Supervised Learning of Universal Sentence Representations from Natural Language Inference Data

Unsupervised learning approach seems like a normal way to build word, sentence or document embeddings in NLP area, because it is more generalized such that pre-trained embeddings result can be transfer to other NLP downstream problems. For example, skip-gram in word embeddings and skip-though in sentence embeddings and distributed bag-of-words in paragraph embeddings. However, some unsupervised representations of sentences have not reached satisfactory enough performance to be widely adopted.

The author -- Conneau et al. noticed that supervised learning in ImageNet (Image Classification) doing good job in transferring result to downstream. Some features can be transferred to downstream somehow. For the same reason, Conneau et al. used textual entailment data to train a sentence embeddings layer which calls InferSent.

InferSent Design

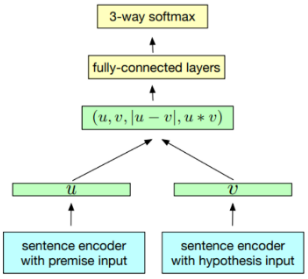
The idea is that team uses SNLI (Standford Natural Language Inference) data to train a model for Natural Language Inference (NLI) problem. NLI target to find the relationship between sentence 1 (premise) and sentence 2 (hypothesis). There are three categories which are entailment, contradiction, and neutral.

Authors believe that NLI is a suitable task to understand semantic relationships within sentences such that it helps to build a good embeddings for sentence embeddings for downstream NLP.

Architecture

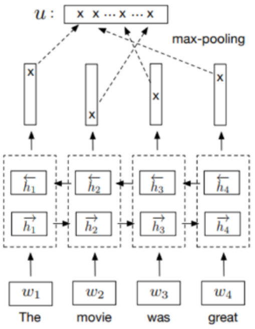
The overall idea is that two sentences input (premise and hypothesis) will be transformed by sentence encoder (by same weights). After that leveraging 3 matching methods to recognize relations between premise input and hypothesis input. Concatenation of two vectors→Element-wise product two vectors →Absolute element-wise difference of two vectors.

Then the author Conneau et al. evaluated 7 different architectures: (1) Standard LSTM, (2) Standard GRU, (3) Concatenation of last hidden states of forward and backward GRU, (4) Bi-directional LSTM with mean polling, (5) Bi-directional LSTM with max polling, (6) Self-attentive Network (Attention with BiLSTM), (7) Hierarchical convolutional networks

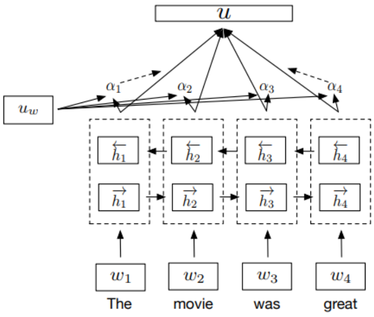

3-way softmax
fully-connected layers
$(u, v, |u-v|, u*v)$
u
v
sentence encoder with premise input
sentence encoder with hypothesis input

Conneau et al. (2017)

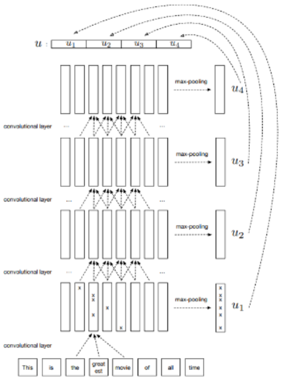
Before concluding the best approach first, we may believe that Attention with BiLSTM should be the best approach as attention mechanism helps to identify important weight. Actually, it may harm when using it in transfer learning. On the other hand, BiLSTM with mean polling perform not very good may due to unable to locate the important part.


u : x x ... x ... x
max-pooling
x
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_4$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_4$
w_1 w_2 w_3 w_4
The movie was great

#5 Bi-directional LSTM with max polling (Conneau et al., 2017)

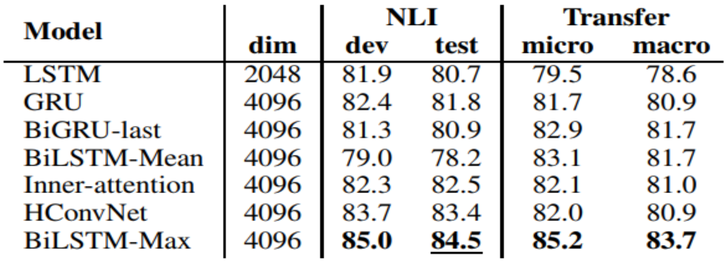

u
u_w
α_1 α_2 α_3 α_4
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_4$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_4$
w_1 w_2 w_3 w_4
The movie was great

#6 Self-attentive Network Architecture (Conneau et al., 2017)



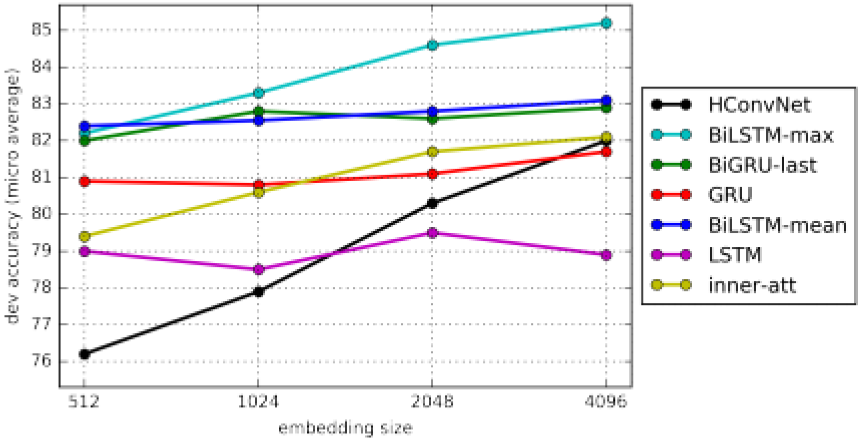
Hierarchical convolutional networks (Conneau et al., 2017)

From the experiment results, the best approach is Bi-directional LSTM with max-polling as following：



Model	dim	NLI		Transfer	
		dev	test	micro	macro
LSTM	2048	81.9	80.7	79.5	78.6
GRU	4096	82.4	81.8	81.7	80.9
BiGRU-last	4096	81.3	80.9	82.9	81.7
BiLSTM-Mean	4096	79.0	78.2	83.1	81.7
Inner-attention	4096	82.3	82.5	82.1	81.0
HConvNet	4096	83.7	83.4	82.0	80.9
BiLSTM-Max	4096	**85.0**	**84.5**	**85.2**	**83.7**

Conneau et al. (2017)



Implementation

There are 2 ways to use InferSent. One is using a pre-trained embeddings layer in NLP problems. Another is building InferSent using a original data. Since this InferSent uses supervised learning approach to generate sentence embeddings, we need to have a annotated (labeled) data first.

The most typical feature of InferSent is supervised learning to compute word vectors. InferSent leverages word embeddings (GloVe/ fastText) to build sentence embeddings. Pretrained model supports both GloVe (version 1) and fasttext (version 2)

Conclusion

This paper studies the effects of training sentence embeddings with supervised data by testing on different transfer tasks. It shows that models learned on NLI can perform better than models trained in unsupervised conditions or on other supervised tasks. For this paper, BiLSTM network with max pooling makes the best current universal sentence encoding methods, outperforming existing approaches like SkipThought vectors