

Deep Fake Detection

ΒΑΡΕΛΑΣ ΑΠΟΣΤΟΛΟΣ - ΦΟΙΒΟΣ
mtn2402
ΚΩΝΣΤΑΝΤΟΠΟΥΛΟΣ ΒΑΪΟΣ
mtn2407
Δ.Π.Μ.Σ. ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Ιούνιος 2025



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS



ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΕΡΕΥΝΑΣ
ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ «ΔΗΜΟΚΡΙΤΟΣ»

Μάθημα: Βαθιά Μάθηση

Καθηγητής: Γ. Μπουρίτσας

Περιεχόμενα

Εισαγωγή	3
Dataset και Data Preparation	3
Περιγραφή Dataset	3
Data Preparation	3
No Crop	3
Crop to Faces	4
Μοντέλα και μέθοδος αξιολόγησης	4
Γενικό πλαίσιο εκπαίδευσης και αξιολόγησης των μοντέλων	4
ResNet50	5
Xception	6
Swin Transformer	6
Μέθοδος Αξιολόγησης	6
Αποτελέσματα	7
Ανάλυση Αποτελεσμάτων	9
Επιλεγμένες ROC καμπύλες και Ανάλυση	9
Confusion Matrices και Ανάλυση	10
Heat Maps	11
Συμπεράσματα	14

Εισαγωγή

Τα Deep Fake είναι βίντεο που έχουν τροποποιηθεί με τη χρήση τεχνητής νοημοσύνης, έτσι ώστε να εμφανίζουν ανθρώπους να λένε ή να κάνουν πράγματα που στην πραγματικότητα δεν έγιναν ποτέ. Αυτό γίνεται μέσω ειδικών αλγορίθμων που «μαθαίνουν» πώς να αλλάξουν τα χαρακτηριστικά του προσώπου ή της φωνής κάποιου με πολύ ρεαλιστικό τρόπο.

Τα τελευταία χρόνια, η δημιουργία τέτοιων παραποτημένων βίντεο έχει γίνει πιο εύκολη και διαδεδομένη, με αποτέλεσμα να δημιουργούνται σοβαρά προβλήματα, όπως παραπληροφόρηση ή αλλοίωση της δημόσιας εικόνας κάποιου προσώπου.

Σκοπός αυτής της εργασίας είναι η ανίχνευση Deep Fake βίντεο με τη βοήθεια τεχνικών βαθιάς μάθησης. Πιο συγκεκριμένα, μελετήθηκαν τέσσερα διαφορετικά μοντέλα νευρωνικών δικτύων και εφαρμόστηκαν σε ένα σύνολο δεδομένων που περιλαμβάνει πραγματικά και παραποτημένα βίντεο.

Η εργασία περιλαμβάνει την προετοιμασία των δεδομένων, την εκπαίδευση των μοντέλων, την αξιολόγησή τους και τη σύγκριση των αποτελεσμάτων, με στόχο να εντοπιστεί ποιο μοντέλο είναι πιο αποτελεσματικό στην αναγνώριση Deep Fake περιεχομένου.

Dataset και Data Preparation

Περιγραφή Dataset

Για την εκπαίδευση και αξιολόγηση των μοντέλων χρησιμοποιήθηκε το σύνολο δεδομένων **FaceForensics++**, το οποίο είναι διαθέσιμο μέσω της πλατφόρμας Kaggle. Η συγκεκριμένη έκδοση περιλαμβάνει συνολικά 400 βίντεο, εκ των οποίων τα 200 είναι πραγματικά (real) και τα 200 παραποτημένα (fake). Τα βίντεο οργανώνονται σε δύο φακέλους με ονομασίες **real** και **fake**, με διάρκεια να κυμαίνεται από 1.38 δευτερόλεπτα μέχρι 67.50 δευτερόλεπτα.

Data Preparation

Η προετοιμασία δεδομένων είναι κοινό σημείο εκκίνησης για όλα τα μοντέλα που εκπαιδεύτηκαν. Ανεξάρτητα από το αν χρησιμοποιούμε ολόκληρα τα καρέ ή περικοπές προσώπου, η βασική διαδικασία είναι η ίδια: κάθε βίντεο διασχίζεται για να επιλεγούν ένα σταθερό πλήθος καρέ (συνήθως 30), τα οποία αναπαριστούν με αντιπροσωπευτικό τρόπο τη χρονική του εξέλιξη.

Η δειγματοληψία γίνεται σε σταθερά χρονικά διαστήματα: υπολογίζεται πόσα καρέ έχει συνολικά το βίντεο και στη συνέχεια επιλέγονται 30 καρέ με ίση απόσταση μεταξύ τους. Αυτή η προσέγγιση εξασφαλίζει ότι δεν επηρεάζεται η κάλυψη ενός βίντεο είτε είναι μικρής είτε μεγάλης διάρκειας.

Αφού συγκεντρωθούν όλα τα καρέ για κάθε βίντεο, ακολουθεί ο χωρισμός σε τρία υποσύνολα: - Εκπαίδευσης (75%) - Επικύρωσης (12.5%) - Δοκιμών (12.5%)

Η διαίρεση γίνεται με τυχαίο τρόπο, αλλά χρησιμοποιείται σταθερός random seed ώστε η διαδικασία να είναι αναπαραγώγιμη. Στο τέλος, το κάθε καρέ περνάει από μια σειρά από μετασχηματισμούς εικόνας, ανάλογα με το αν προέρχεται από το σύνολο εκπαίδευσης ή από τα άλλα δύο (όπου χρησιμοποιούνται πιο «ουδέτεροι» μετασχηματισμοί).

No Crop

Η απλούστερη προσέγγιση είναι να χρησιμοποιηθεί ολόκληρο το καρέ όπως εμφανίζεται στο βίντεο, χωρίς να προσπαθήσουμε να ανιχνεύσουμε κάποιο πρόσωπο. Αυτό κάνει την

όλη διαδικασία πιο γρήγορη και πιο σταθερή, αφού δεν εξαρτάται από την ποιότητα της ανίχνευσης προσώπου. Έτσι, ακόμα και αν το πρόσωπο είναι κρυμμένο, λοξό ή λείπει εντελώς, το δείγμα εξακολουθεί να μπορεί να χρησιμοποιηθεί.

Τα καρέ μετατρέπονται σε RGB και μετά εφαρμόζονται τυπικές τεχνικές προεπεξεργασίας: αλλαγή μεγέθους, περικοπή, τυχαία περιστροφή και παραμόρφωση χρώματος (μόνο στην εκπαίδευση). Η λογική εδώ είναι ότι το μοντέλο μπορεί να ανιχνεύσει deepfake χαρακτηριστικά που δεν περιορίζονται αποκλειστικά στο πρόσωπο (π.χ. παραμορφώσεις στο λαιμό, περίεργο φωτισμό, artefacts στο φόντο).

Crop to Faces

Η προσέγγιση «Crop to Faces» βασίζεται στην υπόθεση ότι τα σημαντικότερα σημάδια παραποίησης εντοπίζονται στο ίδιο το πρόσωπο. Άρα πριν χρησιμοποιήσουμε ένα καρέ, προσπαθούμε πρώτα να εντοπίσουμε το πρόσωπο.

Η διαδικασία ξεκινάει με την ανίχνευση περιοχής προσώπου σε κάθε καρέ, με χρήση δύο εργαλείων: 1. **MediaPipe Face Detection**, με δύο εσωτερικά μοντέλα (`model_selection=0` και `1`). 2. **MTCNN**, ως εναλλακτική σε περίπτωση που τα παραπάνω αποτύχουν.

Αν βρεθεί πρόσωπο, υπολογίζεται το bounding box, προστίθεται ένα περιθώριο γύρω του και το περιεχόμενο περικόπτεται και αναδιαστασιογείται σε 224×224 . Αν δεν εντοπιστεί πρόσωπο, γίνεται έως και 5 φορές προσπάθεια με άλλο τυχαίο καρέ από το ίδιο βίντεο. Αν και πάλι δεν υπάρχει επιτυχία, τότε το καρέ κρατείται ολόκληρο (χωρίς cropping) και προσαρμόζεται αναγκαστικά στο ίδιο μέγεθος.

Οι περικομμένες εικόνες μετατρέπονται σε tensors και δέχονται σειρά από τυχαίους μετασχηματισμούς (όπως περιστροφή, flip, jitter χρωμάτων), με σκοπό να αυξήσουμε τη γενίκευση του μοντέλου.

Ο συνδυασμός ανίχνευσης προσώπου και δυναμικών augmentations βοηθάει τα μοντέλα να εστιάσουν καλύτερα σε σημεία που πιθανώς προδίδουν την ύπαρξη παραποίησης. Παρ' όλα αυτά, η προσέγγιση αυτή είναι πιο ευαίσθητη στην ποιότητα των frames και πιο αργή λόγω των ελέγχων ανίχνευσης.

Μοντέλα και μέθοδος αξιολόγησης

Γενικό πλαίσιο εκπαίδευσης και αξιολόγησης των μοντέλων

Σε όλα τα μοντέλα που δοκιμάζουμε χρησιμοποιούμε το DeepFakeFrameDataset, το οποίο αποσπά καρέ από τα βίντεο και παρέχει τρεις DataLoader: train, val και test. Στο train εφαρμόζουμε τυχαίες μετατροπές (augmentations), όπως περιστροφή, τυχαίο κλιμάρισμα, οριζόντια αντιστροφή και color jitter, ενώ στα val/test απλώς αλλάζουμε μέγεθος, μετατρέπουμε σε τανσορ και κανονικοποιούμε με σταθερούς μέσους όρους και τυπικές αποκλίσεις RGB.

Κατά την εκπαίδευση, ορίζουμε batch size = 8 και learning rate = $1e-4$, χρησιμοποιώντας Adam optimizer. Ενεργοποιούμε mixed precision training (μέσω `torch.cuda.amp` ή `autocast` και `GradScaler`) για ταχύτερη εκτέλεση και μικρότερη χρήση μνήμης. Η συνάρτηση απώλειας είναι η Cross-Entropy. Έχουμε scheduler (ReduceLROnPlateau) που μειώνει το learning rate κατά 0.1 όταν η validation loss δεν βελτιώνεται για δύο συνεχόμενες εποχές, και early stopping αν δεν υπάρχει βελτίωση για πέντε συνεχόμενες εποχές.

Μετά από κάθε εποχή:

1. Βάζουμε το μοντέλο σε eval() και περνάμε όλα τα καρέ του validation set.
2. Για κάθε καρέ παίρνουμε πιθανότητες και προβλέψεις.

3. Συγκεντρώνουμε τα αποτελέσματα ανά βίντεο και εφαρμόζουμε έξι μεθόδους για να ενώσουμε τις frame level προβλέψεις σε video level:

- mean_prob
- max_prob
- min_prob
- median_prob
- trimmed_mean_prob
- vote_percentage

4. Για κάθε μέθοδο υπολογίζουμε accuracy, precision, recall, F1, Matthews correlation, AUC, Average Precision και Equal Error Rate.
5. Επιλέγουμε ως καλύτερη μέθοδο αυτή με το υψηλότερο F1.
6. Αν η τρέχουσα validation loss είναι η χαμηλότερη που έχουμε μέχρι στιγμής, σώζουμε τα βάρη.

Όταν ολοκληρωθεί η εκπαίδευση (ή ενεργοποιηθεί early stopping), σχεδιάζουμε τα γραφήματα:

- train vs val loss ανά εποχή
- train vs val accuracy ανά εποχή
- F1 curve στο validation ανά εποχή
- F1 vs validation loss

και τα αποθηκεύουμε σε PNG αρχεία.

ResNet50

Το ResNet50 είναι ένα δημοφιλές συνελικτικό νευρωνικό δίκτυο, γνωστό για τη χρήση υπολειμματικών συνδέσεων (residual connections), οι οποίες επιτρέπουν την εκπαίδευση πολύ βαθιών μοντέλων χωρίς να εμφανίζεται το πρόβλημα της υποβάθμισης του σήματος. Ουσιαστικά, κάθε μπλοκ του μοντέλου δεν μαθαίνει το πλήρες ουτρυτ, αλλά τη διαφορά (υπόλειμμα) μεταξύ εισόδου και εξόδου, με αποτέλεσμα την ταχύτερη και σταθερότερη εκπαίδευση.

Στο πλαίσιο της παρούσας εργασίας, χρησιμοποιούμε το προεκπαιδευμένο ResNet50 και το προσαρμόζουμε για ταξινόμηση δύο κατηγοριών (αληθινό ή ψεύτικο). Τοποθετούμε ένα νέο πλήρως συνδεδεμένο σύστημα στο τέλος του μοντέλου, το οποίο αποτελείται από ένα επίπεδο με 512 νευρώνες, ενεργοποίηση ReLU, dropout για regularization και τελική έξοδο 2 τιμών (μία για κάθε κατηγορία). Κατά την εκπαίδευση, «παγώνουμε» τα περισσότερα από τα αρχικά επίπεδα του ResNet και επιτρέπουμε την προσαρμογή μόνο των τελευταίων στρωμάτων, ώστε το μοντέλο να μάθει τα χαρακτηριστικά που σχετίζονται με την ανίχνευση deep fake χωρίς να ξεχάσει τη γενική γνώση που απέκτησε από τα αρχικά δεδομένα.

Xception

Το Xception είναι ένα εξελιγμένο συνελικτικό μοντέλο που βασίζεται στην ιδέα των depthwise separable convolutions. Αντί να εφαρμόζει μία κανονική συνελικτική λειτουργία, η οποία συνδυάζει ταυτόχρονα τα χωρικά και καναλικά χαρακτηριστικά, το Xception διαχωρίζει αυτά τα δύο στάδια. Πρώτα εφαρμόζει χωριστά συνελικτικές πράξεις για κάθε κανάλι (depthwise convolution) και στη συνέχεια συνδυάζει τα αποτελέσματα μέσω σημειακής συνελικτικής πράξης (pointwise convolution). Αυτό οδηγεί σε καλύτερη αποδοτικότητα και δυνατότητα μοντελοποίησης πολύπλοκων σχέσεων στα δεδομένα.

Για τις ανάγκες της παρούσας εργασίας, αξιοποιούμε το προεκπαιδευμένο Xception και το επεκτείνουμε προσθέτοντας έναν μηχανισμό προσοχής τύπου SE (Squeeze-and-Excitation). Ο μηχανισμός αυτός προσαρμόζει δυναμικά τη σημασία κάθε καναλιού εξόδου, ενισχύοντας έτσι τα πιο κρίσιμα χαρακτηριστικά του καρέ. Στην έξοδο του μοντέλου προσθέτουμε ένα απλό πλήρως συνδεδεμένο επίπεδο που προβλέπει την τελική κατηγορία (αληθινό ή ψεύτικο).

Κατά την εκπαίδευση, αφήνουμε παγωμένα τα περισσότερα στρώματα του Xception και ξεπαγώνουμε μόνο τα τελευταία (block6, block7, block8), καθώς και το τελικό conv layer. Με αυτό τον τρόπο διατηρούμε τις γενικές ικανότητες του μοντέλου, ενώ του επιτρέπουμε να μάθει τα χαρακτηριστικά που σχετίζονται ειδικά με την ανίχνευση deep fake.

Swin Transformer

Ο Swin Transformer αποτελεί μία σύγχρονη προσέγγιση στην επεξεργασία εικόνων, βασισμένη όχι σε συνελικτικά φίλτρα αλλά σε μηχανισμούς προσοχής (attention). Σε αντίθεση με τα κλασικά Vision Transformers, ο Swin Transformer χωρίζει την εικόνα σε μικρά παράθυρα (windows) και εφαρμόζει self-attention τοπικά σε κάθε παράθυρο. Αυτό μειώνει σημαντικά το υπολογιστικό κόστος και επιτρέπει στο μοντέλο να μαθαίνει ιεραρχικά χαρακτηριστικά, όπως κάνουν και τα συνελικτικά μοντέλα.

Για την εργασία μας, χρησιμοποιούμε το προεκπαιδευμένο μοντέλο Swin-Base, στο οποίο προσθέτουμε έναν επιπλέον μηχανισμό προσοχής τύπου SE (Squeeze-and-Excitation), όπως και στο Xception. Ο SE μηχανισμός εφαρμόζεται στο τελικό διάνυσμα χαρακτηριστικών πριν την έξοδο, ενισχύοντας τα πιο σημαντικά κανάλια πληροφορίας. Τέλος, προσθέτουμε έναν μικρό ταξινομητή που προβλέπει την κατηγορία του καρέ.

Κατά την εκπαίδευση, αρχικά αφήνουμε παγωμένα σχεδόν όλα τα επίπεδα του backbone, εκτός από τον ταξινομητή, τον μηχανισμό SE και το τελευταίο block του Swin. Καθώς προχωρούν οι εποχές, ξεπαγώνουμε σταδιακά και τα υπόλοιπα blocks (από layer 3 προς τα layer 1), ώστε το μοντέλο να μπορεί να προσαρμοστεί πιο στοχευμένα στο πρόβλημα ανίχνευσης deep fake, χωρίς να χάσει τη γενική γνώση που έχει ήδη μάθει.

Μέθοδος Αξιολόγησης

Η αξιολόγηση των μοντέλων γίνεται με τρόπο που λαμβάνει υπόψη όχι μόνο την ακρίβεια σε επίπεδο καρέ, αλλά και την τελική απόφαση σε επίπεδο βίντεο, αφού το κάθε καρέ είναι μόνο ένα μέρος του συνόλου. Το τελικό αποτέλεσμα για κάθε βίντεο προκύπτει με βάση την απόδοση του μοντέλου σε πολλαπλά καρέ και τη συνάθροισή τους με διαφορετικές μεθόδους.

Αρχικά, για κάθε καρέ που περνάει από το δίκτυο, κρατιούνται η πρόβλεψη, η πιθανότητα (softmax) και η πραγματική ετικέτα. Όλα αυτά αποθηκεύονται οργανωμένα ανά βίντεο, ώστε να μπορούμε μετά να πάρουμε συνολικά σκορ για κάθε ένα από αυτά.

Η συνάθροιση των προβλέψεων σε επίπεδο βίντεο γίνεται με έξι διαφορετικές στρατηγικές:

- μέσος όρος των πιθανοτήτων (mean_prob),

- μέγιστη ή ελάχιστη πιθανότητα,
- διάμεση τιμή (median_prob),
- trimmed mean (μέσος όρος αφού αφαιρεθούν τα άκρα),
- ποσοστό ψήφων για fake label (vote_percentage).

Κάθε μέθοδος αξιολογείται ξεχωριστά με τις βασικές μετρικές (accuracy, precision, recall, F1, MCC, AP και EER). Στο τέλος, κρατιέται εκείνη που είχε το καλύτερο F1 score ως η “βέλτιστη στρατηγική συνάθροισης” για το συγκεκριμένο μοντέλο.

Με βάση τη βέλτιστη αυτή μέθοδο, παράγονται επίσης:

- ο **πίνακας σύγχυσης** (confusion matrix),
- η **καμπύλη ROC**,
- η **καμπύλη precision-recall**.

Επιπλέον, γίνεται καταγραφή των αποτελεσμάτων σε δύο αρχεία CSV: ένα για τις σωστές προβλέψεις και ένα για τα λάθη. Αυτά τα αρχεία περιλαμβάνουν πληροφορίες όπως το όνομα του βίντεο, το πλήθος των καρέ του, η πραγματική και η προβλεπόμενη ετικέτα, η εμπιστοσύνη του μοντέλου και αν το αποτέλεσμα ήταν TP, TN, FP ή FN.

Για πιο αναλυτική κατανόηση της συμπεριφοράς του μοντέλου, δημιουργούνται επίσης **εικόνες Grad-CAM**. Για τέσσερις κατηγορίες (TP, TN, FP, FN), επιλέγονται τυχαία 3 παραδείγματα και για κάθε βίντεο παράγεται μια σύνθετη εικόνα που δείχνει θερμικούς χάρτες (heatmaps) για όλα τα επιλεγμένα καρέ. Οι θερμικοί χάρτες παράγονται με χρήση του GradCAM και δείχνουν σε ποια σημεία της εικόνας το δίκτυο «κοίταξε» περισσότερο όταν πήρε την απόφαση του.

Η συνολική διαδικασία αξιολόγησης είναι κοινή είτε πρόκειται για δεδομένα με περικομένα πρόσωπα είτε για πλήρη καρέ. Η μόνη διαφορά είναι στο αν εφαρμόζεται ή όχι face cropping πριν την πρόβλεψη.

Αποτελέσματα

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα από τα πειράματα που εκτελέστηκαν με τα τρία μοντέλα: **ResNet50**, **Xception** και **Swin Transformer**. Κάθε μοντέλο αξιολογήθηκε σε τέσσερα διαφορετικά σενάρια, ανάλογα με τον τύπο δεδομένων που χρησιμοποιήθηκε κατά την εκπαίδευση (faces ή frames) και τον τύπο δεδομένων που χρησιμοποιήθηκε κατά τη δοκιμή. Οι τέσσερις συνδυασμοί είναι:

- Εκπαίδευση σε πρόσωπα και δοκιμή σε πρόσωπα (faces → faces)
- Εκπαίδευση σε πρόσωπα και δοκιμή σε ολόκληρα καρέ (faces → frames)
- Εκπαίδευση σε καρέ και δοκιμή σε πρόσωπα (frames → faces)
- Εκπαίδευση σε καρέ και δοκιμή σε καρέ (frames → frames)

Για κάθε μοντέλο παρουσιάζεται ένας πίνακας με την απόδοσή του σε αυτούς τους τέσσερις συνδυασμούς. Στις γραμμές περιλαμβάνονται οι βασικές μετρικές αξιολόγησης:

- **Aggregation**: η βέλτιστη μέθοδος συνάθροισης από frame-level σε video-level
- **Accuracy**: ποσοστό σωστών προβλέψεων
- **Precision**: πόσα από τα προβλεπόμενα fake ήταν πράγματι fake

- **Recall:** πόσα από τα πραγματικά fake θρέθηκαν
- **F1-score:** ο αρμονικός μέσος Precision και Recall
- **MCC:** Συντελεστής Συσχέτισης του Matthews, συνολική ποιότητα δυαδικής ταξινόμησης
- **AP:** Μέση Ακρίβεια (εμβαδό κάτω από την καμπύλη precision-recall)
- **EER:** Ρυθμός Ίσου Σφάλματος (FAR = FRR)

Οι πίνακες που ακολουθούν δείχνουν αναλυτικά τις τιμές για κάθε συνδυασμό, επιτρέποντας τη σύγκριση της απόδοσης των μοντέλων υπό διαφορετικές συνθήκες.

Metric	faces→faces	faces→frames	frames→faces	frames→frames
Aggregation	vote_percentage	max_prob	max_prob	max_prob
Accuracy	0.8800	0.5400	0.5000	0.5400
Precision	0.8148	0.5000	0.4792	0.5000
Recall	0.9565	0.6522	1.0000	0.9130
F1 Score	0.8800	0.5660	0.6479	0.6462
MCC	0.7713	0.0983	0.1884	0.1839
Average Precision	0.9750	0.4899	0.4951	0.4405
EER	0.0741	0.4815	0.4444	0.5185

Πίνακας 1: Αποτελέσματα μοντέλου - ResNet50.

Metric	faces→faces	faces→frames	frames→faces	frames→frames
Aggregation	mean_prob	max_prob	max_prob	max_prob
Accuracy	0.9000	0.4400	0.4800	0.5400
Precision	0.8750	0.4194	0.4516	0.5000
Recall	0.9130	0.5652	0.6087	0.7826
F1 Score	0.8936	0.4815	0.5185	0.6102
MCC	0.8000	-0.1042	-0.0215	0.1287
Average Precision	0.9698	0.4733	0.4775	0.4161
EER	0.0370	0.5185	0.5556	0.5185

Πίνακας 2: Αποτελέσματα μοντέλου - Xception.

Metric	faces→faces	faces→frames	frames→faces	frames→frames
Aggregation	median_prob	max_prob	mean_prob	max_prob
Accuracy	0.8600	0.4600	0.6600	0.5400
Precision	0.7857	0.4231	0.5833	0.5000
Recall	0.9565	0.4783	0.9130	0.7826
F1 Score	0.8627	0.4490	0.7119	0.6102
MCC	0.7373	-0.0771	0.3968	0.1287
Average Precision	0.9678	0.5081	0.5763	0.3930
EER	0.1111	0.5185	0.3704	0.5556

Πίνακας 3: Αποτελέσματα μοντέλου - Swin Transformer.

Ανάλυση Αποτελεσμάτων

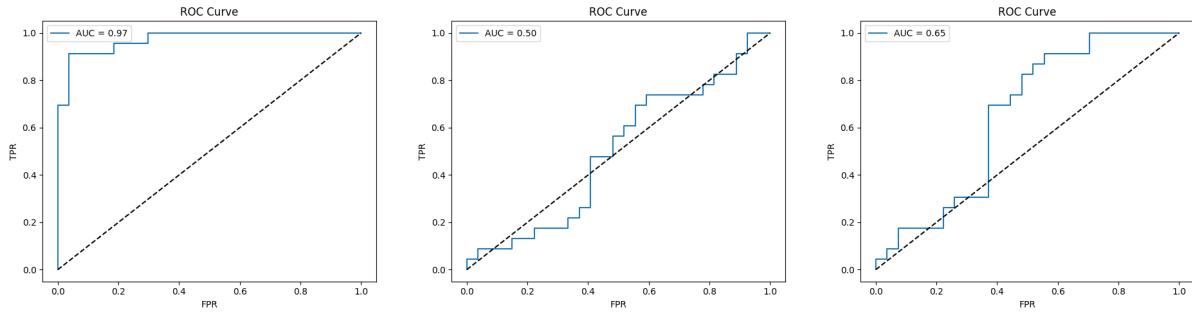
1) Σύγκριση μεταξύ των μοντέλων Στο σενάριο faces→faces, όπου όλα τα μοντέλα αξιολογούνται στις περικομμένες περιοχές προσώπου, το Xception ξεχωρίζει με F1-score 0.8936, ακολουθούμενο από το ResNet50 με 0.8800 και το Swin Transformer με 0.8627. Η μέση ακρίβεια (AP) και ο δείκτης Equal Error Rate (EER) επιβεβαιώνουν αυτή τη σειρά: το Xception πετυχαίνει την υψηλότερη AP (0.9698) και το χαμηλότερο EER (0.0370), το ResNet50 ακολουθεί με AP=0.9750 και EER=0.0741, ενώ το Swin Transformer παρουσιάζει ελαφρώς υποδεέστερη AP=0.9678 και υψηλότερο EER=0.1111. Αυτό δείχνει ότι το Xception, χάρη στις depthwise separable convolutions και το SE block, είναι πιο αποδοτικό στην ανίχνευση λεπτομερειών που χαρακτηρίζουν τα deepfakes σε επίπεδο προσώπου.

2) Σύγκριση μεταξύ των πειραμάτων Σε όλες τις αρχιτεκτονικές η καλύτερη απόδοση παρατηρείται στο faces→faces. Όταν όμως εκπαιδεύουμε σε πρόσωπα και δοκιμάζουμε σε ολόκληρα καρέ (faces→frames), η απόδοση καταποντίζεται (F1 0.48–0.57), καθώς τα μοντέλα δεν έχουν δει ποτέ κόντινγκ φόντου ή άλλα εκτός προσώπου χαρακτηριστικά. Αντίθετα, στο σενάριο frames→frames, όπου η εκπαίδευση και η δοκιμή γίνονται σε πλήρη καρέ, η πρόβλεψη βελτιώνεται (F1 0.61–0.65), υποδεικνύοντας καλύτερη γενίκευση σε “πραγματικά” βίντεο με ποικιλία περιεχομένου. Το πιο ενδιαφέρον εύρημα είναι το frames→faces για το Swin Transformer, με F1 = 0.7119, που υποδεικνύει ότι τα attention based χαρακτηριστικά του Swin μπορούν να εντοπίσουν χρήσιμα μοτίβα προσώπου ακόμη και όταν έχουν εκπαίδευτεί σε πλήρη καρέ.

3) Ερμηνεία των αποτελεσμάτων Η απόδοση στα σενάρια faces faces δείχνει ότι η εστίαση μόνον στο πρόσωπο (crop) μειώνει το “θόρυβο” από φόντο και περιφερειακά artefacts, επιτρέποντας στα μοντέλα να μάθουν τα πραγματικά deepfake χαρακτηριστικά. Όταν όμως απαιτείται γενίκευση σε ολόκληρα καρέ, η έλλειψη εκπαίδευσης σε background patterns επιδεινώνει τα αποτελέσματα (faces frames). Αντίστοιχα, εκπαίδευση σε καρέ (frames frames) δίνει πιο ισορροπημένη απόδοση, καθώς τα augmentations και οι τυχαίες μεταμορφώσεις εκπαιδεύουν τα δίκτυα σε μεγαλύτερη ποικιλία οπτικών συνθηκών. Τέλος, το μοντέλο Xception αποδίδει καλύτερα σε επίπεδο face crop χάρη στην ικανότητά του να αντιλαμβάνεται λεπτομερή pixel level artefacts, ενώ το Swin Transformer φαίνεται πιο ευέλικτο όταν πρόκειται για cross domain σενάρια (frames faces), πιθανότατα λόγω του μηχανισμού self attention που εντοπίζει σχέσεις μεταξύ απομακρυσμένων περιοχών εικόνας.

Επιλεγμένες ROC καμπύλες και Ανάλυση

Παρακάτω παρουσιάζονται τρεις αντιπροσωπευτικές ROC καμπύλες, για κάθε ένα από τα “άκρα” και το ενδιαφέρον σενάριο generalization:



(a') Xception στο faces→faces (b') ResNet50 στο faces→frames (γ') Swin Transformer στο frames→faces (AUC = 0.65)

Σχημα 1: ROC καμπύλες για τρία επιλεγμένα σενάρια: (a) Xception στο ιδανικό faces→faces, (b) ResNet50 στο domain shift faces→frames, (γ) Swin Transformer στο αντίστροφο shift frames→faces.

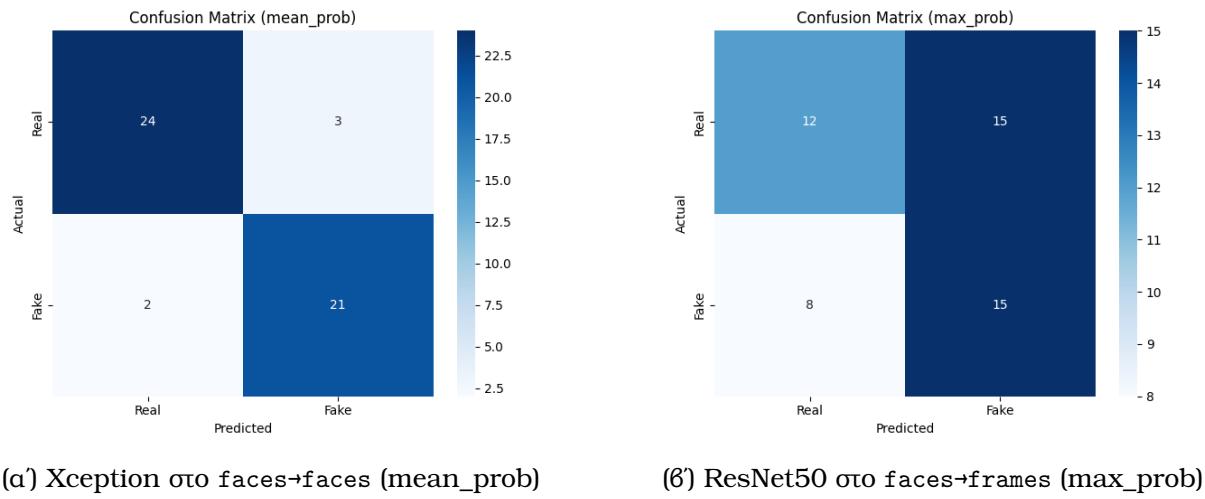
Ερμηνεία αποτελεσμάτων

- **Xception (faces→faces, AUC = 0.97):** Εξαιρετική καμπύλη με υψηλή TPR και πολύ χαμηλή FPR σε όλο το εύρος threshold, επιβεβαιώνοντας ότι το Xception μαθαίνει αξιόπιστα τα pixel level χαρακτηριστικά των deepfakes όταν εστιάζει μόνο στο πρόσωπο.
- **ResNet50 (faces→frames, AUC = 0.50):** Η καμπύλη σχεδόν συμπίπτει με τη διαγώνιο, υποδεικνύοντας τυχαία πρόβλεψη. Το ResNet50 δεν γενικεύει καθόλου από cropped σε πλήρη καρέ, καθώς δεν έχει εκπαιδευτεί σε περιφερειακά στοιχεία του πλαισίου.
- **Swin Transformer (frames→faces, AUC = 0.65):** Ενδιάμεση απόδοση: το self attention συλλαμβάνει μερικά χρήσιμα μοτίβα προσώπου ακόμα και όταν έχει εκπαιδευτεί σε full frame inputs, αλλά δεν φτάνει το επίπεδο του Xception στο καθαρό face crop σενάριο.

Αυτές οι καμπύλες δείχνουν ξεκάθαρα πώς: (a) το Xception είναι κυρίαρχο στο ελεγχόμενο περιβάλλον προσώπου, (b) το ResNet50 “καταρρέει” με domain shift από πρόσωπο σε full frame, και (γ) το Swin Transformer προσφέρει καλύτερη generalization στο αντίστροφο σενάριο.

Confusion Matrices και Ανάλυση

Στο παρακάτω σχήμα φαίνονται δύο αντιπροσωπευτικές confusion matrices:



Σχημα 2: Confusion matrices για δύο αντιπροσωπευτικά σενάρια: (a) Xception στο καθαρό crop προσώπου, (b) ResNet50 με domain shift σε full-frame.

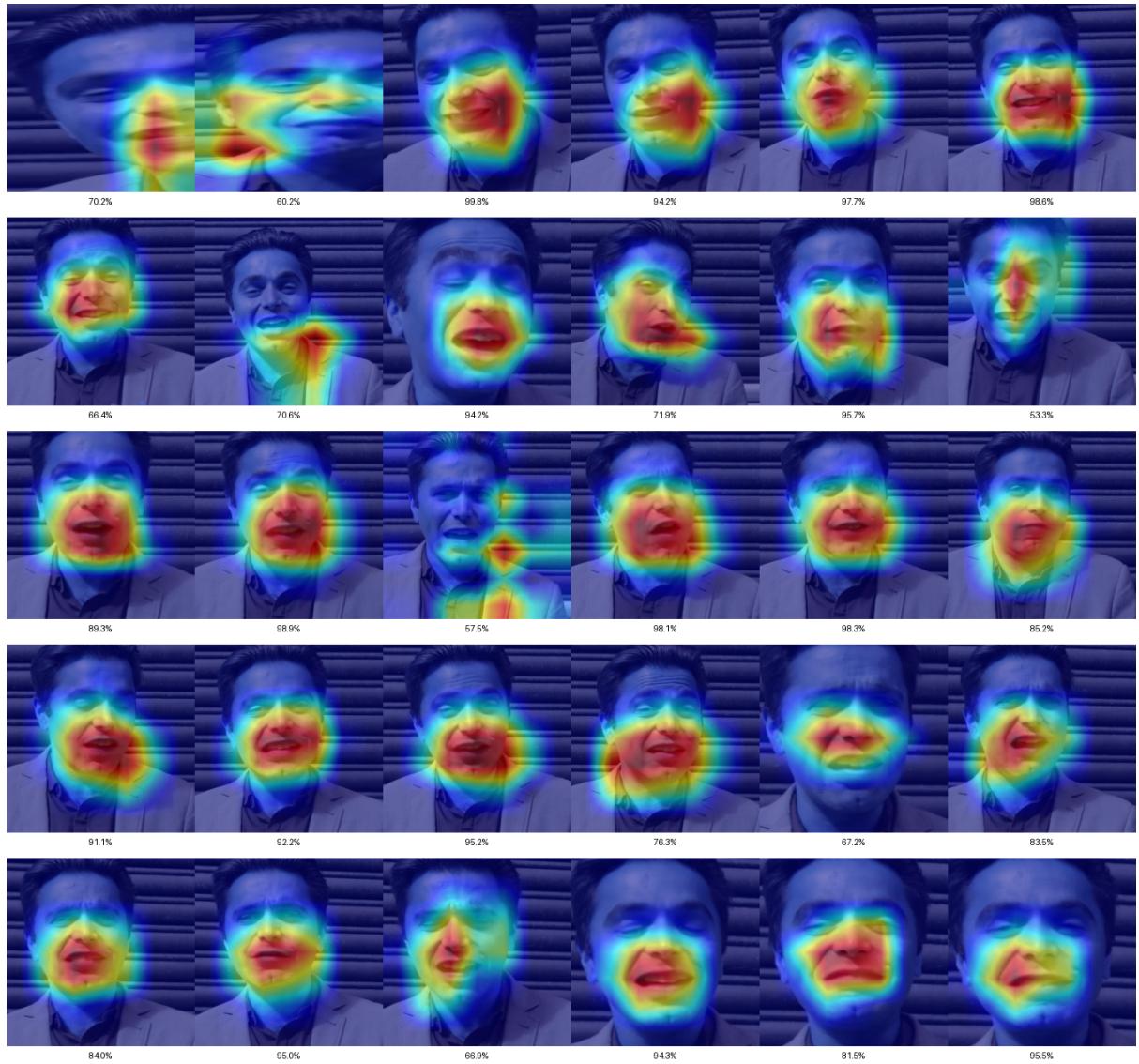
Ερμηνεία

- **Xception (faces→faces):** Από τα 27 πραγματικά πρόσωπα, 24 ταξινομήθηκαν σωστά (True Negatives) και μόνο 3 σημειώθηκαν False Positives. Από τα 23 παραπομένα, 21 εντοπίστηκαν (True Positives) και 2 χάθηκαν (False Negatives). Αυτό οδηγεί σε πολύ υψηλή ακρίβεια και recall, επιβεβαιώνοντας την ισχυρή απόδοση του Xception σε cropped input.
- **ResNet50 (faces→frames):** Σε 27 πραγματικά frames, μόνο 12 αναγνωρίστηκαν σωστά, ενώ 15 θεωρήθηκαν λάθος (False Positives). Σε 23 fake frames, μόλις 15 επισημάνθηκαν σωστά (True Positives) και 8 χάθηκαν (False Negatives). Η άνιση κατανομή των σφαλμάτων δείχνει ότι το ResNet50 δυσκολεύεται να ξεχωρίσει το fake από το πραγματικό όταν το context αλλάζει — υποδεικνύοντας την ανάγκη εκπαίδευσης σε full-frame δεδομένα ή την προσθήκη πιο ισχυρών augmentations.

Heat Maps

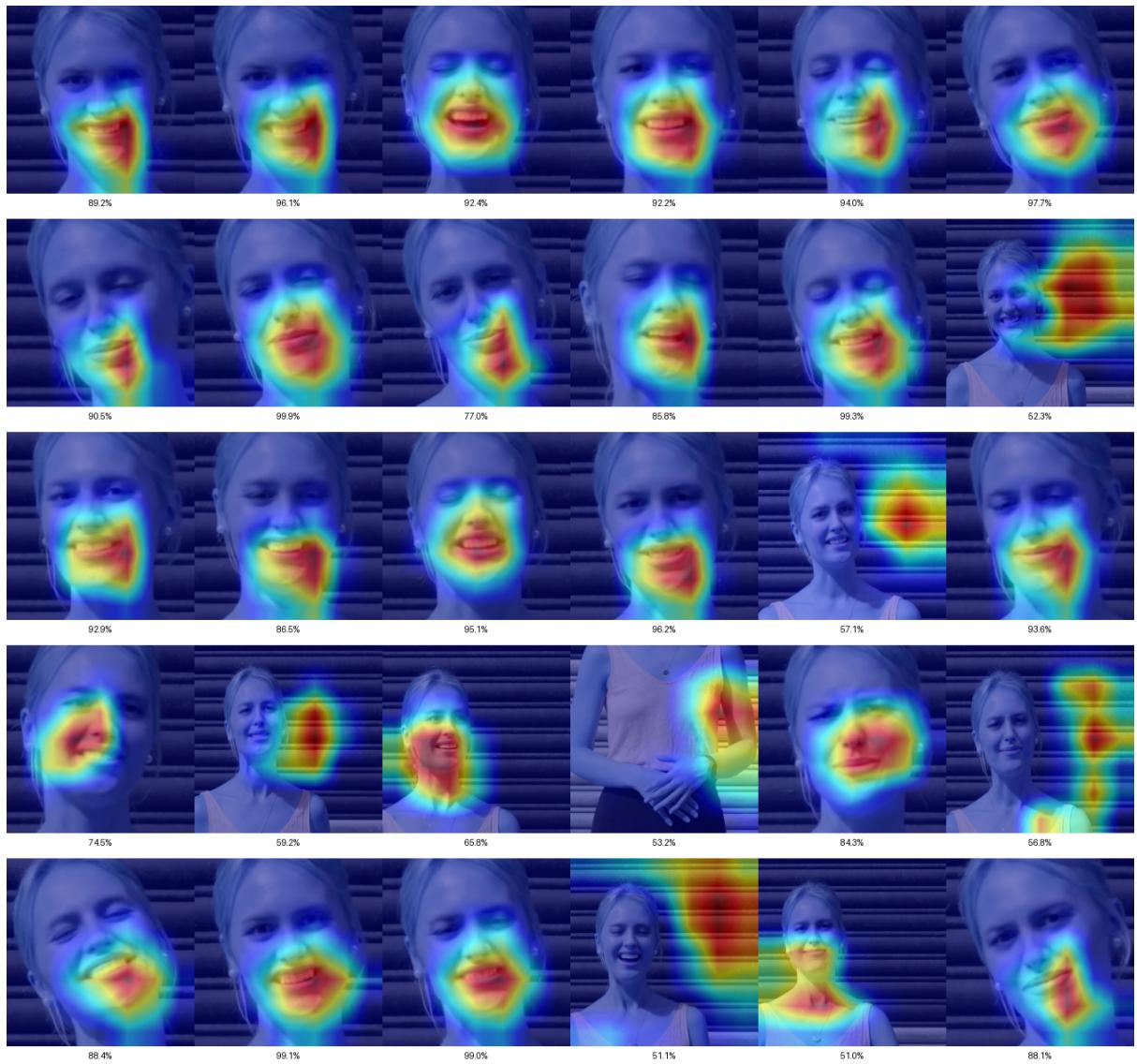
Παρακάτω παρουσιάζονται μερικά heatmaps που μας βοηθούν να καταλάβουμε καλύτερα πώς σκέφτονται τα μοντέλα. Κάθε εικόνα δείχνει τα heatmaps όλων των frames από ένα βίντεο, καθώς και την "βεβαιότητα" του μοντέλου για το κάθε frame (πόσο πιθανό είναι να είναι fake).

Ξεκινάμε με δύο παραδείγματα από το μοντέλο Xception. Στην πρώτη εικόνα, βλέπουμε ένα βίντεο που ήταν όντως ψεύτικο και το μοντέλο κατάφερε να το εντοπίσει σωστά. Παρατηρούμε ότι έχει εστιάσει στο στόμα, που είναι και το σημείο που έχει τροποποιηθεί. Το συγκεκριμένο πείραμα ήταν face→face, δηλαδή το μοντέλο εκπαιδεύτηκε και δοκιμάστηκε σε πρόσωπα.



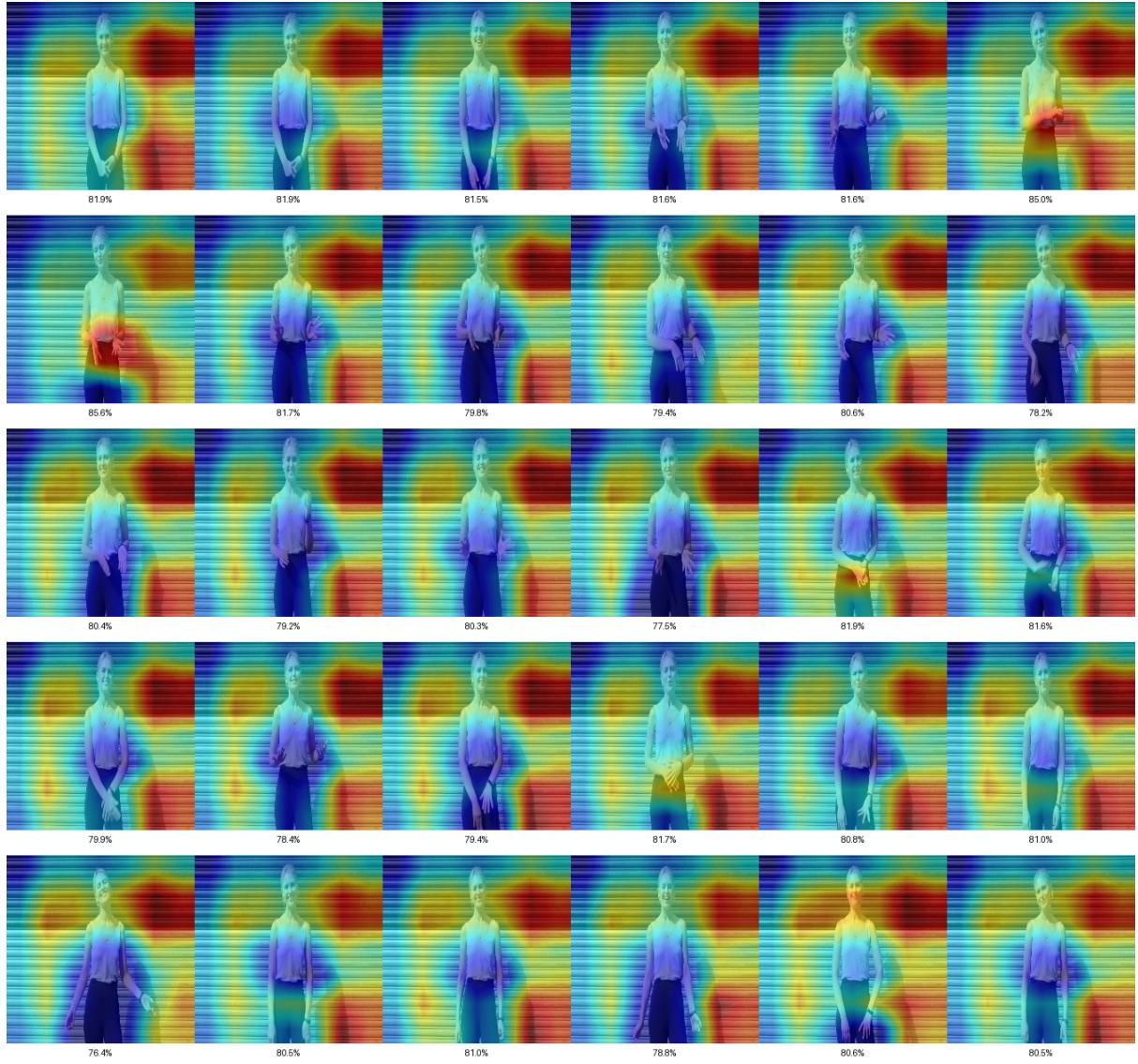
Σχημα 3: Grad-CAM heatmaps για ένα TP του Xception (faces→faces).

Στη δεύτερη εικόνα, έχουμε ένα βίντεο που ήταν fake, αλλά το Xception δεν κατάφερε να το αναγνωρίσει. Και πάλι μιλάμε για το πείραμα face→face. Το πρόβλημα εδώ είναι ότι η παραποίηση είναι πολύ λεπτή και δεν φαίνεται ξεκάθαρα στα μεμονωμένα καρέ. Συγκεκριμένα, η παραμόρφωση γίνεται στο στόμα, το οποίο φαίνεται να έχει glitch μόνο όταν το βίντεο παίζει, κάτι που δεν αποτυπώνεται καλά όταν βλέπουμε ένα frame τη φορά.



Σχημα 4: Grad-CAM heatmaps για ένα FN του Xception (faces→faces).

Τέλος, έχουμε ένα παράδειγμα από το ResNet50 σε ένα πείραμα όπου εκπαιδεύτηκε μόνο σε πρόσωπα, αλλά κατά το testing δώσαμε ολόκληρα καρέ (face→frames). Όπως φαίνεται, το μοντέλο δεν καταφέρνει καν να εντοπίσει το πρόσωπο, και έτσι βασίζεται σε άλλα σημεία της εικόνας που δεν σχετίζονται με την παραποίηση. Αυτός είναι και ένας λόγος που το συγκεκριμένο μοντέλο απέδωσε χειρότερα σε αυτό το σενάριο.



Σχηματικό 5: Grad-CAM heatmaps για ένα FN του ResNet50 (faces→frames).

Συμπεράσματα

Στην εργασία αυτή δοκιμάστηκαν τρία μοντέλα—ResNet50, Xception και Swin Transformer—για την ανίχνευση Deep Fake βίντεο, βασισμένα σε επιλεγμένα καρέ. Κάθε μοντέλο αξιολογήθηκε σε τέσσερα σενάρια ανάλογα με το αν χρησιμοποίησε crop προσώπου ή ολόκληρα frames στην εκπαίδευση και τη δοκιμή.

Τα βασικά ευρήματα είναι τα εξής. Κατάλληλη επιλογή input (faces vs. frames) επηρεάζει σημαντικά την απόδοση: όταν εκπαίδευσμε και δοκιμάζουμε σε cropped πρόσωπα, τα μοντέλα πετυχαίνουν F1 γύρω στο 0.86–0.89, με κορυφαίο το Xception. Αντίθετα, όταν απαιτείται γενίκευση από cropped σε full frame inputs, η επίδοση πέφτει δραματικά (F1 0.5). Η εκπαίδευση σε full frame (frames frames) δίνει πιο σταθερά αποτελέσματα (F1 0.62), καθώς τα augmentations καλύπτουν μεγαλύτερη ποικιλία στιγμιότυπων. Ιδιαίτερο ενδιαφέρον έχει το Swin Transformer, που μπορεί να «μεταφέρει» μέρος της γνώσης από full frame inputs σε cropped πρόσωπα (F1 0.71), χάρη στον μηχανισμό self attention.

Η κύρια δυσκολία που αντιμετωπίσαμε ήταν ότι η ταξινόμηση frame by frame δεν εκμε-

ταλλεύεται την πληροφορία κίνησης. Αυτό οδηγεί σε:

- προβλήματα όπου χαρακτηριστικές αλλαγές εμφανίζονται μόνο σε συνεχόμενα καρέ,
- προβλήματα γενίκευσης σε διαφορετικές ταχύτητες καρέ ή μοτίβα κίνησης,
- μεγάλη καθυστέρηση λόγω ανίχνευσης προσώπου σε κάθε frame.

Για τη βελτίωση προτείνουμε μελλοντικά:

- χρήση μοντέλων που επεξεργάζονται ακολουθίες (3D CNN, LSTM, Video Transformer),
- συνδυασμό εικόνας και ήχου για ανίχνευση lip sync artefacts,
- πιο ευέλικτα σχήματα συνάθροισης προβλέψεων (π.χ. attention based pooling).

Συνολικά, η έρευνα τονίζει την ανάγκη για temporal πληροφορία και multimodal χαρακτηριστικά, ώστε οι ανιχνευτές Deep Fake να αποδίδουν αξιόπιστα σε πραγματικά βίντεο.

Πηγές Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε είναι διαθέσιμο στη διεύθυνση: FaceForensics++