ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Δ.Π.Μ.Σ. ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

ΔΗΜΟΚΡΙΤΟΣ
DEMOKRITOS

ΙΝΣΤΙΤΟΥΤΟ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Προτεινόμενα Θέματα Διπλωματικής Εργασίας

Από

Θεόδωρο Γιαννακόπουλο

A. Adversarial Attacks and Robustness in Deep Neural Networks for Sound Event Detection

Abstract:

Η ραγδαία ανάπτυξη και η αποτελεσματικότητα της Βαθιάς Μάθησης έχουν οδηγήσει πολλές εφαρμογές να βασίζονται στα νευρωνικά δίκτυα για τις βασικές τους λειτουργίες. Παρ' όλα αυτά, έχει αποδειχθεί ότι αυτά τα δίκτυα είναι ευάλωτα σε επιθέσεις που αποσκοπούν στο να τα αναγκάσουν να κάνουν λάθος προβλέψεις. Σε αυτή τη διπλωματική εργασία, εστιάζουμε σε νευρωνικά δίκτυα που έχουν εκπαιδευτεί για την αναγνώριση ηχητικών γεγονότων. Πιο συγκεκριμένα, εκπαιδεύουμε ένα συνελικτικό δίκτυο στο AudioSet, ένα ευρείας κλίμακας dataset που αποτελείται από περισσότερα από 500 ηχητικά γεγονότα. Στη συνέχεια, εφαρμόζουμε κλασικές επιθέσεις που αναδεικνύουν τις αδυναμίες αυτών των μοντέλων. Ένα εντυπωσιακό αποτέλεσμα αυτών των επιθέσεων είναι ότι μπορούν να αναγκάσουν το μοντέλο (με μεγάλη πιθανότητα) να προβλέπει ότι δεν υπάρχει κάποιο ηχητικό γεγονός, ενώ παράλληλα στο ανθρώπινο αυτί η παρουσία του γεγονότος είναι εμφανής. Αυτή η αδυναμία είναι απαγορευτική για εφαρμογές που βασίζονται σε τέτοιου είδους μοντέλα για τις λειτουργίες τους. Έτσι, στα πλαίσια της εργασίας, εξετάζουμε διάφορες μεθόδους με τις οποίες μπορούμε να βελτιώσουμε την ανοχή των νευρωνικών δικτύων. Πιο συγκεκριμένα, εξετάζουμε την επίδραση του adversarial training σε αυτές τις επιθέσεις. Τέλος, πειραματιζόμαστε με μεθόδους ομαδοποίησης (clustering) στον χώρο αναπαραστάσεων (representation space) του μοντέλου, με στόχο να διαχωρίσουμε τα κανονικά δείγματα από τα αντίστοιχα adversarial δείγματα.

B. Automatic music captioning

Abstract:

In recent years a lot of research has been done on the task of automatic audio captioning, where given a non-speech audio sample a (deep) learning model returns a brief caption in free text form that describes the acoustic scene. Music captioning comprises a domain-specific application of audio captioning, where the machine learning mode now has to describe a given music sample. Such models can be helpful and important for people with accessibility issues, as well as for automatically extracting information about music samples.

C. Machine Unlearning

Abstract:
Machine unlearning allows a trained model to selectively remove some unwanted samples (“forget set”) while minimizing any adverse effects on the performance of the remaining data (“retain set”) and without retraining the model from scratch [1]. Current methodologies in machine unlearning have primarily focused on models trained through supervised learning. However, the applicability of these methodologies to self-supervised models, which power many of today's practical applications, remains questionable. The scope of this thesis is to investigate the feasibility and effectiveness of applying machine unlearning techniques to self-supervised models.

[1] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. 2023. Machine Unlearning: A Survey. ACM Comput. Surv. 56, 1, Article 9 (January 2024), 36 pages. https://doi.org/10.1145/3603620

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΙΝΣΤΙΤΟΥΤΟ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ