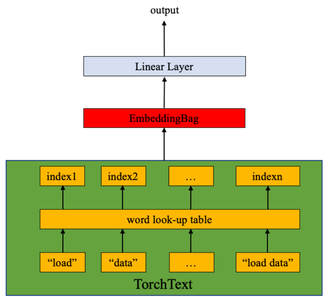
AUTHORSHIP VERIFICATION

Introduction

Authorship verification consists of comparing 2 or more text documents for their style to determine if the same person wrote the documents. It is typically performed by experts who rely on the analysis of spelling/grammar, stylistic mannerisms, dialects, sociolects, and registers of language that hint at the authorship of a disputed document. Developing a performing network would help in many criminal cases and detect deceptive intent and fake news in e-commerce and social media.

Method

In order to convert the stories to tensors of numerical values, I built a complex vocab covering them using spacy and torchtext.vocab. Then I standardized the size of texts and mapped them to an array of values based on the vocab, similarly to the bag of words technique.
For the neural network I went for a simple implementation, 3 fully connected layers, 4096 to 128, 128 to 64, and then concatenate and return the output. One way of improvement I think would be implementing LSTM layers. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. This is a behavior required in complex problem domains like machine translation, speech recognition, and more.



Experiments

At first glance, I opted for the bag of words method and then word2vec, but I encounter many problems regarding embeddings and training speed. As so, the final form was built with torch.vocab. Another difficult problem to solve was managing a dual input neural network, which implied a bit of data manipulation and creating layers for each one.
For parameter optimization, I chose Adam for the loss function Cross-Entropy Loss. Adam optimizer involves a combination of two gradient descent methodologies: Momentum: This algorithm is used to accelerate the gradient descent algorithm by taking into consideration the 'exponentially weighted average' of the gradients. Using averages makes the algorithm converge towards the minima in a faster pace. Cross entropy loss is a metric used to measure how well a classification model in machine learning performs. The loss (or error) is measured as a number between 0 and 1, with 0 being a perfect model. The goal is generally to get your model as close to 0 as possible.
Each pair of text is fed to the network and goes through 2 linear layers, then the result is concatenated and classified, and sent as output. The best accuracy obtained was 55

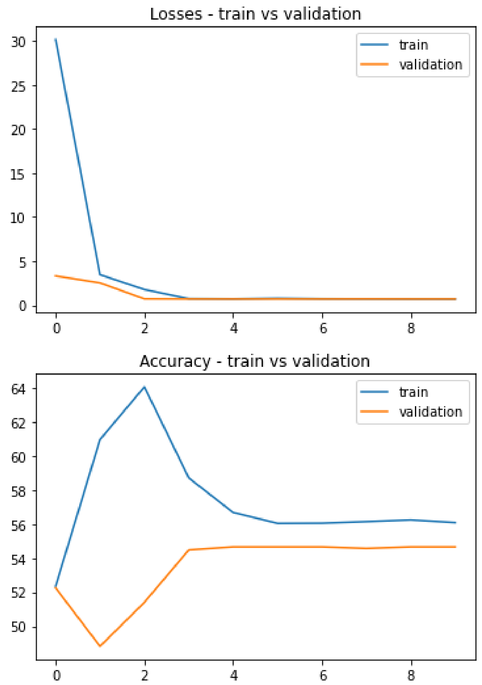


Fig. 2: Big fancy graphic.

Dataset

The dataset consists of fandom stories about the Guardians of Ga'Hoole, a fantasy series of 16 books centered in a universe ruled by owls. They are pairs of 2 stories, together with the author's ids, followed by another set of labels connected to the pairs.

Conclusions

Most of the problems encountered were technical, getting thousands of stories in the dual input network wasn't as easy as expected. For future work, I am thinking of constructing multiple NN each with a focus: on grammar, vocab, punctuation, and emotion. After combining all of them, I think I can get much better results.