαταραχές αυτής της μεθόδου γενικεύονται αρκετά καλά σε διάφορες βαθιές αρχιτεκτονικές (π.χ. VGG, GoogLeNet, ResNet, CaffeNet).

4.7.2.10: Carlini-Wagner Attack (C&W)

Η μέθοδος αυτή [35] χρησιμοποιείται για την παραγωγή αντιπαλικών δειγμάτων και προτείνει τρεις επαναληπτικές επιθέσεις με βάση τις l_0 , l_2 και l_∞ νόρμες. Είναι ικανή για παραγωγή τόσο στοχευμένων όσο και μη-στοχευμένων αντιπαλικών δειγμάτων. Ενδεικτικά, το πρόβλημα βελτιστοποίησης που επιλύει η μέθοδος για την παραγωγή στοχευμένων αντιπαλικών δειγμάτων είναι το ακόλουθο:

$$\underset{\mathbf{n}}{\text{minimize}} \quad \|\mathbf{n}\|_p + c \cdot g(\hat{\mathbf{x}}) \quad \text{έτσι ώστε} \quad \hat{\mathbf{x}} \in [0, 1]^n$$

όπου $\hat{\mathbf{x}}$ η αντιπαλική εικόνα και ορίζεται ως $\hat{\mathbf{x}} = \mathbf{x} + \mathbf{n}$, \mathbf{x} η αρχική εικόνα, \mathbf{n} η ελάχιστη αντιπαλική διαταραχή, $\|\mathbf{n}\|_p$ η l_p -νόρμα για τον υπολογισμό του μεγέθους της αντιπαλικής διαταραχής, c μία υπερπαράμετρος που λειτουργεί σαν συντελεστής βάρους, g μία αντικειμενική συνάρτηση για την οποία ισχύει $g(\hat{\mathbf{x}}) \leq 0$ μόνο όταν ισχύει $f(\hat{\mathbf{x}}) = \hat{y}$ και ορίζεται ως $g(\mathbf{v}) = \max(\max(\{Z(\mathbf{v})_i : i \neq \hat{y}\}) - Z(\mathbf{v})_{\hat{y}}, -\kappa)$, Z η έξοδος του τελευταίου κρυφού επιπέδου ενός ταξινομητή f , κ μια σταθερά που ελέγχει το επίπεδο βεβαιότητας της εσφαλμένης κατηγοριοποίησης του ταξινομητή και \hat{y} η λανθασμένη κλάση στην οποία στοχεύουμε.

Όσο μεγαλύτερη τιμή θέτουμε στην παράμετρο κ τόσο πιο ισχυρά γίνονται τα αντιπαλικά δείγματα. Επίσης, όσο αφορά την υπερπαράμετρο c η τιμή της βρίσκεται όπως και στην μέθοδο L-BFGS-B.

4.7.2.11: One Pixel Attack

Η μέθοδος αυτή [36] χρησιμοποιείται για την παραγωγή στοχευμένων αντιπαλικών δειγμάτων και το ιδιαίτερο χαρακτηριστικό της είναι ότι προσπαθεί να αλλάξει μόνο ένα εικονοστοιχείο της εικόνας εισόδου \mathbf{x} ώστε να την μετατρέψει σε αντιπαλική. Η αντιπαλική εικόνα $\hat{\mathbf{x}}$ ορίζεται ως $\hat{\mathbf{x}} = \mathbf{x} + \mathbf{n}$ όπου \mathbf{n} η αντιπαλική διαταραχή. Έτσι, η αντιπαλική διαταραχή που προστίθεται στην αρχική εικόνα είναι ελάχιστη με αποτέλεσμα οι δύο εικόνες να είναι πανομοιότυπες πλην του εικονοστοιχείου που αλλοιώθηκε. Το πρόβλημα βελτιστοποίησης που επιλύει η μέθοδος είναι το ακόλουθο:

$$\underset{\mathbf{n}}{\text{maximize}} \quad f(\hat{\mathbf{x}}, \hat{y}) \quad \text{έτσι ώστε} \quad \|\mathbf{n}\|_0 \leq \varepsilon_0$$

όπου $\varepsilon_0 = 1$ ώστε να αλλαχθεί μόνο ένα εικονοστοιχείο, f ένα μοντέλο ταξινόμησης και \hat{y} η λανθασμένη κλάση στην οποία στοχεύουμε.

Κεφάλαιο 5: Προτεινόμενη Μέθοδος

Σε αυτό το κεφάλαιο θα περιγραφεί και θα αναλυθεί η προτεινόμενη μέθοδος, θα αναφερθούμε στα πρωτότυπα στοιχεία της καθώς και στα προβλήματα που είναι ικανή να λύσει συγκριτικά με τις προϋπάρχουσες μεθόδους.

5.1: Εισαγωγή

Στα προηγούμενα κεφάλαια αναλύσαμε τις διάφορες προϋπάρχουσες μεθόδους που χρησιμοποιούνται για αποταυτοποίηση προσώπου καθώς και για παραγωγή αντιπαλικών δειγμάτων. Σε αυτό το κεφάλαιο προτείνουμε την μέθοδο «Penalized Fast Gradient Value Method» (P-FGVM) η οποία υλοποιεί την αποταυτοποίηση προσώπου χρησιμοποιώντας αντιπαλικά δείγματα. Αυτή η προσέγγιση είναι νέα και καινοτόμα καθώς επιλύει σημαντικά προβλήματα που οι προηγούμενες μέθοδοι αποταυτοποίησης προσώπου αδυνατούν να χειριστούν.

Για παράδειγμα, είδαμε πως σε εφαρμογές όπου θέλουμε οι αποταυτοποιημένες εικόνες να είναι όσο το δυνατόν ρεαλιστικές και όμοιες με τις αρχικές, οι προϋπάρχουσες μέθοδοι δεν μπορούν να θεωρηθούν αποδεκτές διότι αλλοιώνουν έντονα την εμφάνιση του προσώπου. Ακόμη, αδυνατούν να διατηρήσουν ικανοποιητικά στο αποταυτοποιημένο πρόσωπο τα διάφορα μη-ταυτοτικά χαρακτηριστικά του αρχικού, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα, με αποτέλεσμα να είναι εξίσου μη αποδεκτές σε εφαρμογές που τα χρειάζονται.

Η μέθοδος P-FGVM λειτουργεί στον χώρο της εικόνας και μπορεί να αποταυτοποιήσει μία εικόνα προσώπου αλλοιώνοντας την ελάχιστα με τέτοιο τρόπο ώστε η αποταυτοποιημένη εικόνα να είναι ανεπαίσθητα διαφορετική από την αρχική αλλά και να μην μπορεί να αναγνωριστεί σωστά από σύγχρονα αυτόματα συστήματα αναγνώρισης προσώπου. Στην πραγματικότητα αποτελεί μέθοδος αντιπαλικής επίθεσης και κατασκευάζει τις αποταυτοποιημένες εικόνες ως στοχευμένα αντιπαλικά δείγματα.

5.2: Περιγραφή Μεθόδου

Η μέθοδος «Penalized Fast Gradient Value Method» (P-FGVM) βασίζεται στις αρχές λειτουργίας της μεθόδου «Iterative Fast Gradient Value Method» (IFGVM) και χρησιμοποιείται αποκλειστικά για την παραγωγή στοχευμένων αντιπαλικών δειγμάτων. Όπως όλες οι μέθοδοι αντιπαλικής επίθεσης έτσι και αυτή είναι ικανή να πάρει στην είσοδο μία εικόνα x και να παράγει στην έξοδο μια αντιπαλική εικόνα \hat{x} η οποία είναι παρόμοια με την x αλλά κατηγοριοποιείται λανθασμένα στην κλάση \hat{y} από έναν ταξινομητή f . Η μέθοδος ελαχιστοποιεί την παρακάτω πολυκριτηριακή αντικειμενική συνάρτηση:

$$J_f(\hat{x}, \hat{y}) + \lambda \cdot \frac{1}{2} \cdot \|\hat{x} - x\|_2^2 \quad \text{έτσι ώστε } \hat{x} \in [0, 1]^n$$

όπου J_f η συνάρτηση σφάλματος του ταξινομητή f και $\lambda \cdot \frac{1}{2} \cdot \|\hat{x} - x\|_2^2$ ένας όρος ποινής που εξαρτάται από το μέγεθος της αντιπαλικής διαταραχής βάση της l_2 -νόρμας καθώς και από την υπερπαράμετρο λ που λειτουργεί ως συντελεστής βάρους.

Επίσης, η ελαχιστοποίηση της παραπάνω αντικειμενικής συνάρτησης γίνεται με την μέθοδο της κατάβασης δυναμικού. Για την παραγωγή του στοχευμένου αντιπαλικού δείγματος χρησιμοποιούνται οι παρακάτω εξισώσεις:

$$\hat{x}_0 = x \quad \text{και} \quad \hat{x}_{i+1} = clip_{[0,1]}(\hat{x}_i - \alpha \cdot (\nabla_x J_f(\hat{x}_i, \hat{y}) + \lambda \cdot (\hat{x}_i - x)))$$

όπου \hat{x}_i η αντιπαλική εικόνα στην επανάληψη i , α το μέγεθος του βήματος και $clip_{[0,1]}$ ο περιορισμός των φωτεινοτήτων των εικονοστοιχείων εντός των ορίων $[0,1]$ στα ενδιάμεσα αποτελέσματα της κάθε επανάληψης ώστε η εικόνα να είναι έγκυρη.

Όταν ο ταξινομητής f είναι νευρωνικό δίκτυο τότε ο υπολογισμός της κλίσης της συνάρτησης σφάλματος J_f μπορεί να γίνει με τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

5.3: Πρωτότυπα Στοιχεία

Η μέθοδος P-FGVM έχει τρία καινοτόμα και πρωτότυπα στοιχεία:

1. Είναι η πρώτη μέθοδος που χρησιμοποιεί αντιπαλικά δείγματα για να επιλύσει το πρόβλημα της αποταυτοποίησης προσώπου. Οι προϋπάρχουσες μέθοδοι αποταυτοποίησης προσώπου επιλύουν το πρόβλημα με διαφορετικό τρόπο. Έτσι, ένα νέο εξελικτικό μονοπάτι γνωστοποιείται στο ευρύ κοινό για το πώς μπορεί να υλοποιηθεί η αποταυτοποίηση προσώπου με την βοήθεια των αντιπαλικών δειγμάτων.
2. Σε σχέση με τις προϋπάρχουσες μεθόδους αποταυτοποίησης προσώπου προσφέρει καλύτερα αποτελέσματα αφού το ποσοστό λανθασμένης κατηγοριοποίησης που πετυχαίνει είναι μεγαλύτερο και οι αποταυτοποιημένες εικόνες που παράγει είναι ανεπαίσθητα διαφορετικές από τις αρχικές. Επίσης, στις αποταυτοποιημένες εικόνες διατηρούνται στο μέγιστο τα μη-ταυτοτικά χαρακτηριστικά των αρχικών προσώπων, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα.
3. Οι αντιπαλικές εικόνες που παράγονται από την μέθοδο είναι ρεαλιστικές και όμοιες με τις αρχικές λόγο του όρου ποινής $\lambda \cdot (\hat{x}_i - x)$ στην εξίσωση της μεθόδου. Όσο περισσότερο ελαχιστοποιείται αυτός ο όρος τόσο η αντιπαλική εικόνα μοιάζει με την αρχική. Με λίγα λόγια ελαχιστοποιείται η αντιπαλική διαταραχή. Η ιδέα του όρου ποινής ως προσέγγιση είναι καινούργια και δίδει καλύτερα αποτελέσματα από τον μηχανισμό $clip_{[x-\varepsilon, x+\varepsilon]}$ που χρησιμοποιείται για τον περιορισμό των φωτεινοτήτων των εικονοστοιχείων στην αντιπαλική εικόνα. Επίσης, με την προσέγγιση του όρου ποινής η αντιπαλική εικόνα μπορεί να αρχικοποιηθεί με τυχαίο θόρυβο και στην συνέχεια να εξελιχθεί προσεγγιστικά στην αρχική εικόνα.

Κεφάλαιο 6: Πειραματική Διαδικασία

Σε αυτό το κεφάλαιο θα περιγράψουμε την συγκριτική ανάλυση που διεξήγαμε, τα σύνολα δεδομένων που χρησιμοποιήσαμε και τον τρόπο με τον οποίο τα προεπεξεργαστήκαμε, τα μοντέλα αναγνώρισης προσώπου που επιτεθήκαμε, τις υπερπαραμέτρους που επιλέξαμε για τις μεθόδους της συγκριτικής ανάλυσης καθώς και το βοηθητικό υλικό.

6.1: Συγκριτική Ανάλυση

Στα πειράματα μας προσπαθήσαμε να αποταυτοποιήσουμε εικόνες προσώπου από το σύνολο δεδομένων CelebA [38] χρησιμοποιώντας τις μεθόδους IFGVM, ITCM καθώς και την προτεινόμενη μέθοδο P-FGVM. Συγκεκριμένα, με αυτές τις μεθόδους στοχεύσαμε στο να κατασκευάσουμε τις αποταυτοποιημένες εικόνες προσώπου ως στοχευμένα αντιπαλικά δείγματα. Γενικά, αυτές οι μέθοδοι παράγουν αντιπαλικά δείγματα κάνοντας επίθεση σε κάποιο μοντέλο ταξινόμησης. Για αυτό δημιουργήσαμε δύο συνελικτικά νευρωνικά δίκτυα και τα εκπαιδεύσαμε με ένα μικρό υποσύνολο του συνόλου δεδομένων CelebA ώστε να κάνουν αναγνώριση προσώπου.

Αρχικά, χρησιμοποιώντας την μέθοδο P-FGVM στοχεύσαμε στο να κατασκευάσουμε ρεαλιστικές αποταυτοποιημένες εικόνες προσώπου με μεγάλο ποσοστό εσφαλμένης κατηγοριοποίησης. Συγκεκριμένα, έχοντας ως εικόνα στόχου την εικόνα προσώπου προς αποταυτοποίηση στοχεύσαμε στο να παράγουμε την αποταυτοποιημένη χρησιμοποιώντας ως είσοδο είτε την εικόνα προσώπου προς αποταυτοποίηση είτε τυχαίο Γκαουσιανό Θόρυβο. Στην συνέχεια, χρησιμοποιώντας τις μεθόδους IFGVM και ITCM στοχεύσαμε εξίσου στο να κατασκευάσουμε ρεαλιστικές αποταυτοποιημένες εικόνες προσώπου με μεγάλο ποσοστό εσφαλμένης κατηγοριοποίησης. Ωστόσο, στις μεθόδους αυτές χρησιμοποιήσαμε ως είσοδο μόνο την εικόνα προσώπου προς αποταυτοποίηση.

Στα πειράματα μας χρησιμοποιήσαμε ως μετρικές αξιολόγησης τον δείκτη ομοιότητας CW-SSIM [37] μεταξύ των αποταυτοποιημένων και των αρχικών εικόνων προσώπου, την l_2 -νόρμα για τον υπολογισμό του μεγέθους της αντιπαλικής διαταραχής που προστίθεται στην αρχική εικόνα καθώς και το ποσοστό εσφαλμένης κατηγοριοποίησης. Με αυτές τις μετρικές έγινε και η σύγκριση των μεθόδων.

6.2: Σύνολα Δεδομένων

Οι περισσότεροι μέθοδοι αποταυτοποίησης προσώπου καθώς και αρκετά μοντέλα αναγνώρισης προσώπου λειτουργούν με εικόνες που απεικονίζονται πρόσωπα σε εμπρόσθια όψη. Για αυτό το λόγο στα πειράματα μας χρησιμοποιήσαμε εικόνες που προέρχονται από τα σύνολα δεδομένων «CelebFaces Attributes Dataset (CelebA)» και «VGG Face Dataset».

6.2.1: CelebFaces Attributes Dataset (CelebA)

Αυτό το σύνολο δεδομένων περιέχει εικόνες στις οποίες εμφανίζονται πρόσωπα διαφόρων διασημοτήτων καθώς και χρήσιμα μεταδεδομένα που σχετίζονται με αυτά. Οι εικόνες αυτές έχουν ανάλυση 178x218, είναι έγχρωμες και τα διάφορα πρόσωπα που απεικονίζονται σε αυτές είναι ευθυγραμμισμένα. Επίσης, έχουν μεγάλη ποικιλία όσο αφορά τα ταυτοικά ή μη-ταυτοικά χαρακτηριστικά προσώπου και τον θόρυβο παρασκηνίου. Συγκεκριμένα, το

σύνολο δεδομένων περιέχει 202.599 εικόνες που αντιστοιχούν σε 10.177 ταυτότητες και διατηρεί για κάθε μια εικόνα 5 ορόσημα προσώπου και 40 δυαδικά επισημειωμένα χαρακτηριστικά.

6.2.2: VGG Face Dataset

Αυτό το σύνολο δεδομένων περιέχει διαδικτυακούς συνδέσμους σε εικόνες στις οποίες εμφανίζονται πρόσωπα διαφόρων διασημοτήτων καθώς και χρήσιμα μεταδεδομένα που σχετίζονται με αυτά. Οι εικόνες αυτές δεν έχουν κάποια κοινή ανάλυση, μπορεί να είναι έγχρωμες ή διαβάθμισης του γκρι και τα διάφορα πρόσωπα που απεικονίζονται σε αυτές δεν είναι ευθυγραμμισμένα. Επίσης, έχουν μεγάλη ποικιλία όσο αφορά τα ταυτοτικά ή μη-ταυτοτικά χαρακτηριστικά προσώπου και τον θόρυβο παρασκηνίου. Συγκεκριμένα, το σύνολο δεδομένων περιέχει 2.6 εκατομμύρια εικόνες που αντιστοιχούν σε 2.622 ταυτότητες και διατηρεί για κάθε μια εικόνα πληροφορίες που σχετίζονται με την πόζα του προσώπου και το ορθογώνιο οριοθέτησης που ανιχνεύτηκε από τον ανιχνευτή προσώπου.

6.3: Προεπεξεργασία Δεδομένων

Η αρχική μας προσέγγιση ήταν να εκπαιδεύσουμε τα μοντέλα αναγνώρισης προσώπου με όλες τις εικόνες του συνόλου δεδομένων CelebA. Ωστόσο, η διαδικασία αυτή είχε πολύ μεγάλη χρονική πολυπλοκότητα και για αυτό έπρεπε να χρησιμοποιήσουμε ένα μικρό υποσύνολο του συνόλου δεδομένων. Έτσι, δημιουργήσαμε ένα πρόγραμμα με την γλώσσα προγραμματισμού C# το οποίο είναι ικανό να εξάγει ένα τυχαίο υποσύνολο από το σύνολο δεδομένων. Στο πρόγραμμα αυτό μπορούμε να ορίσουμε το πλήθος των κλάσεων αλλά και των πλήθος των εικόνων ανά κλάση που θέλουμε να εξάγουμε ώστε να έχουμε κλάσεις ίσου μεγέθους. Το υποσύνολο που εξάγαμε και χρησιμοποιήσαμε για την εκπαίδευση των μοντέλων μας περιέχει συνολικά 900 εικόνες (30 κλάσεις με 30 εικόνες η κάθε μια).

6.4: Μοντέλα Αναγνώρισης Προσώπου

Για τις ανάγκες των πειραμάτων μας δημιουργήσαμε δύο συνελικτικά νευρωνικά δίκτυα και τα εκπαιδεύσαμε με ένα μικρό υποσύνολο του συνόλου δεδομένων CelebA ώστε να κάνουν αναγνώριση προσώπου. Παρακάτω, θα επεξηγήσουμε αναλυτικά πώς τα εκπαιδεύσαμε καθώς και πως καταλήξαμε στην αρχιτεκτονική τους.

6.4.1: Μοντέλο A

Το μοντέλο αυτό ανακαλύφθηκε κάνοντας εξαντλητική αναζήτηση στις τιμές ενός συνόλου παραμέτρων. Για κάθε μοναδικό συνδυασμό τιμών κατασκευάσαμε ένα καινούργιο μοντέλο και αφού το εκπαιδεύσαμε μετρήσαμε την ακρίβεια του στα σύνολα δεδομένων εκπαίδευσης, επικύρωσης και ελέγχου. Ως μέτρο σύγκρισης των διαφόρων μοντέλων χρησιμοποιήσαμε τις ακρίβειες στα σύνολα δεδομένων επικύρωσης και ελέγχου.

Η αναζήτηση έγινε σε δύο φάσεις. Στην πρώτη φάση κάναμε αναζήτηση σε μία ομάδα παραμέτρων που σχετίζονται με την αρχιτεκτονική και την εκπαίδευση του μοντέλου. Στην δεύτερη φάση χρησιμοποιήσαμε ως βάση τον καλύτερο συνδυασμό τιμών της πρώτης φάσης και κάναμε επιπλέον αναζήτηση σε μια ομάδα παραμέτρων που σχετίζονται με μεθόδους συστηματοποίησης-γενίκευσης. Τέλος, με τον καλύτερο συνδυασμό τιμών όλων

των παραμέτρων εκπαιδεύσαμε 100 διαφορετικά μοντέλα και κρατήσαμε το καλύτερο από αυτά ως το τελικό.

Η εκπαίδευση του νευρωνικού δικτύου έγινε στην GPU κάρτα γραφικών NVIDIA GeForce GTX 1080 χρησιμοποιώντας τον βελτιστοποιητή «Adam» (σύγχρονη μέθοδος κατάβασης δυναμικού) καθώς και τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

Ακολουθούν οι τιμές των παραμέτρων που δοκιμάσαμε στην πρώτη φάση:

Πίνακας 1: Οι τιμές των παραμέτρων της πρώτης φάσης αναζήτησης του Μοντέλου A

Παράμετρος	Τιμές
Ρυθμός Μάθησης (Learning Rate)	1e-1, 1e-2, 1e-3, 1e-4, 1e-5
Μέγεθος Δεσμίδας (Batch Size)	8, 16, 32, 64, 128
Μέγεθος Συνελικτικού Πυρήνα (Convolution Kernel Size)	3x3, 5x5
Επιπλέον Συνελικτικά Μπλοκ (Extra Convolution Blocks)	0, 1, 2
Φίλτρα Πρώτου Συνελικτικού Μπλοκ (First Convolution Block Filters)	8, 16, 32
Νευρώνες Προτελευταίου Επιπέδου (Penultimate Level Neurons)	16, 32, 64, 128, 256, 512

Ακολουθούν οι τιμές των παραμέτρων που δοκιμάσαμε στην δεύτερη φάση:

Πίνακας 2: Οι τιμές των παραμέτρων της δεύτερης φάσης αναζήτησης του Μοντέλου A

Παράμετρος	Τιμές
Πιθανότητα Περιορισμού Ενεργοποίησης (Dropout Rate)	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Παράγοντας Κανονικοποίησης l_2 -νόρμας (L2 Regularization Factor)	1e-1, 1e-2, 1e-3, 1e-4, 1e-5
Χρήση Κανονικοποίησης Παρτίδας (Batch Normalization Usage)	Ναι, Όχι

Ακολουθεί ο καλύτερος συνδυασμός τιμών παραμέτρων για το Μοντέλο A:

Πίνακας 3: Ο καλύτερος συνδυασμός τιμών παραμέτρων για το Μοντέλου A

Παράμετρος	Τιμή
Ρυθμός Μάθησης (Learning Rate)	1e-4
Μέγεθος Δεσμίδας (Batch Size)	16
Μέγεθος Συνελικτικού Πυρήνα (Convolution Kernel Size)	5x5
Επιπλέον Συνελικτικά Μπλοκ (Extra Convolution Blocks)	1
Φίλτρα Πρώτου Συνελικτικού Μπλοκ (First Convolution Block Filters)	32
Νευρώνες Προτελευταίου Επιπέδου (Penultimate Level Neurons)	512

Πιθανότητα Περιορισμού Ενεργοποίησης (Dropout Rate)	0.9
Παράγοντας Κανονικοποίησης l_2 -νόρμας (L2 Regularization Factor)	1e-3
Χρήση Κανονικοποίησης Παρτίδας (Batch Normalization Usage)	Ναι

Ακολουθεί η αρχιτεκτονική του Μοντέλου Α:

Πίνακας 4: Η αρχιτεκτονική του Μοντέλου Α

Conv(32, Kernel(5, 5), Padding(Same), L2Regularizer(1e-3)) BatchNormalization+Relu MaxPooling(PoolSize(2, 2), Strides(2, 2)) Conv(64, Kernel(5, 5), Padding(Same), L2Regularizer(1e-3)) BatchNormalization+Relu MaxPooling(PoolSize(2, 2), Strides(2, 2)) FC(512, L2Regularizer(1e-3)) BatchNormalization+Relu Dropout(0.9) FC(30)+Softmax

Ακολουθεί η συνοπτική αναπαράσταση του Μοντέλου Α σε Keras:

Πίνακας 5: Η συνοπτική αναπαράσταση του Μοντέλου Α σε Keras

Layer	Output Shape	Parameters
input_1	(None, 218, 178, 3)	0
conv2d_1	(None, 218, 178, 32)	2432
batch_normalization_1	(None, 218, 178, 32)	128
activation_1	(None, 218, 178, 32)	0
max_pooling2d_1	(None, 109, 89, 32)	0
conv2d_2	(None, 109, 89, 64)	51264
batch_normalization_2	(None, 109, 89, 64)	256
activation_2	(None, 109, 89, 64)	0
max_pooling2d_2	(None, 54, 44, 64)	0
flatten_1	(None, 152064)	0
dense_1	(None, 512)	77857280
batch_normalization_3	(None, 512)	2048
activation_3	(None, 512)	0
dropout_1	(None, 512)	0
dense_2	(None, 30)	15390
Total params: 77,928,798		
Trainable params: 77,927,582		
Non-trainable params: 1,216		

Ακολουθούν οι πληροφορίες εκπαίδευσης του Μοντέλου Α:

Πίνακας 6: Οι πληροφορίες εκπαίδευσης του Μοντέλου Α

Σύνολο Δεδομένων (Dataset)	CelebA
Συνολικές Κλάσεις (Total Classes)	30

Συνολικές Εικόνες (Total Images)	900
Ανάλυση Εικόνων (Images Resolution)	178x218
Αναλογία Εκπαίδευσης (Training Ratio)	70%
Αναλογία Ελέγχου (Testing Ratio)	15%
Αναλογία Επικύρωσης (Validation Ratio)	15%
Στρωματοποιημένη Δειγματοληψία (Stratified Sampling)	Ναι
Κανονικοποίηση Εικόνων (Images Normalization)	MinMax
Ρυθμός Μάθησης (Learning Rate)	1e-4
Αλγόριθμος Εκπαίδευσης (Training Algorithm)	Οπισθοδιάδοση Σφάλματος
Μέθοδος Βελτιστοποίησης (Optimization Method)	Adam
Συνάρτηση Σφάλματος (Loss Function)	Διεντροπία
Μέγεθος Δεσμίδας (Batch Size)	16
Εποχές Εκπαίδευσης (Training Epochs)	147
Ακρίβεια Ελέγχου (Testing Accuracy)	80.7%
Ακρίβεια Εκπαίδευσης (Training Accuracy)	100%
Ακρίβεια Επικύρωσης (Validation Accuracy)	80%

6.4.2: Μοντέλο Β

Το μοντέλο αυτό έχει προκύψει με μεταφορά μάθησης και αξιοποιεί αυτούσια τα προεκπαιδευμένα VGG-16 συνελικτικά επίπεδα του μοντέλου «VGG-Face CNN descriptor» [39]. Αυτά τα επίπεδα έχουν εκπαιδευτεί στο σύνολο δεδομένων «VGG Face Dataset» και λειτουργούν ως γενικοί εξαγωγείς χαρακτηριστικών προσώπου.

Συγκεκριμένα, συνθέσαμε ένα νέο συνελικτικό νευρωνικό δίκτυο το οποίο αποτελείται από τα προεκπαιδευμένα συνελικτικά επίπεδα του μοντέλου «VGG-Face CNN descriptor» καθώς και από μερικά πλήρως διασυνδεδεμένα επίπεδα εμπρόσθιας τροφοδότησης. Κατά την εκπαίδευση κρατήσαμε σταθερά τα προεκπαιδευμένα συνελικτικά επίπεδα και εκπαιδεύσαμε τα υπόλοιπα επίπεδα του νευρωνικού δικτύου.

Η εκπαίδευση του νευρωνικού δικτύου έγινε στην GPU κάρτα γραφικών NVIDIA GeForce GTX 1080 χρησιμοποιώντας τον βελτιστοποιητή «Adam» (σύγχρονη μέθοδος κατάβασης δυναμικού) καθώς και τον αλγόριθμο της οπισθοδιάδοσης σφάλματος.

Ακολουθεί η αρχιτεκτονική του Μοντέλου Β:

Πίνακας 7: Η αρχιτεκτονική του Μοντέλου B

VGG-Face CNN descriptor (VGG-16)
FC(256, L2Regularizer(1e-3))
BatchNormalization+Relu
FC(30)+Softmax

Ακολουθεί η συνοπτική αναπαράσταση του Μοντέλου B σε Keras:

Πίνακας 8: Η συνοπτική αναπαράσταση του Μοντέλου B σε Keras

Layer	Output Shape	Parameters
input_1	(None, 218, 178, 3)	0
conv1_1	(None, 218, 178, 64)	1792
conv1_2	(None, 218, 178, 64)	36928
pool1	(None, 109, 89, 64)	0
conv2_1	(None, 109, 89, 128)	73856
conv2_2	(None, 109, 89, 128)	147584
pool2	(None, 54, 44, 128)	0
conv3_1	(None, 54, 44, 256)	295168
conv3_2	(None, 54, 44, 256)	590080
conv3_3	(None, 54, 44, 256)	590080
pool3	(None, 27, 22, 256)	0
conv4_1	(None, 27, 22, 512)	1180160
conv4_2	(None, 27, 22, 512)	2359808
conv4_3	(None, 27, 22, 512)	2359808
pool4	(None, 13, 11, 512)	0
conv5_1	(None, 13, 11, 512)	2359808
conv5_2	(None, 13, 11, 512)	2359808
conv5_3	(None, 13, 11, 512)	2359808
pool5	(None, 6, 5, 512)	0
flatten_1	(None, 15360)	0
dense_1	(None, 256)	3932416
batch_normalization_1	(None, 256)	1024
activation_1	(None, 256)	0
dense_2	(None, 30)	7710
Total params: 18,655,838		
Trainable params: 3,940,638		
Non-trainable params: 14,715,200		

Ακολουθούν οι πληροφορίες εκπαίδευσης του Μοντέλου B:

Πίνακας 9: Οι πληροφορίες εκπαίδευσης του Μοντέλου B

Σύνολο Δεδομένων (Dataset)	CelebA
Συνολικές Κλάσεις (Total Classes)	30
Συνολικές Εικόνες (Total Images)	900
Ανάλυση Εικόνων (Images Resolution)	178x218
Αναλογία Εκπαίδευσης	70%

(Training Ratio)	
Αναλογία Ελέγχου (Testing Ratio)	15%
Αναλογία Επικύρωσης (Validation Ratio)	15%
Στρωματοποιημένη Δειγματοληψία (Stratified Sampling)	Ναι
Κανονικοποίηση Εικόνων (Images Normalization)	MinMax
Ρυθμός Μάθησης (Learning Rate)	1e-4
Αλγόριθμος Εκπαίδευσης (Training Algorithm)	Οπισθοδιάδοση Σφάλματος
Μέθοδος Βελτιστοποίησης (Optimization Method)	Adam
Συνάρτηση Σφάλματος (Loss Function)	Διεντροπία
Μέγεθος Δεσμίδας (Batch Size)	16
Εποχές Εκπαίδευσης (Training Epochs)	144
Ακρίβεια Ελέγχου (Testing Accuracy)	95.4%
Ακρίβεια Εκπαίδευσης (Training Accuracy)	100%
Ακρίβεια Επικύρωσης (Validation Accuracy)	97.8%

6.5: Τιμές Παραμέτρων

Στα πειράματα μας συγκρίναμε την προτεινόμενη μέθοδο P-FGVM με τις μεθόδους IFGVM και ITCM. Κάθε μία από αυτές τις μεθόδους χρησιμοποιεί ένα πλήθος παραμέτρων. Εφαρμόζοντας την διαδικασία δοκιμής και σφάλματος μπορέσαμε να βρούμε τιμές για αυτές τις παραμέτρους που οδηγούν σε καλά αποτελέσματα αποταυτοποίησης προσώπου.

Συγκεκριμένα, η παράμετρος N είναι το πλήθος των επαναλήψεων, η παράμετρος α είναι το μέγεθος του βήματος, η παράμετρος ε είναι η τιμή που ελέγχει το μέγεθος της αντιπαλικής διαταραχής στις μεθόδους IFGVM, ITCM και η παράμετρος λ είναι ο συντελεστής βάρους του όρου ποινής στην μέθοδο P-FGVM.

Πίνακας 10: Οι τιμές παραμέτρων για τις μεθόδους της συγκριτικής ανάλυσης

Μέθοδος	Μοντέλο A				Μοντέλο B			
	α	N	ε	λ	α	N	ε	λ
P-FGVM	1.0	50	n/a	0.22	0.55	58	n/a	0.28
IFGVM	1.0	50	0.022	n/a	0.1	40	0.022	n/a
ITCM	$\varepsilon \div N$	20	0.026	n/a	$\varepsilon \div N$	20	0.026	n/a

6.6: Βοηθητικό Υλικό

Τα πειράματα που πραγματοποιήσαμε υλοποιήθηκαν με την γλώσσα προγραμματισμού Python και εκτελέστηκαν μέσα στο προγραμματιστικό περιβάλλον ανάπτυξης JetBrains PyCharm.

Η εκπαίδευση των νευρωνικών δικτύων έγινε με CUDA στην GPU κάρτα γραφικών NVIDIA GeForce GTX 1080. Με την βοήθεια του προγράμματος DatasetsTransformation που υλοποιήσαμε στο Visual Studio με την γλώσσα προγραμματισμού C# μπορέσαμε να εξάγουμε το υποσύνολο του συνόλου δεδομένων CelebA.

Επίσης, χρησιμοποιήσαμε τις παρακάτω βιβλιοθήκες Python στα πειράματα μας:

Πίνακας 11: Οι βιβλιοθήκες Python που χρησιμοποιήσαμε στα πειράματα μας

Βιβλιοθήκη	Έκδοση
scikit-learn	0.20.1
scikit-image	0.14.0
numpy	1.15.4
scipy	1.1.0
matplotlib	3.0.2
keras-gpu	2.2.4
tensorflow-gpu	1.12.0
cudatoolkit	9.0
cudnn	7.1.2
keras_vggface	0.5
pyssim	0.4

Κεφάλαιο 7: Αποτελέσματα Κ Συμπεράσματα

Σε αυτό το κεφάλαιο θα παρουσιάσουμε τα αποτελέσματα της συγκριτικής ανάλυσης που εφαρμόσαμε καθώς και διάφορα παραδείγματα αποταυτοποιημένων εικόνων. Επιπλέον, θα σχολιάσουμε τα πειραματικά αποτελέσματα των μεθόδων.

7.1: Αποτελέσματα Συγκριτικής Ανάλυσης

Οι μετρικές αξιολόγησης που υπολογίσαμε στα πειράματα μας για την συγκριτική ανάλυση των μεθόδων P-FGVM, IFGVM και ITCM είναι ο δείκτης ομοιότητας CW-SSIM μεταξύ των αποταυτοποιημένων και των αρχικών εικόνων προσώπου, η l_2 -νόρμα της αντιπαλικής διαταραχής που προστίθεται στην αρχική εικόνα καθώς και το ποσοστό εσφαλμένης κατηγοριοποίησης των αποταυτοποιημένων εικόνων προσώπου.

Στον παρακάτω πίνακα βλέπουμε τα αποτελέσματα των μετρικών αξιολόγησης (κάθε τιμή του πίνακα αποτελεί μέσος όρος πολλών δοκιμών) των μεθόδων της συγκριτικής ανάλυσης:

Πίνακας 12: Τα αποτελέσματα της αξιολόγησης των μεθόδων της συγκριτικής ανάλυσης

Μοντέλο Α			Μοντέλο Β		
l_2 -νόρμα	CW-SSIM	Ποσοστό Εσφαλμένης Κατηγοριοποίησης	l_2 -νόρμα	CW-SSIM	Ποσοστό Εσφαλμένης Κατηγοριοποίησης
P-FGVM					
3.39	0.438	99.6%	2.11	0.456	95.9%
IFGVM					
5.32	0.421	99.4%	2.71	0.441	93.2%
ITCM					
5.68	0.424	98.9%	5.75	0.423	94.4%

Στον παρακάτω πίνακα βλέπουμε τα ποσοστά βελτίωσης της μεθόδου P-FGVM:

Πίνακας 13: Τα ποσοστά βελτίωσης της μεθόδου P-FGVM στις μετρικές αξιολόγησης

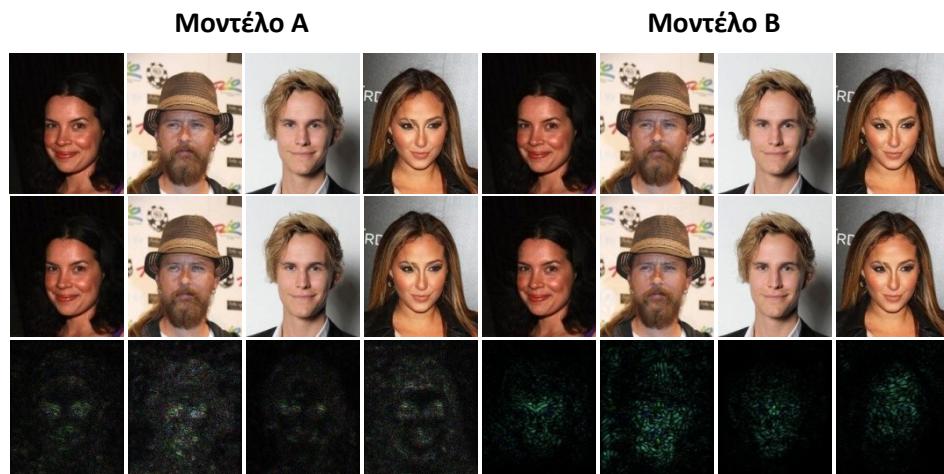
Μοντέλο Α			Μοντέλο Β		
l_2 -νόρμα	CW-SSIM	Ποσοστό Εσφαλμένης Κατηγοριοποίησης	l_2 -νόρμα	CW-SSIM	Ποσοστό Εσφαλμένης Κατηγοριοποίησης
Σε σχέση με την μέθοδο IFGVM					
36.3%	4.0%	0.2%	22.1%	3.4%	2.9%
Σε σχέση με την μέθοδο ITCM					
40.3%	3.3%	0.7%	63.3%	7.8%	1.6%

7.2: Παραδείγματα Αποταυτοποιημένων Εικόνων

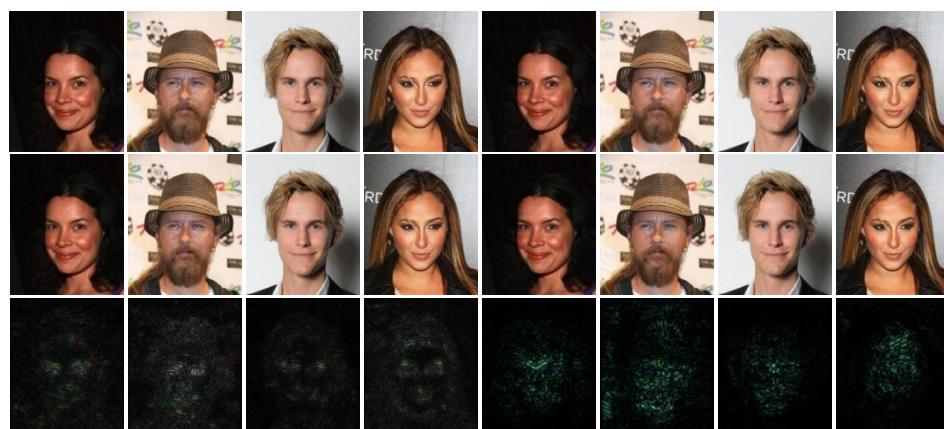
Στους παρακάτω τρεις πίνακες βρίσκονται μερικά παραδείγματα αποταυτοποιημένων εικόνων προσώπου του συνόλου δεδομένων CelebA που παρήγαγαν οι μέθοδοι αντιπαλικής επίθεσης P-FGVM, IFGVM και ITCM κάνοντας επίθεση «λευκού κουτιού» στα μοντέλα αναγνώρισης προσώπου A και B.

Πιο συγκεκριμένα, σε κάθε στήλη των πινάκων βρίσκεται μια ενδεικτική αρχική εικόνα προσώπου, η αντίστοιχη αποταυτοποιημένη καθώς και η απόλυτη τιμή της αντιπαλικής διαταραχής (μεγεθυμένη κατά 10 φορές) που μετατρέπει την αρχική εικόνα προσώπου σε αντιπαλική.

Πίνακας 14: Παραδείγματα αποταυτοποιημένων εικόνων με την μέθοδο P-FGVM

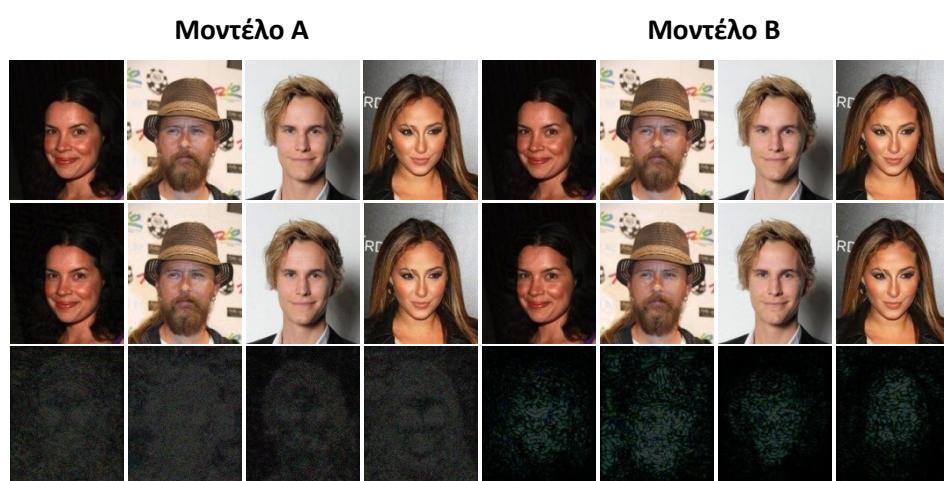


Έχοντας ως είσοδο τυχαίο Γκαουσιανό θόρυβο

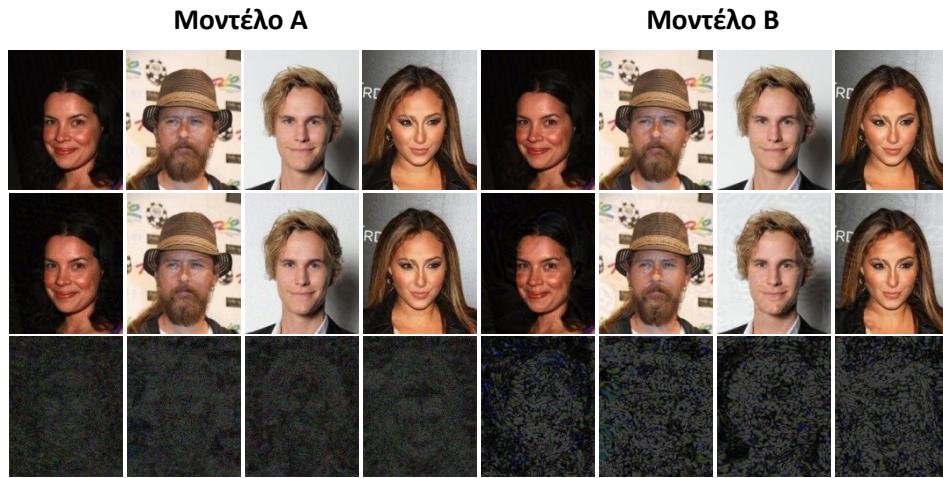


Έχοντας ως είσοδο την αρχική εικόνα προσώπου

Πίνακας 15: Παραδείγματα αποταυτοποιημένων εικόνων με την μέθοδο IFGVM

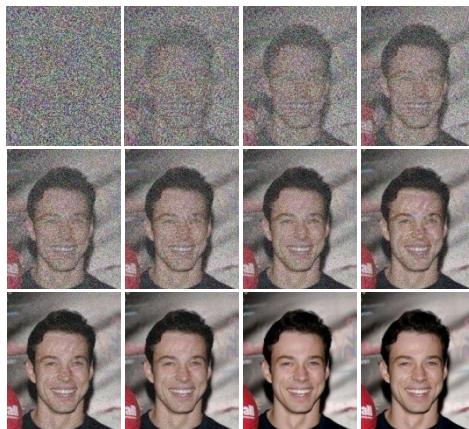


Πίνακας 16: Παραδείγματα αποταυτοποιημένων εικόνων με την μέθοδο ITCM



Στον παρακάτω πίνακα βλέπουμε ένα ενδεικτικό παράδειγμα του πως η μέθοδος P-FGVM μπορεί να παράγει μια αποταυτοποιημένη εικόνα προσώπου παίρνοντας ως είσοδο μόνο τυχαίο Γκαουσιανό θόρυβο.

Πίνακας 17: Χρήση της μεθόδου P-FGVM με τυχαίο Γκαουσιανό θόρυβο ως είσοδο



7.3: Σχόλια Πειραματικών Αποτελεσμάτων

Από τα αποτελέσματα της συγκριτικής ανάλυσης φαίνεται ξεκάθαρα πως η προτεινόμενη μέθοδος P-FGVM είναι καλύτερη από τις μεθόδους IFGVM, ITCM και στα δύο μοντέλα αναγνώρισης προσώπου (Μοντέλο Α και Μοντέλο Β) διότι οι αποταυτοποιημένες εικόνες που παράγει μοιάζουν περισσότερο με τις αρχικές και έχουν μεγαλύτερη πιθανότητα να κατηγοριοποιηθούν εσφαλμένα.

Για το Μοντέλο Α, συγκριτικά με την μέθοδο IFGVM το ποσοστό βελτίωσης της l_2 -νόρμας είναι 36.3%, του δείκτη ομοιότητας CW-SSIM είναι 4.0% και του ποσοστού εσφαλμένης κατηγοριοποίησης είναι 0.2%. Επίσης, συγκριτικά με την μέθοδο ITCM το ποσοστό βελτίωσης της l_2 -νόρμας είναι 40.3%, του δείκτη ομοιότητας CW-SSIM είναι 3.3% και του ποσοστού εσφαλμένης κατηγοριοποίησης είναι 0.7%.

Για το Μοντέλο Β, συγκριτικά με την μέθοδο IFGVM το ποσοστό βελτίωσης της l_2 -νόρμας είναι 22.1%, του δείκτη ομοιότητας CW-SSIM είναι 3.4% και του ποσοστού εσφαλμένης

κατηγοριοποίησης είναι 2.9%. Επίσης, συγκριτικά με την μέθοδο ITCM το ποσοστό βελτίωσης της l_2 -νόρμας είναι 63.3%, του δείκτη ομοιότητας CW-SSIM είναι 7.8% και του ποσοστού εσφαλμένης κατηγοριοποίησης είναι 1.6%.

Επίλογος

Οι υπάρχουσες μέθοδοι αποταυτοποίησης προσώπου αδυνατούν να κατασκευάσουν αποταυτοποιημένες εικόνες προσώπου που είναι ρεαλιστικές και μοιάζουν με τις αρχικές. Επίσης, αδυνατούν να διατηρήσουν στο αποταυτοποιημένο πρόσωπο τα μη-ταυτοτικά χαρακτηριστικά του αρχικού, όπως π.χ. το χρώμα του δέρματος, την φυλή, το φύλο, την ηλικία, την συναισθηματική έκφραση ή την πόζα. Σε αυτήν την εργασία προτείναμε την μέθοδο αντιταλικής επίθεσης P-FGVM η οποία επιλύει τα παραπάνω προβλήματα αφού παράγει ως στοχευμένα αντιταλικά δείγματα ρεαλιστικές αποταυτοποιημένες εικόνες προσώπου που μοιάζουν με τις αρχικές. Με τις μεθόδους P-FGVM, IFGVM και ITCM επιτεθήκαμε δυο βαθιά συνελικτικά νευρωνικά δίκτυα που κάνουν αναγνώριση προσώπου σε ένα υποσύνολο του συνόλου δεδομένων CelebA. Επίσης, από την συγκριτική ανάλυση των μεθόδων καταλήξαμε στο συμπέρασμα πως οι αποταυτοποιημένες εικόνες προσώπου της μεθόδου P-FGVM έχουν καλύτερη οπτική ποιότητα καθώς και μεγαλύτερη πιθανότητα να κατηγοριοποιηθούν εσφαλμένα.

Βιβλιογραφία

- [1] Javier Ruiz-del-Solar, Patricio Loncomilla and Naomi Soto. "A Survey on Deep Learning Methods for Robot Vision". arXiv preprint arXiv:1803.10862, 2018.
- [2] Mei Wang and Weihong Deng. "Deep Face Recognition: A Survey". arXiv preprint arXiv:1804.06655, 2018.
- [3] K. W. Bowyer. "Face recognition technology: security versus privacy". In: IEEE Technology and Society Magazine, vol. 23, no. 1, pp. 9-19, 2004.
- [4] J. L. Crowley, J. Coutaz and F. Berard. "Things that See: Machine Perception for Human Computer Interaction". In: Communications of the Association for Computing Machinery, 2000.
- [5] C. Neustaedter and S. Greenberg. "Balancing Privacy and Awareness in Home Media Spaces". In: Workshop on Ubicomp Communities: Privacy as Boundary Negotiation, 2003.
- [6] M. Boyle, C. Edwards and S. Greenberg. "The Effects of Filtered Video on Awareness and Privacy". In: Proceedings of Computer Supported Cooperative Work, 2000.
- [7] E. M. Newton, L. Sweeney and B. Malin. "Preserving privacy by de-identifying face images". In: IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 2, pp. 232-243, 2005.
- [8] Ralph Gross, Edoardo Airola, Bradley Malin and Latanya Sweeney. "Integrating utility into face de-identification". In: International Workshop on Privacy Enhancing Technologies, pp. 227-242, 2005.
- [9] R. Gross, L. Sweeney, F. de la Torre and S. Baker. "Model-based face de-identification". In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, p. 161, 2006.
- [10] Meden B, Emeršič T, Štruc V and Peer P. "k-Same-Net: k-Anonymity with Generative Deep Neural Networks for Face Deidentification". In: Entropy, 20(1):60, 2018.
- [11] Latanya Sweeney. "k-anonymity: A model for protecting privacy". In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002.
- [12] R. Gross, L. Sweeney, J. F. Cohn, F. de la Torre and S. Baker. "Face de-identification". In: Protecting Privacy in Video Surveillance, pp. 129-146, 2009.

- [13] Mosaddegh, Saleh, Loïc Simon and Frédéric Jurie. "Photorealistic Face De-Identification by Aggregating Donors' Face Components". In: Asian Conference on Computer Vision, 2014.
- [14] Blaž Meden, Refik Can Mallı, Sebastjan Fabijan, Hazım Kemal Ekenel, Vitomir Štruc and Peter Peer. "Face Deidentification with Generative Deep Neural Networks". arXiv preprint arXiv:1707.09376, 2017.
- [15] K. Brkić, T. Hrkać, Z. Kalafatić and I. Sikirić. "Face, hairstyle and clothing colour de-identification in video sequences". In: IET Signal Processing, vol. 11, no. 9, pp. 1062-1068, 2017.
- [16] K. Brkic, I. Sikiric, T. Hrkac and Z. Kalafatic. "I Know That Person: Generative Full Body and Face De-identification of People in Images". In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1319-1328, 2017.
- [17] Yifan Wu, Fan Yang and Haibin Ling. "Privacy-Protective-GAN for Face De-identification". arXiv preprint arXiv:1806.08906, 2018.
- [18] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele and Mario Fritz. "Natural and Effective Obfuscation by Head Inpainting". arXiv preprint arXiv:1711.09001, 2017.
- [19] I. Matthews and S. Baker. "Active Appearance Models Revisited". In: International Journal of Computer Vision, vol. 60, no. 2, pp. 135-164, 2004.
- [20] A. Oussidi and A. Elhassouny. "Deep generative models: Survey". In: International Conference on Intelligent Systems and Computer Vision, Fez, pp. 1-8, 2018.
- [21] R. Gross. "Face De-Identification using Multi-Factor Active Appearance Models". PhD thesis, Carnegie Mellon University, 2008.
- [22] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". arXiv preprint arXiv:1409.1556, 2014.
- [23] Ying Nian Wu, Ruiqi Gao, Tian Han and Song-Chun Zhu. "A Tale of Three Probabilistic Families: Discriminative, Descriptive and Generative Models". arXiv preprint arXiv:1810.04261, 2018.
- [24] S. B. Maind and P. Wankar. "Research Paper on Basic of Artificial Neural Network". In: International Journal on Recent and Innovation Trends in Computing and Communication, vol. 2, no. 1, pp. 96-100, 2014.

- [25] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow and Rob Fergus. “Intriguing properties of neural networks”. arXiv preprint arXiv:1312.6199, 2013.
- [26] Alexey Kurakin, Ian Goodfellow and Samy Bengio. “Adversarial examples in the physical world”. arXiv preprint arXiv:1607.02533, 2016.
- [27] Ian J. Goodfellow, Jonathon Shlens and Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. arXiv preprint arXiv:1412.6572, 2014.
- [28] Alexey Kurakin, Ian Goodfellow and Samy Bengio. “Adversarial Machine Learning at Scale”. arXiv preprint arXiv:1611.01236, 2016.
- [29] Andras Rozsa, Ethan M. Rudd and Terrance E. Boult. “Adversarial Diversity and Hard Positive Generation”. arXiv preprint arXiv:1605.01775, 2016.
- [30] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu and Dawn Song. “Generating Adversarial Examples with Adversarial Networks”. arXiv preprint arXiv:1801.02610, 2018.
- [31] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu and Jianguo Li. “Boosting Adversarial Attacks with Momentum”. arXiv preprint arXiv:1710.06081, 2017.
- [32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi and Pascal Frossard. “DeepFool: a simple and accurate method to fool deep neural networks”. arXiv preprint arXiv:1511.04599, 2015.
- [33] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik and Ananthram Swami. “The Limitations of Deep Learning in Adversarial Settings”. arXiv preprint arXiv:1511.07528, 2015.
- [34] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi and Pascal Frossard. “Universal adversarial perturbations”. arXiv preprint arXiv:1610.08401, 2016.
- [35] Nicholas Carlini and David Wagner. “Towards Evaluating the Robustness of Neural Networks”. arXiv preprint arXiv:1608.04644, 2016.
- [36] Jiawei Su, Danilo Vasconcellos Vargas and Sakurai Kouichi. “One pixel attack for fooling deep neural networks”. arXiv preprint arXiv:1710.08864, 2017.
- [37] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik and M. K. Markey. “Complex Wavelet Structural Similarity: A New Image Similarity Index”. In: IEEE Transactions on Image Processing, vol. 18, no. 11, pp. 2385-2401, 2009.

- [38] Ziwei Liu, Ping Luo, Xiaogang Wang and Xiaoou Tang. “Deep Learning Face Attributes in the Wild”. arXiv preprint arXiv:1411.7766, 2014.
- [39] O. M. Parkhi, A. Vedaldi and A. Zisserman. “Deep Face Recognition”. In: British Machine Vision Conference, 2015.
- [40] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. “Generative Adversarial Networks”. arXiv preprint arXiv:1406.2661, 2014.
- [41] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan Yuille, Sangxia Huang, Yao Zhao, Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui and Motoki Abe. “Adversarial Attacks and Defences Competition”. arXiv preprint arXiv:1804.00097, 2018.