Fake News Detection by Learning Convolution Filters through Contextualized Attention

Ekagra Ranjan
Indian Institute Of Technology Guwahati
ekagra.ranjan@iitg.ac.in

Methodology

Overview

Instead of directly extracting features from Statement, we employ an attention mechanism to use the given side information (subject, speaker, job, state, party, context and justification) to attend over the given statement to check its truthfulness. The attention mechanism makes the process of feature extraction from statement contextualized based on side information. See Fig. 1 for the graphical representation of the architecture.

Formation of Context Query

First the words are passed through a learnable embedding matrix (randomly initialized initially) to convert the input sparse representation of words to meaningful dense representations $w_i \in \mathbb{R}^d$. Subject, Job, Context and Justification are passed through their respective 2 layer BiLSTM model for modeling the dependency between the words in them. Speaker, State and Party are not passed though LSTM as they are extremely short sentence and there is not much sequence information. These hidden features from side information are concatenated and passed through a fully connected (fc) layer to give us a *context query* vector q. This query vector contains the summarised side information which will be used for attention mechanism explained the following sections.

Word Attention

The query vector $q \in \mathbb{R}^{1\times d}$ is used to attend over all the words $w_i \in \mathbb{R}^{1\times d}$ in the statement to produce a score $\alpha_i \in \mathbb{R}$ which models the relevance of the word w_i for finding the truthfulness of the statement. α_i is computed as :

$$\alpha_i = \frac{\langle fc(q), w_i \rangle}{d} \tag{1}$$

where fc is a fully connected layer. Each word embedding w_i is then updated as:

$$e_i = \alpha * w_i \tag{2}$$

A Convolution Neural Network (CNN) is applied over this new word representation matrix $E \in \mathbb{R}^{n\times d}$ where n is the number of words in the statement.

Contextual Attention

Given the kernel size k and number of filter as F, the weights $W \in \mathbb{R}^{F\times k\times d}$ of the filters in the CNN is *contextualized* using the query q_i as:

$$W^c = fc(q) * W \tag{3}$$

where $*$ is elementwise broadcastable multiplication operator. This formulation of kernel weights allows the side information to govern the pattern to be searched for from the statements (Shi, Rao, and Lin 2018). The CNN with weights W^c extracts features $f^c \in \mathbb{R}^{F\times n\times 1}$ from E. f^c is maxpooled to give the final representation $f^c_m \in \mathbb{R}^F$ of the statement. f^c_m and q are concatenated followed by an fc layer to give the output class of the statement.

Experimentation

Baseline

To find out the usefulness of the attention mechanism, a similar hybrid model of CNN and LSTM but without attention mechanism is compared to Fake-Net.

Setup

We train all models for 10 epochs with no early stopping or learning rate scheduling. The learning rate was chosen 0.001. Dropouts value was searched over $0.2, 0.5, 0.7$. The embedding dimention d was set to 100.

Results

From 1 we can see that on 6-way classification our model has an improvement of 2% wheras on binary classification this gain increases to 5%. This shows that contextual attention helps in finding useful features from relevant parts of the statement.

Areas of Improvement

(Alhindi, Petridis, and Muresan 2018) and (Wang 2017) use Glove embeddings as starting point for their embedding matrix. In current the embedding was randomly initialised initially. Using Glove or word2vec embeddings will surely help in increasing the performance. (Alhindi, Petridis, and Muresan 2018) also uses SentiStrength for capturing sentiment

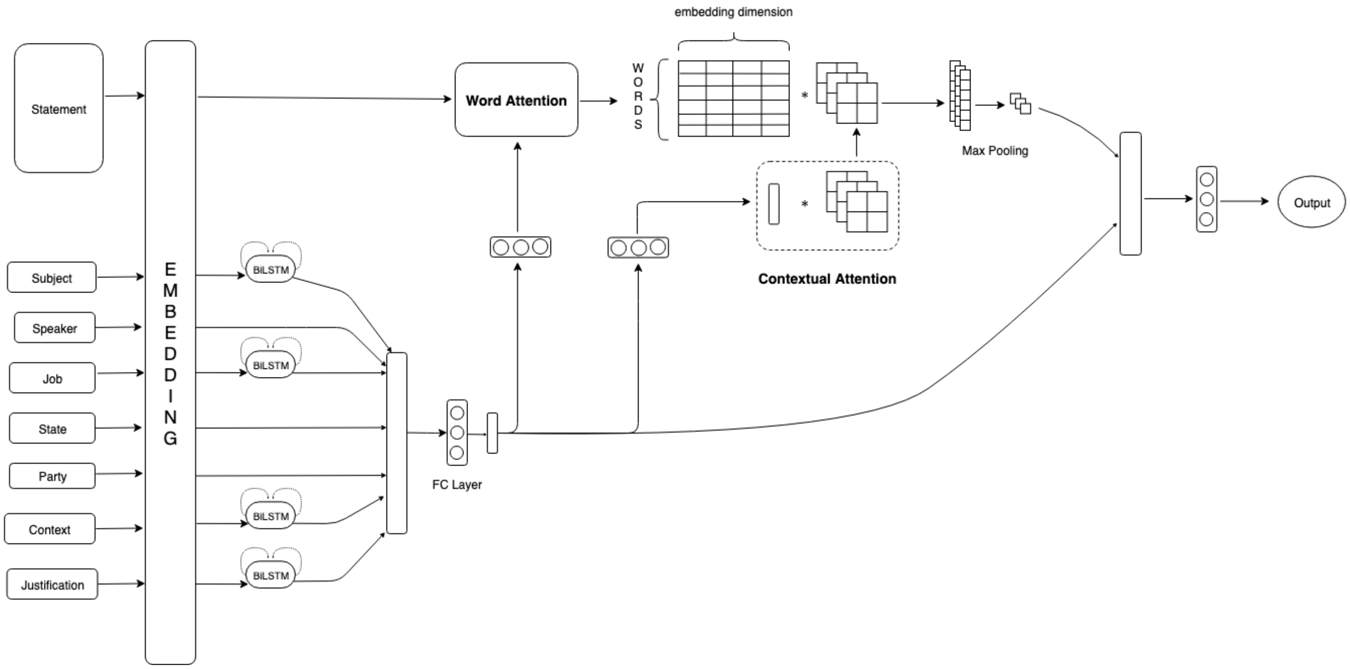

Figure 1: Fake-Net Architecture Overview

Method	2-way classification	6-way classification
Baseline	0.578	0.231
Fake-Net	**0.633**	**0.249**

Table 1: Ablation study for context attention from side information

and EmoLex for emotions as additional features. Incorporating these will also boost the performacne of Fake-Net. State of art techniques for Natural Language Processing like Transformers (Vaswani et al. 2017) will be useful for complex tasks like fake news detection given that the there is an increase in the number of training samples in future datasets. Considring the LIAR-PLUS dataset (Alhindi, Petridis, and Muresan 2018) which has only 12k samples, transfer learning by finetuning pretrained models like Bert (Devlin et al. 2018) and XLNet (Yang et al. 2019) will be very helpful.

Acknowledgement

I would like to thank the developer of Pytorch for their deep learning framework. I would also like to acknowledge the efforts of Fangjun Zhang (Zhang 2017) for open sourcing the code for LIAR dataset which served as my starting point over which I have built my approach.

References

Alhindi, T.; Petridis, S.; and Muresan, S. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 85–90.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shi, P.; Rao, J.; and Lin, J. 2018. Simple attention-based representation learning for ranking short social media posts. *arXiv preprint arXiv:1811.01013*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, W. Y. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zhang, F. 2017. fake-news-detection. In *https://github.com/zfjmike/fake-news-detection*.