The Case for Multi-Task Learning: Using Natural Language Inference to enhance performance on Fake News Detection

Introduction

This academic poster concerns itself with our research into to the case for Multi-Task Learning (MTL) in the context of Fake News Detection (FND) and Natural Language Inference (NLI), Specifically, the research question under investigation is whether training and evaluating a deep neural network (DNN) model on NLI as an *auxiliary* task enhances performance on FND as the *main* task.

Model

The model we designed and implemented can be reviewed in Fig. 1, and draws inspiration from the Hierarchical Attention Network (HAN) as introduced by Yang et al. (2016), as well as from the Multi-Task Deep Neural Network (MT-DNN) as introduced by Liu et al. (2019).

As such, our model boasts deep, contextual word embeddings and an attentive sentence encoder, which are shared for both tasks. The sentence embeddings produced by this encoder are used as input for either the NLI classifier for our *auxiliary* task, or the attentative document encoder used in the FND classifier.

Both the attentative sentence and document encoder use a bidirectional LSTM (Hochreiter & Schmidhuber, 1997), combined with an attention mechanism at the word- and sentence level, respectively, to produce their embeddings.

Platform	Fake / real	# of articles
Gossipcop	Fake	5,323
	Real	17,817
PolitiFact	Fake	432
	Real	624
	Total	**23,196**

Table 1. FNN dataset statistics

Experiments

Datasets

For the FND task, we used the *FakeNewsNet* dataset as published by Shu et al. (2018), some statistics of which can be reviewed in Tab. 1. This dataset did not come with the articles themselves, but instead with the article titles and URLs (as well as tweets, which we did not use), as well as software to scrape them.

After scraping, the data collected was quite noisy. The scraper ran into some paywalls and privacy notices, for example. Another issue with the data, was that

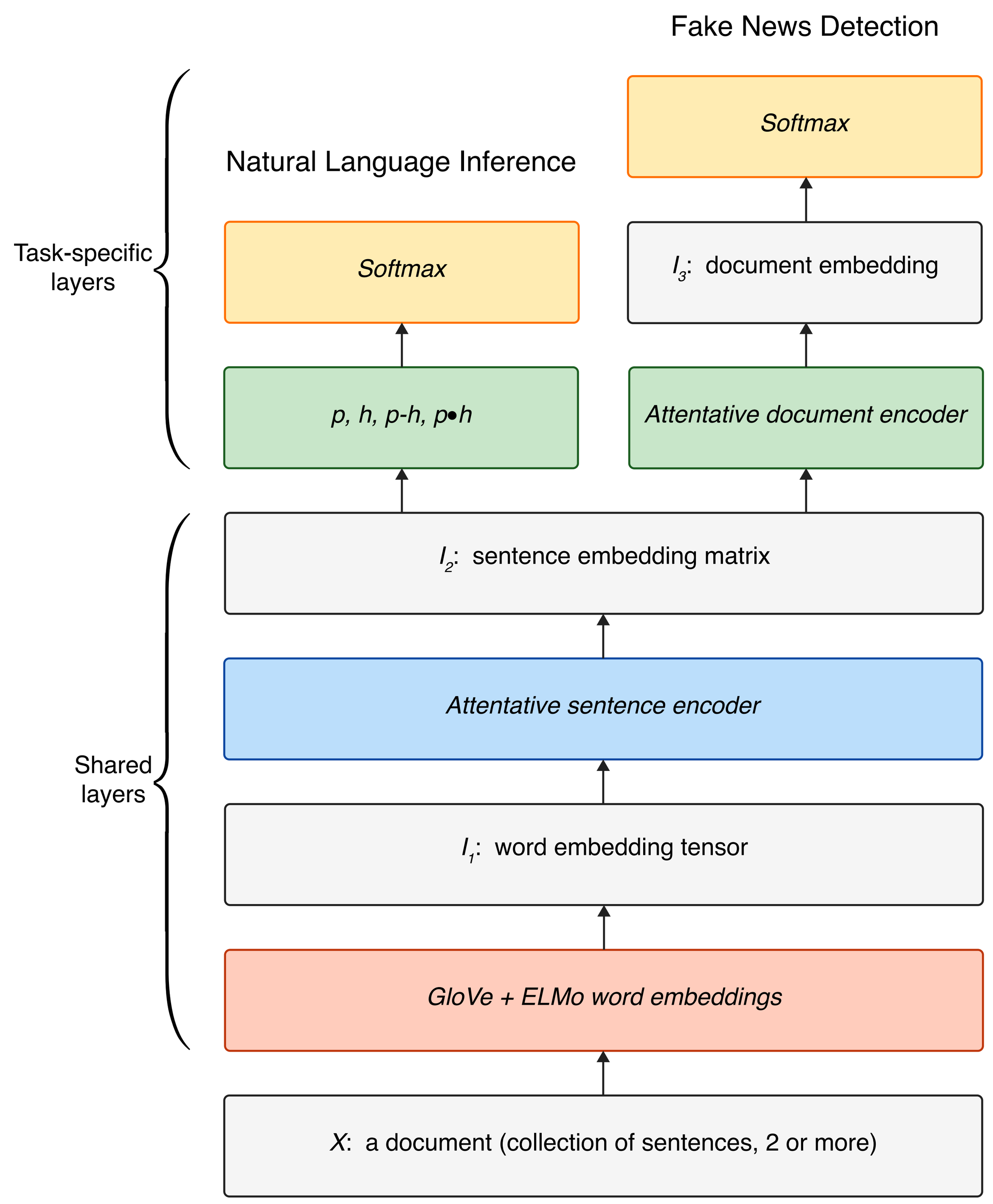

Figure 1. Our model architecture

some documents were extremely long (in terms of the number of sentences), as can be reviewed in Fig. 2. As such, we performed data cleaning to remove any such articles that either ran into paywalls / privacy notices or were too long (> 40 sentences), leaving 20,467 articles with 375k sentences in total.

For the NLI task, we used the *Stanford NLI* (SNLI) corpus, as introduced by Bowman et al. (2015), which contains 570k sentence-pairs with entailment labels.

Training procedure

In order to investigate our research question, we trained our model as introduced and described previously twice: Once with our Single-Task Learning (STL) objective, where we just trained and evaluated on our FND task, and once with our MTL objective, where we trained and evaluated our model on both the FND task, as well as the NLI task. In both cases, we trained our models for 10 epochs.

For the MTL objective, in any given epoch, the task to be trained was sampled uniformly at random, with a probability proportional to the dataset size and the batch size, to ensure a balanced training process. Furthermore, to maintain focus on our *main* FND task and achieve higher performance, we incorporated two weighting schemes for combining the losses. Following Liu et al. (2018), we introduce dynamic weighted averages to weight the losses for the FND and NLI task. Next to that, we add a weight that balances the dataset sizes, namely the complement of the probability of sampling a task k.

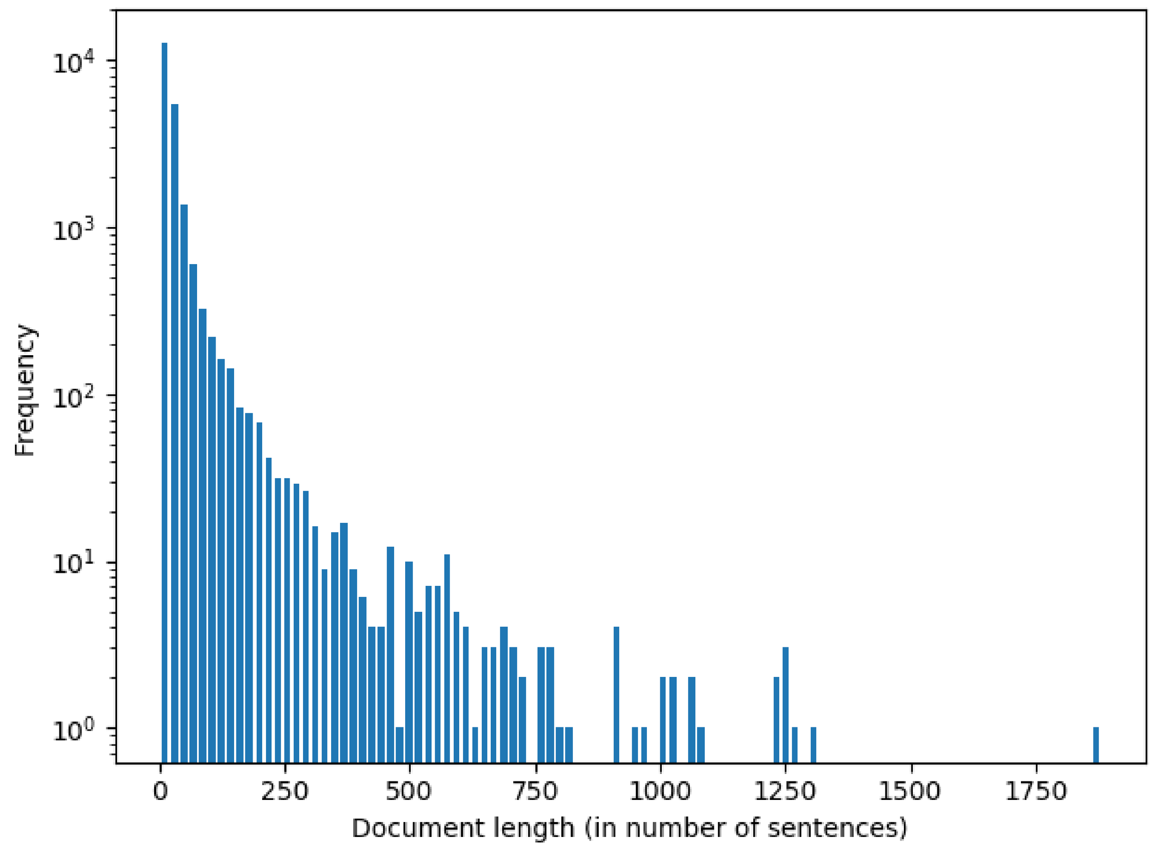

Figure 2. Plot of document length frequencies in FNN

Baselines

The first and most obvious baseline for our MTL learning objective, is that of the STL objective, as the only difference between the two is the addition of the *auxiliary* NLI task during training, thus allowing for direct comparisons between STL and MTL for FND.

A second baseline comes from existing research. Shu et al. (2019) combine linguistic and structural models, exploiting both the textual and semantic information from the articles and the user interaction data from the tweets corpus. Their best performing model that is trained solely on news content is LIWC, a simple bag-of-words approach that categorizes lexicon into psycholinguistic categories to learn a feature representation. Their best performing model overall is LIWC_HPNF, that extends the LIWC model by adding features of the hierarchical propagation network built from Twitter interaction data.

Results

After training our model for the STL objective for 7 epochs, the accuracies obtained for both the train and validation set can be reviewed in Fig. 3.

Unfortunately, due to time constraints, we have not yet been able to train our model for the MTL objective for 10 (or any) epochs. As such, we cannot present any results for this objective yet.

Analysis

As can be inferred from Fig 3., with the exception of a single epoch, our STL objective seems to be learning from the data, as the validation accuracy increases. Moreover, our STL objective performs better than the majority baseline (0.767), which reinforces our conclusion about model learning.

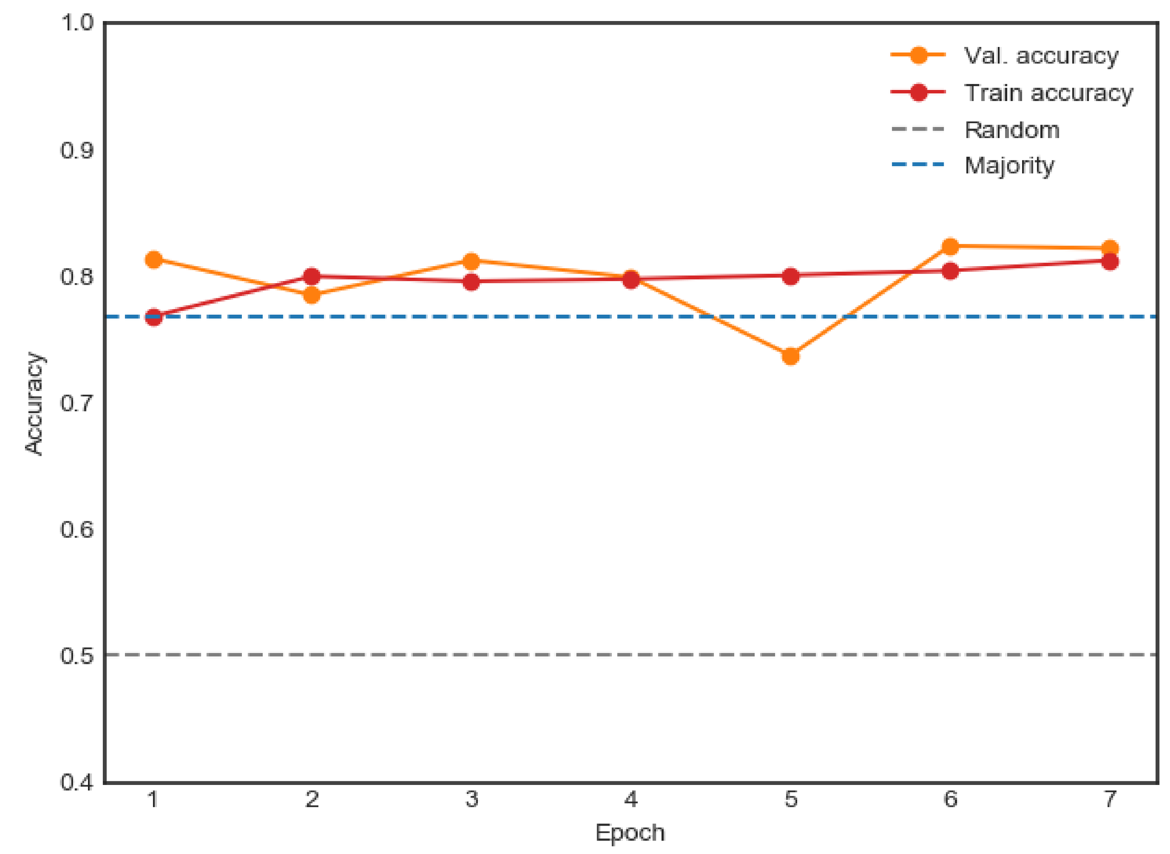

Figure 3. Plot of the accuracies for our STL objective

References

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Liu, S., Johns, E., & Davison, A. J. (2018). End-to-end multi-task learning with attention.

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-Task Deep Neural Networks for Natural Language Understanding.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media.

Shu, K., Mahudeswaran, D., Wang, S., & Liu, H. (2019). Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1480-1489).