ICANN19
28th International Conference on Artificial Neural Networks

Hierarchical Attentional Hybrid Neural Networks for Document Classification

Jader Abreu*, Luis Fred*, David Macêdo**, and Cleber Zanchettin**
Centro de Informática - Universidade Federal de Pernambuco - 50.740-560, Recife, PE, Brazil
{jaoa,lfgs,dlm,cz}@cin.ufpe.br

Centro de Informática UFPE

Universidade Federal de Pernambuco

Abstract

Document classification is a challenging task with important applications. The deep learning approaches to the problem have gained much attention recently. Despite the progress, the proposed models do not incorporate the knowledge of the document structure in the architecture efficiently and not take into account the contexting importance of words and sentences. In this paper, we propose a new approach based on a combination of convolutional neural networks, gated recurrent units, and attention mechanisms for document classification tasks. We use of convolution layers varying window sizes to extract more meaningful, generalizable and abstract features by the hierarchical representation. The proposed method in improves the results of the current attention-based approaches for document classification.

Keywords: Text classification · Attention mechanisms · Document classification · Convolutional Neural Networks.

Hierarchical Attentional Hybrid Neural Networks

In this paper, we propose a new approach for document classification based on CNN, GRU hidden units and attentional mechanisms to improve the model performance by selectively focusing the network on essential parts of the text sentences during the model training. Inspired by [1], we have used the hierarchical concept to better representation of document structure. Temporal convolutions [2], which give us more flexible receptive field sizes, was also used in some experiments . We call our model as Hierarchical Attentional Hybrid Neural Networks (HAHNN). We evaluate the proposed approach comparing its results with state-of-the-art models and the model shows an improved accuracy.

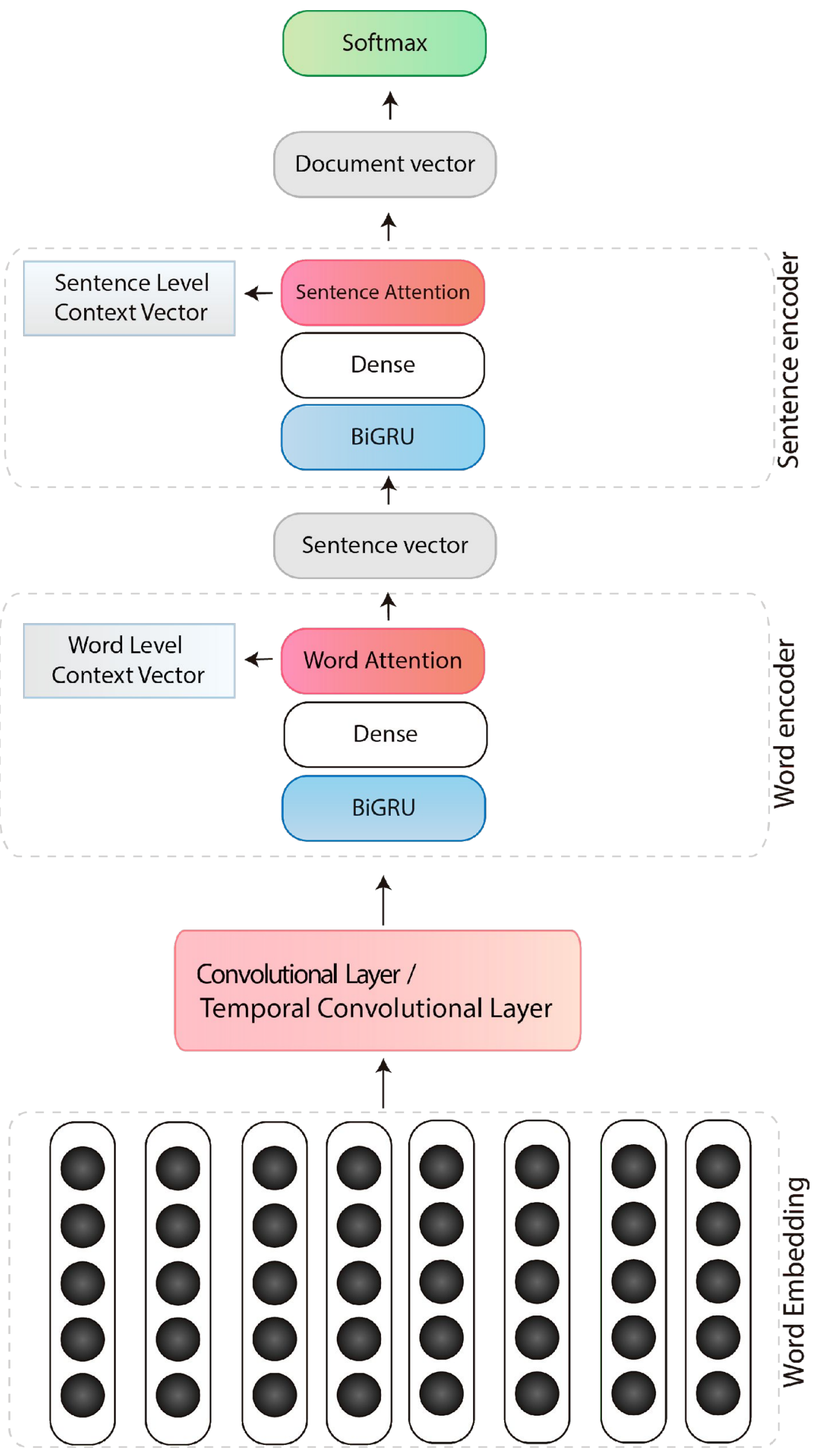


Fig. 1: Our HAHNN Architecture include an CNN layer after the embedding layer. In addition, we have created a variant which includes a temporal convolutional layer [2] after the embedding layer.

The HAHNN model combines convolutional layers, Gated Recurrent Units, and attention mechanisms. Figure 1 shows the proposed architecture. The first layer of HAHNN is a pre-processed word embedding layer (black circles in the Figure 1). The second layer contains a stack of CNN layers that consist of convolutional layers with multiple filters (varying window sizes) and feature maps. We also have performed some trials with temporal convolutional layers with dilated convolutions and gotten promising results. Besides, we used Dropout for regularization. In the next layers, we use a word encoder applying the attention mechanism on word level context vector. In sequence, a sentence encoder applying the attention on sentence-level context vector. The last layer uses a softmax function to generate the output probability distribution over the classes.

We use CNN to extract more meaningful, generalizable and abstract features by the hierarchical representation. Combining convolutional layers in different filter sizes with both word and sentence encoder in a hierarchical architecture let our model extract more rich features and improves generalization performance in document classification. We investigate two variants of the proposed architecture. There is a basic version, as described in Figure 1, and there is another which implements a TCN [2] layer. The goal is to simulate RNNs with very long memory size by adopting a combination of dilated and regular convolutions with residual connections. Dilated convolutions are considered beneficial in longer sequences as they enable an exponentially larger receptive field in convolutional layers

The proposed model takes into account that the different parts of a document have no similar relevant information. Moreover, determining the relevant sections involves modeling the interactions among the words, not just their isolated presence in the text. Therefore, to consider this aspect, the model includes two levels of attention mechanisms. One structure at the word level and other at the sentence level, which let the model pay more or less attention to individual words and sentences when constructing the document representation.

*Authors contributed equally and are both first authors.
**Corresponding authors
The proposed method is openly available in the github on repository:
https://github.com/luisfredgs/cnn-hierarchical-network-for-document-classification

Main components of HAHNN architecture

Our architecture consists of different parts: 1) A word sequence encoder and a word-level attention layer; and 2) A sentence encoder and a sentence-level attention layer. In the word encoder, the model uses bidirectional GRU to produce annotations of words by summarizing contextual information from both directions. The attention levels let the model pay more or less attention to individual words and sentences when constructing the representation of the document.

Experiments and Results

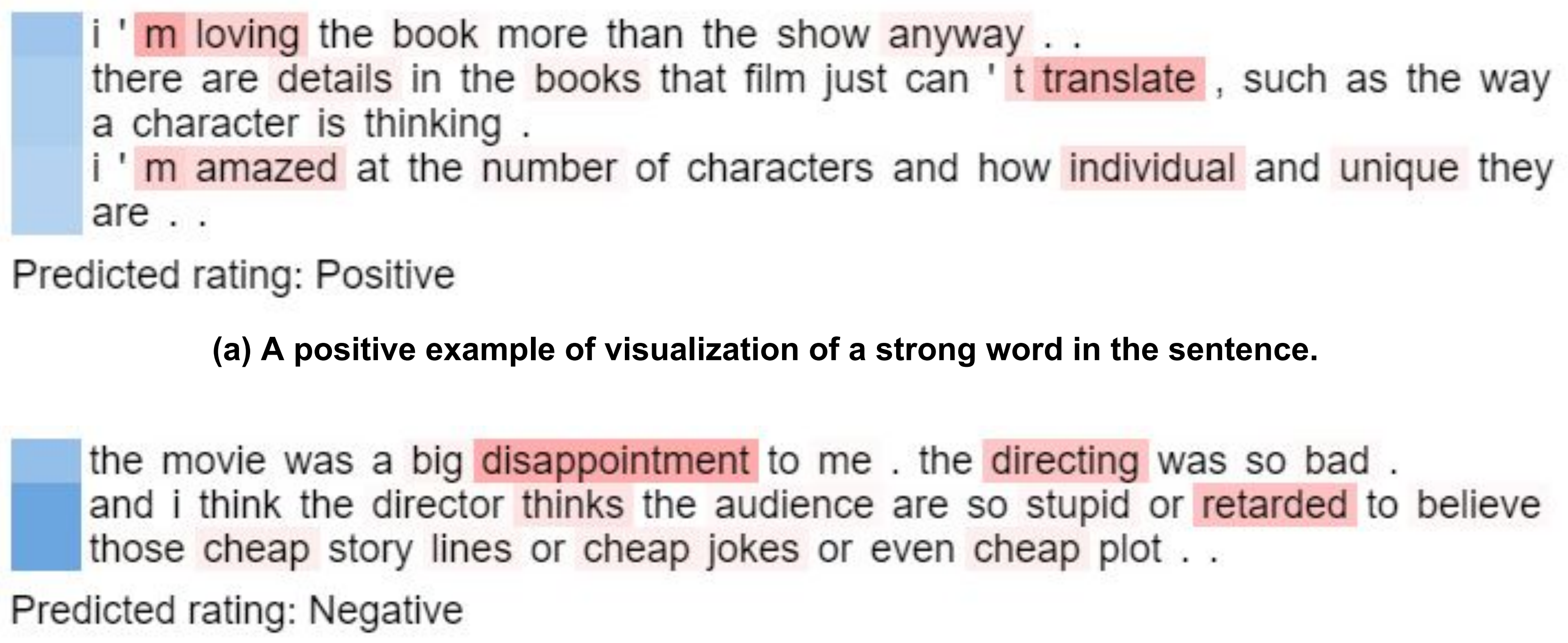
We evaluate our model on two document classification datasets using 90% of the data for training and the remaining 10% for tests. The word embeddings have dimension 200 and we use Adam optimizer with a learning rate of 0.001. The datasets used are the IMDb Movie Reviews and Yelp 2018. The former contains a set of 25k highly polar movie reviews for training and 25k for testing, whereas the classification involves detecting positive/negative reviews. The latter include users ratings and write reviews about stores and services on Yelp, being a dataset for multiclass classification (ratings from 0-5 stars). Yelp 2018 contains around 5M full review text data, but we fix in 500k the number of used samples for computational purposes.

Table 1: Classification accuracies.

Method	Accuracy on test set	
	Yelp 2018 (five classes)	IMDb (two classes)
VDNN [3]	62.14	79.47
HN-ATT [1]	72.73	89.02
CNN [4]	71.81	91.34
Our model with CNN	**73.28**	92.26
Our model with TCN	72.63	**95.17**

Attention Weights Visualizations

To validate the model performance in select informative words and sentences, we present the visualizations of attention weights in Figure 2. There is an example of the attention visualizations for a positive and negative class in test reviews. Every line is a sentence. Blue color denotes the sentence weight, and red denotes the word weight in determining the sentence meaning. There is a greater focus on more important features despite some exceptions. For example, the word "loving" and "amazed" in Figure 2 (a) and "disappointment" in Figure 2 (b).


i ' m loving the book more than the show anyway . .
there are details in the books that film just can ' t translate , such as the way a character is thinking .
i ' m amazed at the number of characters and how individual and unique they are . .
Predicted rating: Positive

(a) A positive example of visualization of a strong word in the sentence.

the movie was a big disappointment to me . the directing was so bad .
and i think the director thinks the audience are so stupid or retarded to believe those cheap story lines or cheap jokes or even cheap plot . .
Predicted rating: Negative

(b) A negative example of visualization of a strong word in the sentence.

Fig. 2: Visualization of attention weights computed by the proposed model

Occasionally, we have found issues in some sentences, where fewer important words are getting higher importance. For example, in Figure 2 (a) notes that the word "translate" has received high importance even though it represents a neutral word. These drawbacks will be taken into account in future works.

References

[1] YANG, Zichao et al. Hierarchical attention networks for document classification. In: Conf. North Am. Chapter of the Assoc. for Comp. Ling. 2016. p.1480-1489, San Diego, CA, USA. doi: 10.18653/v1/N16-1174
[2] BAI, Shaojie; KOLTER, J. Zico; KOLTUN, Vladlen. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271, 2018.
[3] Conneau, Alexis, et al. "Very deep convolutional networks for text classification." arXiv preprint arXiv:1606.01781 (2016). doi:10.18653/v1/E17-1104
[4] KIM, Yoon. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014. doi:10.3115/v1/D14-1181

Manuscript Number 288