Learned in Translation: Contextualized Word Vectors

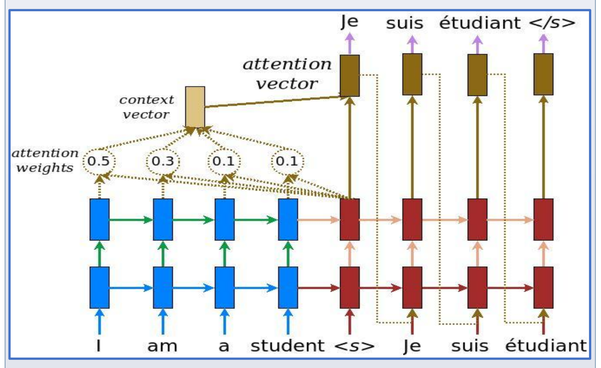
Aditya Agarwal, Hans Hanley, Adam Hare, Lisa Schut

Department of Computer Science, University of Oxford, Oxford, UK

Introduction

Recently, word embeddings have emerged as a topic of research in and of themselves, with the realization that they can be used as standalone trainable features in many NLP tasks. They encode surprisingly accurate syntactic and semantic word relationships and this work illustrates transfer learning from Machine Translation to another task of sentiment classification in form of the output encodings from the Translation Encoder.

Machine Translation



- Since 2014, the sequence-to-sequence models have become the state of the art in machine translation.
- In 2015-2016, attention mechanisms were introduced to better capture dependencies in long sentences.
- McCann et al. use a LSTM encoder-decoder model with global hard attention to perform neural machine translation.

Machine Translation Datasets

Machine Translation Datasets

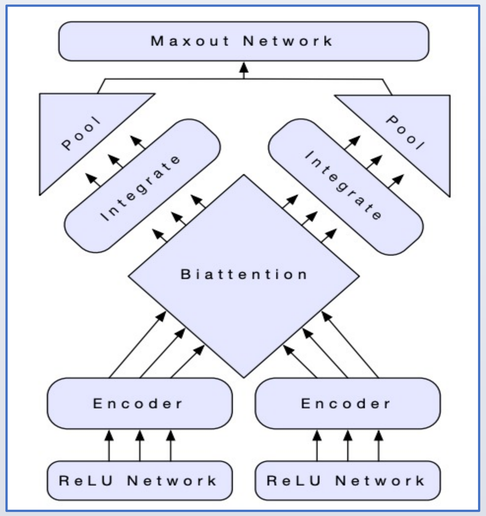
- Multi30K German-English (CoVe)
  - 29,000(train)/1014 (validation)/ 1000 (test)
- IWST German-English (CoVe-L)
  - 98,132/887/1565
- Europarl: WMT French- English (CoVe-F)
  - 100,000/2,000/2,000

Machine Translation Results

Run Number	Test Loss	Test Perplexity	BLEU Score
Run 1	29.76	11.35	32.17
Run 2	28.98	10.65	29.75
Run 3	29.73	10.58	31.27

Bi-Attentive Classification Network

- Task-specific embeddings are created using a RELU and a Bi-LSTM.
- Following Seo et al. (2017) and Xiong et al. (2017), they use a Biattentive mechanism to incorporate the interdependence between the input sequences.
- Next, the sequences are passed through an integration mechanism.
- The resulting sequence is put through an integration layer, undergoes 4 different types of pooling, and then a 3-layer batch-normalized maxout network.



Sentiment Datasets

Sentiment Classification Datasets

- Stanford Sentiment Treebank SST-5
  - 9,017/1,127/1,128
- Stanford Sentiment Treebank SST-2
  - 6,228/ 692/1,821

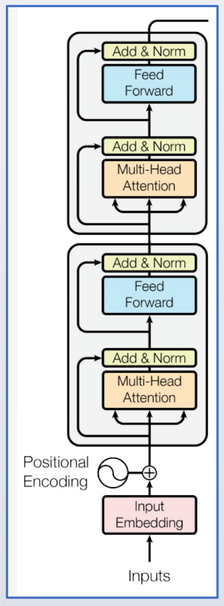
Extensions: ELMo

- ELMo embeddings are learned from the layers of two LSTMS that learn a bidirectional language model.
- ELMo is a character based encoding that like CoVe incorporates contextual information. ELMo's character based encoding is well generalizable to other tasks because it is able to easily handle words outside the original vocabulary.
- We compare how ELMo embeddings of the sentiment datasets compare to CoVe embeddings.

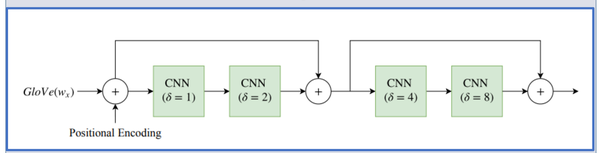
Extensions: Dropout

- We performed grid search for dropout and recurrent dropout for values from 0.1-0.5 for both values.
- We found the values of 0.2 for dropout and 0.1 for recurrent dropout gave us the best result.

Extensions: Transformer & CNN- Encoder



- Both Transformer and CNN benefit from being nonsequential, allowing for more efficient training.
- We kept the decoder from McCann et al. for both novel encoder models.
- Both encoder models use positional embeddings to incorporate temporal information from text.
- Our Transformer model from Vaswani et al. 2017.
- We make use of multi-headed attention. We used six heads instead of eight [Vaswani et al. 2017] for our model in order to obtain a 600-dimension embedding.
- Since 2016, CNNs have become popular in neural machine translation.
- Dilation has been shown to be an effective mechanism to increase the receptive field [Kalchbrenner et al. 2016].



Extensions: BCN Results

	SST-2		SST-5	
Embedding	Val	Test	Val	Test
Random	77.46	76.30	34.43	34.04
GloVe	81.50	78.30	46.81	45.92
Glove+Char	82.80	83.53	43.79	45.61
Baseline CoVe	81.07	80.29	45.16	47.07
CoVe	81.79	82.66	45.92	46.41
CoVe+Char	83.24	81.21	44.6	42.06
CoVe-Hyper	82.65	76.64	45.31	44.61
CoVe-L	80.78	81.94	38.12	37.80
CoVe-F	**83.96**	**83.67**	46.45	**47.20**
ELMo	77.31	74.52	41.61	42.73
Transformer	80.35	67.98	48.54	45.48
CNN	81.50	80.67	43.09	42.15
Baseline+Transformer	81.65	74.35	**48.80**	46.10
Baseline+CNN	82.08	82.54	44.19	42.82

Extensions: CNN for Sentiment Analysis

- We made use of a three-layer CNN for sentiment classification Ouyang et al.
- This network makes use of dropout, max-pooling, and normalization.

	SST-2		SST-5	
Embedding	Val	Test	Val	Test
Random	62.86	65.32	31.29	30.61
GloVe	80.49	80.78	42.20	40.20
Glove+Char	81.50	81.65	43.09	43.48
Baseline CoVe	81.50	82.36	43.79	44.90
CoVe	80.49	82.08	44.20	44.54
CoVe+Char	80.06	80.92	44.24	**46.68**
CoVe-L	80.06	80.35	40.73	42.64
CoVe-F	**83.24**	**85.12**	**49.02**	44.51
ELMo	79.48	76.30	45.92	44.63

Conclusions

- We were unable to reproduce the high accuracy results from McCann et al.
- We found the BCN training to be somewhat difficult and the BCN to sensitive to the hyperparameters and that may explain our lackluster results for sentiment classification.
- The results suggest that including contextual embeddings may improve downstream performance on sentiment classification
- We found the French dataset improved our downstream results compared with the larger German dataset
- The Transformer performed the best consistently on the SST-5 datasets. The Transformer performance improved when also combined with the Baseline CoVe embeddings. This shows that that Transformer may be learning additional underlying information from the dataset.
- We did not see a monotonic relationship with performance with machine translation and improvements in the downstream sentiment classification task

References

- McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. In Advances in Neural Information Processing Systems (pp. 6294-6305).
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
- Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A. V. D., Graves, A., & Kavukcuoglu, K. (2016). Neural machine translation in linear time. arXiv preprint arXiv:1610.10099.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
- J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel forimage captioning. arXiv preprint arXiv:1612.01887, 2016.
- M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. ICLR, 2017.