Σχολη Επιστημης Υπολογιστων-Πανεπιστημιο Κρητης

CS543-Software Systems and Technologies for Big Data Applications

Εαρινο Εξαμηνο 2020-2021

Programming Assignment 3

Φοιτητής
Καλογερακης Στεφανος
ΑΜ:1205
Email:skalogerakis97@gmail.com

Διδάσκων
Χ. Κοζανιτης

Εισαγωγή

Στα πλαίσια του τρίτου και τελευταίου assignment του μαθήματος CS543, σκοπός ήταν το χτίσιμο μιας εφαρμογής σε πλατφόρμα της επιλογής μας για ένα θέμα της επιλογή μας. Η βασικότερη ιδέα ήταν να μας δοθεί η ελευθερία να ανακαλύψουμε κάποια πλατφόρμα/API που είναι εντελώς καινούργια για τον καθένα ξεχωριστά και ανακαλύψουμε καινούργιες δυνατότητες, εναλλακτικές αλλά και περιορισμούς που παρέχει η εκάστοτε τεχνολογία.

Questions

1. What did you build?

Για την συγκεκριμένη εργασία, επέλεξα να δουλέψω σε ένα θέμα σχετικό με machine learning και ένα text classification πρόβλημα. Η συγκεκριμένη ιδέα για άσκηση προέρχεται από το μάθημα ΗΥ563, το οποίο σχετίζεται με information retrieval και NLP όπου είχα δει διάφορα πράγματα σε θεωρητικό πλαίσιο για κάποιες μεθοδολογίες και αρχιτεκτονικές που θα ήθελα να δω και στην πράξη πως υλοποιούνται. Στο πρακτικό κομμάτι της υλοποίησης πραγματοποιήθηκε χρήση του **TensorFlow και του Keras API**. Τελικός σκοπός της υλοποίησης ήταν *η ταξινόμηση του κατα πόσο μια κριτική ήταν καλή ή κακή από ένα Dataset με κριτικές ταινιών(IMDB).*

2. What technology did you choose?

Όπως επισημάνθηκε και παραπάνω η τεχνολογία που επιλεχθηκε είναι το TensorFlow και το Keras API. Το TensorFlow αποτελεί μια ευρέως χρησιμοποιούμενη επιλογή σε προβλήματα machine/deep learning οπότε η εξοικείωση με το συγκεκριμένο εργαλείο είναι αρκετά σημαντική αλλά και ενδιαφέρουσα.

3. How can we install the framework of your choice to run your code?

Η έκδοση του TensorFlow που χρησιμοποιήθηκε κατά τις δοκιμες ήταν το TensorFlow 2. Σύμφωνα με το Documentation οι οδηγίες είναι οι εξής:

- Python 3.6–3.8
- Requires the latest pip
  - pip install –upgrade pip
- Current stable release for CPU and GPU
  - pip install tensorflow
- Install matplotlib(for plots)
  - python3 -m pip install -U matplotlib

4. What was hard about building your application?

Σε προσωπικό στάδιο, η συγκεκριμένη εφαρμογή ήταν challenging σε αρκετά επίπεδα. Ξεκινώντας, πέρα από την ενασχόληση με το συγκεκριμένο μάθημα οι γνώσεις μου σχετικά με το machine learning ήταν αρκετά περιορισμένες. Το θεώρησα όμως σαν μια ακόμα μια καλή ευκαιρία να συναντήσω και να ασχοληθώ με κάποιο αντίστοιχο πρόβλημα. Εν συνεχεία, δεν είχα κανενός είδους εξοικείωση με την συγκεκριμένη τεχνολογία(όπως

ζητούσε και η εκφώνηση) και ήταν αρκετά δύσκολο να βρω ένα σημείο εκκίνησης. Λόγω και της φύσης των προβλημάτων που ασχολείται η συγκεκριμένη τεχνολογία, η πληροφορία και οι δυνατότητες είναι χαοτικές και έπρεπε να αφιερώσω συνολικό χρόνο αρκετά παραπάνω από τις 3-4 ώρες που ζητούσε η εκφώνηση σε πολλές περιπτώσεις απλά διαβάζοντας documentation και demos από αρκετές πηγές. Τελικός περιορισμός που όμως δεν ήταν εν τέλει αποτρεπτικός παράγοντας ήταν η ίδια η γλώσσα python, που δεν έχω ασχοληθεί σε μεγάλο βαθμό στο παρελθόν, αλλά η εξοικείωση μου με τις υπόλοιπες γλώσσες προγραμματισμού έκανε την συγκεκριμένη μετάβαση σχετικά ομαλή.

Να επισημάνω βέβαια ότι τόσο το dataset που επιλέχθηκε όσο και τα αρχικά στάδια του preprocessing ήταν λίγο πολύ τετριμμένα(Dataset έτοιμο απο το keras καθώς και labeled πληροφορία) μόλις υπήρχε κατανόηση για την δομή του Dataset. Ο λόγος ήταν ότι ήθελα να επικεντρωθώ στις διαφορετικές επιλογές από μοντέλα που δύναται να αξιοποιήσει κάποιος με την συγκεκριμένη τεχνολογία.

5. What did you think was intuitive about the API/SDK that you used?

Παρά τις τόσες πολλές διαφορετικές εναλλακτικές και την χαοτική φύση των προβλημάτων που μπορεί κάποιος να ασχοληθεί με το TensorFlow το API έχει εξαιρετική δομή. Αυτό έχει σαν αποτέλεσμα ότι η δημιουργία ενός pipeline σε μια απλή μορφή του είναι σχετικά straightforward ενώ και τα αρκετά πιο σύνθετα προβλήματα αποτελούν εξέλιξη της απλής αυτή μορφής.

6. Why do you think the developers made the API/SDK the way they did?

Θεωρώ ότι το σκεπτικό των developers της συγκεκριμένη τεχνολογίας ήταν η δημιουργία της πιο απλής δυνατής σχεδίασης. Έτσι η πολυπλοκότητα μένει στην επίλυση του προβλήματος με τον χρήστη όμως να έχει μια πληθώρα από διαφορετικές επιλογές προκειμένου να μπορεί παράγει όσο σύνθετη υλοποίηση επιθυμεί. Αξίζει βέβαια να σημειωθεί ότι η συγκεκριμένη τεχνολογία έχει επέλθει απο πολλά στάδια εξέλιξης προκειμένου να φτάσει σε αυτό το στάδιο και εξελίσσεται συνεχώς.

7. What changes would you make to the API/SDK to improve it?

Ίσως το μεγαλύτερο πρόβλημα που συνάντησα κατά την υλοποίηση μου ήταν σχετικά με τις εκδόσεις. Σε αρκετές περιπτώσεις συναντούσα για ίδιες συναρτήσεις διαφορετικά διαθέσιμα αλλά και deprecated πεδία, το οποίο φαντάζομαι σε ένα βαθμό ευθύνεται και η έκδοση της python. Το ίδιο όμως ισχύει και για το documentation όπου σε κάποια σημεία αναμειγνύονται καινούργιες και παλιές εκδόσεις. Σίγουρα λοιπόν θα βελτιωνα την διάρθρωση και την οργάνωση ξεκινώντας απο το documentation για τις διαθέσιμες εκδόσεις.

8. What did you learn in the process of building your application?

Προκειμένου να φτάσω στην τελική υλοποίηση χρειάστηκε να περάσω από αρκετά διαφορετικά στάδια υλοποίησης και δοκιμών. Ξεκινώντας από τα βασικά έπρεπε να εξοικειωθώ με τις έννοιες των μοντέλων που υπάρχουν στο TensorFlow καθώς και την διαχείρισή τους(αρχικοποίηση μοντέλου, προσθήκη Layer, Compile, Fit, Evaluate). Σε αρχικό στάδιο έπρεπε να μάθω τι ακριβώς σημαίνουν και τι γίνεται σε όλα αυτά τα στάδια σε απλά σενάρια. Στο επόμενο στάδιο ασχολήθηκα με τις λεπτομέρειες σε πιο χαμηλό επίπεδο για κάθε επιμέρους Layer(optimizer, activation function, loss function, batch size, epochs). Σε τελικό στάδιο, πειραματίστηκα με τις παραμέτρους για όλα τα παραπάνω προσπαθώντας να ερμηνεύσω την συμπεριφορά των αποτελεσμάτων.

9. What was most surprising?

Η μεγαλύτερη έκπληξη από την ενασχόληση με την συγκεκριμένη τεχνολογία είναι σίγουρα η απλότητα που έχουν καταφέρει να παρέχουν στον χρήστη προκειμένου να δημιουργήσει μοντέλα που στην θεωρία έχουν πολύ μεγάλο βαθμό πολυπλοκότητας. Επίσης είναι σίγουρα εντυπωσιακός ο βαθμός της λεπτομέρειας και των τόσων διαφορετικών δυνατοτήτων που δίνονται από την συγκεκριμένη τεχνολογία.

Resources

https://realpython.com/python-keras-text-classification/
https://www.tensorflow.org/tutorials/keras/text_c250,021250,021lassification
https://www.youtube.com/watch?v=6g4O5UOH304&t=5107s&ab_channel=freeCodeCamp.org