Task-universal sentence embeddings from learning natural language inference

Yao Liu, Jeha Yang, Katherine Yu, Alex Kolchinski

{yaoliu, jeha, yukather, yakolch}@stanford.edu

Abstract

In this project, we show how fixed-dimensional sentence embeddings from encoders trained on Stanford Natural Language Inference (SNLI) dataset [1] and a new dataset we generated, derived from Stanford Question Answering (SQuAD) dataset [2] could be transfered into many other semantic tasks, especially tasks with little training data. The NLI baseline is based on Conneau et al. ([3]). We use following models to learn our sentence embeddings and compare the results on both SNLI and transfer task.

- Siamese model with BiLSTM as encoder, followed by MLP as classifier [3].
- Some variants of siamese model with different encoder: Transformer [4], 2 layer BiLSTM.
- Decomposable attention model [5].

Dataset description

- **SNLI dataset** : *(premise,hypothesis)→{entailment,contradiction,neutral}* ; 549k/10k/10k of "balanced" train/dev/test set
  (Ex) Two blond women are hugging one another.
  + There are women showing affection. (entailment)
  + The women are sleeping. (contradiction)
  + Some women are hugging on vacation. (neutral)
- **"ClassifSQuAD"** - new classification dataset generated from SQuAD, inspired by NLI datasets:
  (question, answer1, answer2) → whether answer1 is correct;
  We generated positive samples from original QA pairs and negative samples from answers to other questions within the same article, checking that they are not the same or substrings. Importantly, we filtered any examples which had OOV's in either the question or answer and any answer with fewer than 5 words since shorter answers were mostly undescriptive named entities.
  745k/128k train/dev; unique questions: 13.2k/2.1k train/dev.
  (Ex) What changes macroscopic closed system energies?
  + **internal energies of the system**(*correct*)
  + directed toward the center of the curving path (wrong)

  (Ex) For what cause is money raised at the Bengal Bouts tournament at Notre Dame?
  + **the holy cross missions in bangladesh**(*correct*)
  + a golden statue of the virgin mary (wrong)

  (Ex) What was the cost for a half minute ad?
  + **$ 5 million for a 30-second**(*correct*)
  + newton was limited by denver 's defense (wrong)
- **Transfer task evaluation data**: from SentEval [3].
  - Sentence classification: sentiment analysis (MR, SST), product reviews (CR), subjectivity/objectivity (SUBJ) and opinion polarity (MPQA).
  - Semantic inference: SICK-E(Entailment), SICK-R(Relatedness).
  - Semantic textual similarity: STS14.

Models

Siamese for generating sentence embeddings
Siamese models are by far not the best-performing models on the two training tasks due to not using intersentence word-by-word attention; however, we need to use siamese training for to produce generic word embeddings since sharing the same encoder parameters allows the single encoder to learn from all sentences in the training data and at inference on a single sentence, we do not have a target sentence (we cannot use seq2seq attention).

We tried the following encoder architectures with Siamese training:

- Bidirectional LSTM [3], 2-layer Bidirectional LSTM
- Transformer [4] [1]

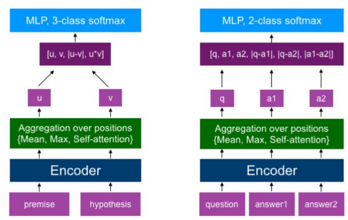
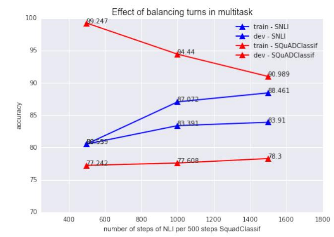


Figure 1: NLI Siamese architecture (left, [3]) and ClassifSQuAD architecture (right). We try to multitask-train these two tasks.

[1] Using code from https://github.com/jadore801120/attention-is-all-you-need-pytorch

Multitask training We (unconventially) train each of the two tasks for many consecutive steps (e.g. 1000 steps NLI, 500 steps ClassifSQuAD) and find that each task can recover and improve on its previous train loss very quickly (within 100 batches) after the other tasks's turn (the reason for this initially was that we wanted to compare in-epoch progress against a reference single-task learning curve). I t seemed actually important to use different batch sizes for the two tasks: 64 for NLI and 128 for ClassifSQuAD; thus, we generally take fewer steps in ClassifSQuAD.



S2S Methods We also tried models with more joint attention between sentences such as decomposable attention model [5]. The main aim here is to verify that although these models could achieve good performance in NLI dataset as shown below, it might not be a good choice for learning transferable representations. Our results were within 2-3 points of published results. [2]

[2] Adapting code from https://github.com/libowen2121/SNLI-decomposable-attention

Transfer task results

Model	MR	CR	MPQA	SUBJ	SST-B	SST-F	SICK-E	SICK-R	STS
Conneau et al. BLSTM(max)	79.9	84.6	**89.8**	92.1	**83.3**	-	**86.3**	**0.885**	.68/.65
BLSTM(max)	81.32	84.11	89.19	**93.02**	81.0	42.08	85.18	0.8727	.68/.65
BLSTM(max,mean)	**81.33**	84.56	89.3	92.19	80.45	40.23	85.83	0.8812	.66/.64
2-BLSTM(max)	80.71	83.21	88.86	91.86	74.9	37.1	83.93	0.868	.68/.64
2-BLSTM(max,mean)	81.2	83.79	89.02	92.49	76.28	39.28	84.96	0.875	.66/.64
Transformer(max)	72.57	76.06	87.12	89.3	75.23	39.68	82.59	0.8467	.66/.64
Transformer(max,mean)	74.22	76.24	87.85	90.6	76.99	41.4	82.79	0.8557	.63/.62
Multitask LSTM(max)	80.9	**84.82**	89.68	92.74	80.23	**42.44**	85.41	0.8695	**.69/.67**
Decomposable Att (Max)	70.59	74.89	86.63	87.38	72.87	35.79	79.26	0.817	.41/.44
Decomposable Att (Sum)	73.52	76.95	86.27	89.66	78.36	39.05	74.1	0.767	.56/.55
Decomposable Att (Max,Mean)	73.4	77.01	87.84	89.75	76.33	39.14	80.6	0.825	.60/.58

SNLI results

Model	Train Acc	Dev Acc	Test Acc
BLSTM	84.122	83.350	83.429
2-BLSTM	85.556	83.062	82.504
Multitask LSTM	87.072	83.393	82.874
Decomposable Attention	83.062	**84.088**	**83.926**
Siamese Transformer	83.565	82.692	82.597

Discussion and future works

- Siamese methods with BiLSTM encoder from [3] achieved best performance in some tasks, while Multitask training derived from SQUAD dataset did so in other tasks.
- Multitask training seems promising, considering we only use vanilla LSTM with max-pooling. This might imply that multitask training could avoid learning features that heavily depends on SNLI dataset or NLI structure.
- Concatenation of max-pool and mean-pool could help many encoders achieve better transfer performance.
- Decomposable attention model does well on NLI dataset but could not learn good transferable embeddings because they rely on inter-sentence attention.

For future work, we would further develop the multitask model! We would try adding additional training data such as MultNLI, Quora Question Pairs, as well as different word embeddings (word2vec, train our own) since they are very important to this model. We also want to work on a decomposition analysis on the set of transfer/SentEval tasks so we can understand performance by important task characteristics like sentence length, OOV rate, and relative size of the training data.

References

[1] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning.
A large annotated corpus for learning natural language inference.
arXiv preprint arXiv:1508.05326, 2015.

[2] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang.
Squad: 100,000+ questions for machine comprehension of text.
arXiv preprint arXiv:1606.05250, 2016.

[3] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes.
Supervised learning of universal sentence representations from natural language inference data.
arXiv preprint arXiv:1705.02364, 2017.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
pages 6000–6010, 2017.

[5] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit.
A decomposable attention model for natural language inference.
arXiv preprint arXiv:1606.01933, 2016.