



## MSENSIS ML Engineering Task Report

Καλπαζίδης Αλέξανδρος

Το παρόν κείμενο αποτελεί ένα σύντομο report για το **MSENSIS MLE task** που μου ανατέθηκε. Σε αυτό περιλαμβάνονται σχόλια, παρατηρήσεις καθώς και τεχνικές λεπτομέρειες σχετικά με την υλοποίηση του task.

### Περιγραφή υλοποίησης

Για την επίτευξη των τεσσάρων βασικών στόχων της άσκησης αναπτύχθηκε ένα Web app το οποίο είναι υπεύθυνο για το classification εικόνων με τη χρήση δύο μοντέλων Vision Transformer αρχιτεκτονικής.

### Αρχιτεκτονική συστήματος

#### 1. Web UI (Streamlit)

Δημιουργήθηκε διαδραστικό UI που επιτρέπει στον χρήστη να επιλέξει το μοντέλο που θέλει να χρησιμοποιήσει μέσω radiobuttons (επιλέγει μεταξύ των δύο μοντέλων που περιγράφονται στη συνέχεια). Ο χρήστης έχει μια διαισθητική εμπειρία καθώς μπορεί να δει και την οπτικοποίηση της αρχιτεκτονικής του μοντέλου που επέλεξε. Επίσης μπορεί να κάνει drag and drop ή να επιλέξει από τα αρχεία του/της την εικόνα που θα δώσει ως είσοδο στο νευρωνικό δίκτυο καθώς και να δει τα αποτελέσματα του νευρωνικού δικτύου μαζί με τα αντίστοιχα confidence scores.

#### 2. Pre-trained ViT model

Η επιλογή του προ εκπαίδευμένου νευρωνικού δικτύου που θα χρησιμοποιηθεί

έγινε με κριτήριο το accuracy του. Μετά από μία έρευνα στο HuggingFace (όπως απαιτούσε η εκφώνηση) το μοντέλο που επιλέχθηκε ήταν <https://huggingface.co/nateraw/vit-base-cats-vs-dogs> το οποίο είχε το υψηλότερο accuracy μεταξύ των Vision Transformers (Overall accuracy: 99.35%)

### 3. Custom-trained MobileViT v3 model

Εφόσον επιλέχθηκε για πρώτο μοντέλο, το μοντέλο με το υψηλότερο accuracy κρίνεται σκόπιμο το δεύτερο μοντέλο να επιλεχθεί βάσει του performance του. Για το λόγο αυτό επιλέχθηκε η MobileViT αρχιτεκτονική. Το συγκεκριμένο μοντέλο παρότι ελαφρύ (5.8 εκατομμύρια παράμετροι) παραμένει αρκετά accurate δεδομένου καλού dataset και αρκετής εκπαίδευσης. Να σημειωθεί ότι λόγω περιορισμένου χρόνου το μοντέλο σταμάτησε να εκπαιδεύεται στις 300 εποχές και θα περιμέναμε καλύτερη απόδοση αν είχε εμπλουτιστεί το training dataset και είχε εκπαιδευτεί για παραπάνω εποχές. Το αποτέλεσμα είναι ένα μοντέλο που έχει Overall Accuracy: 90.86% ενώ ανά κλάση έχει Cat Accuracy: 87.57% και Dog Accuracy: 94.15%. Ένας επιπρόσθετος λόγος για την επιλογή της συγκεκριμένης αρχιτεκτονικής είναι ότι το μοντέλο είναι αρκετά ελαφρύ για να τρέξει on device ακόμα και σε συσκευές περιορισμένου hardware όπως είναι ένα mid range smartphone μεγιστοποιώντας τη δυνατότητα αξιοποίησης του σε μελλοντικά πρότζεκτ.

### Επεξεργασία dataset

Το μοντέλο MobileViT εκπαιδεύτηκε στο dataset που δόθηκε μέσα από την εκφώνηση της άσκησης. Το dataset χωρίστηκε 80/10/10 (80% training set, 10% validation set, 10% test set) δίνοντας έμφαση καθ' ένα από αυτά τα υποσύνολα να είναι balanced. Από το CSV αρχείο αφαιρέθηκαν τα labels που δεν αντιστοιχίζονταν σε εικόνες ή ήταν null. Από το dataset αφαιρέθηκαν δύο εικόνες που ήταν corrupted. Τέλος όλες οι εικόνες μετατράπηκαν σε RGB.

### Οδηγίες εγκατάστασης

Αναλυτικές οδηγίες για την εγκατάσταση και τη χρήση του Web app υπάρχουν στο README αρχείο που περιέχεται στο πρότζεκτ.