Learning General Purpose Distributed Sentence Representations Via Large Scale Multi-task Learning

10-605 Team Project

Kuo Tian, Zehao Guan, Zeyu Peng

Outline

- Background
- Related work
- Model
- Experiments
- Conclusion
- Reference


Published as a conference paper at ICLR 2018

LEARNING GENERAL PURPOSE DISTRIBUTED SENTENCE REPRESENTATIONS VIA LARGE SCALE MULTI-TASK LEARNING

Sandeep Subramanian[1,2,3]*, **Adam Trischler**[3], **Yoshua Bengio**[1,2,4] **& Christopher J Pal**[1,5]
[1] Montréal Institute for Learning Algorithms (MILA)
[2] Université de Montréal
[3] Microsoft Research Montreal
[4] CIFAR Senior Fellow
[5] École Polytechnique de Montréal

ABSTRACT

A lot of the recent success in natural language processing (NLP) has been driven by distributed vector representations of words trained on large amounts of text in an unsupervised manner. These representations are typically used as general purpose features for words across a range of NLP problems. However, extending this success to learning representations of sequences of words, such as sentences, remains an open problem. Recent work has explored unsupervised as well as supervised learning techniques with different training objectives to learn general purpose fixed-length sentence representations. In this work, we present a simple, effective multi-task learning framework for sentence representations that combines the inductive biases of diverse training objectives in a single model. We train this model on several data sources with multiple training objectives on over 100 million sentences. Extensive experiments demonstrate that sharing a single recurrent sentence encoder across weakly related tasks leads to consistent improvements over previous methods. We present substantial improvements in the context of transfer learning and low-resource settings using our learned general-purpose representations.[1]

1 INTRODUCTION

Transfer learning has driven a number of recent successes in computer vision and NLP. Computer vision tasks like image captioning (Xu et al., 2015) and visual question answering typically use CNNs pretrained on ImageNet (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014) to extract representations of the image, while several natural language tasks such as reading comprehension and sequence labeling (Lample et al., 2016) have benefited from pretrained word embeddings (Mikolov et al., 2013; Pennington et al., 2014) that are either fine-tuned for a specific task or held fixed.

Many neural NLP systems are initialized with pretrained word embeddings but learn their representations of *words in context* from scratch, in a task-specific manner from supervised learning signals. However, learning these representations reliably from scratch is not always feasible, especially in low-resource settings, where we believe that using general purpose sentence representations will be beneficial.

Some recent work has addressed this by learning general-purpose sentence representations (Kiros et al., 2015; Wieting et al., 2015; Hill et al., 2016; Conneau et al., 2017; McCann et al., 2017; Jernite et al., 2017; Nie et al., 2017; Pagliardini et al., 2017). However, there exists no clear consensus yet on what training objective or methodology is best suited to this goal.

Understanding the inductive biases of distinct neural models is important for guiding progress in representation learning. Shi et al. (2016) and Belinkov et al. (2017) demonstrate that neural ma-

*Work done while author was an intern at Microsoft Research Montreal
[1] Code will be made available at https://github.com/Maluuba/gensen

Background

- **Recent Success in NLP**
  - Distributed vector representation of words trained on large amount of text in an unsupervised manner
- **Problem**
  - Extending this success to learning representations of sequences of words, such as sentences
- **This project**
  - First large-scale reusable sentence representation model
  - Single model contains inductive bias of diverse training objectives, over 100 million sentences
  - Substantial improvement in transfer learning and low-resources settings

Related Work

- **Learn representations from scratch**
  - Neural architectures for named entity recognition 2016
  - Global vectors for word representation 2014

- **General-purpose sentence representations**
  - Sentence representation learning from explicit discourse relations 2017
  - Unsupervised learning of sentence embeddings using compositional N-gram features 2017

- **Similar work: Multi-task sequence to sequence learning (2015)**
  - Attention mechanism prevents learning a fixed length vector representation for sentence
  - Aims for improvement on the same tasks that the model is trained V.S. re-usable sentence representations

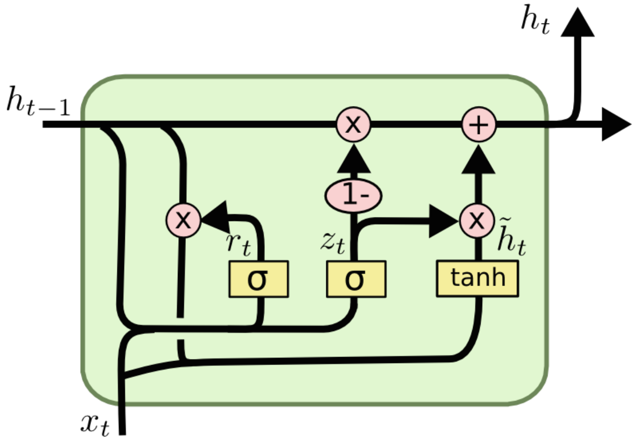
https://arxiv.org/abs/1804.00079

Model -- dataset and architecture

- Datasets for different tasks

Task	Sentence Pairs
En-Fr (WMT14)	40M
En-De (WMT15)	5M
Skipthought (BookCorpus)	74M
AllNLI (SNLI + MultiNLI)	1M
Parsing (PTB + 1-billion word)	4M
Total	124M

- Seq2seq model with GRU



https://arxiv.org/abs/1804.00079

Model -- multi-task algorithm

Require: A set of k tasks with a common source language, a shared encoder $\mathbf{E}$ across all tasks and a set of k task specific decoders $\mathbf{D_1} \ldots \mathbf{D_k}$. Let θ denote each model's parameters, α a probability vector $(p_1 \ldots p_k)$ denoting the probability of sampling a task such that $\Sigma_i^k p_i = 1$, datasets for each task $\mathbb{P}_1 \ldots \mathbb{P}_\mathrm{k}$ and a loss function L.

while θ *has not converged* **do**

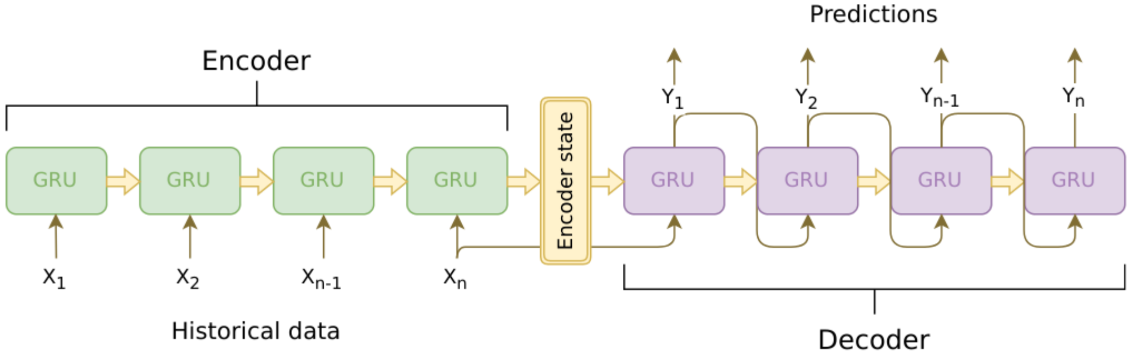
1: Sample task $i \sim \mathbf{Cat}(k, \alpha)$.
2: Sample input, output pairs $\mathbf{x}, \mathbf{y} \sim \mathbb{P}_\mathrm{i}$.
3: Input representation $\mathbf{h}_x \leftarrow \mathbf{E}_\theta(\mathbf{x})$.
4: Prediction $\tilde{\mathbf{y}} \leftarrow \mathbf{D}_{i_\theta}(\mathbf{h}_x)$
5: $\theta \leftarrow \mathrm{Adam}(\nabla_\theta L(\mathbf{y}, \tilde{\mathbf{y}}))$.

end

https://arxiv.org/abs/1804.00079

Model -- detail settings

- **Encoder**
  - Bidirectional GRU
- **Decoder**
  - Conditional GRU
- **Parameters**
  - Hidden units: 2048
  - Minibatch size: 128
  - Optimizer: Adam
  - Word embedding dimension: 512



https://arxiv.org/abs/1804.00079

Experiments -- training objectives

- **Goal**
  - Sufficient diversity
  - Existence of fairly large datasets for training
  - Success as standalone objectives for sentence representations
- **Training tasks**
  - Skip-thought vectors (STP + STN)
  - Neural machine translation (Fr + De)
  - Constituency parsing (linearized parse tree construction)
  - Natural language inference (NLI)

Experiments -- evaluation

- **Text classification**
  - Movie reviews (MR), product reviews (CR), Stanford Sentiment (SST), opinion polarity (MPQA)
  - question type classification (TREC), subjectivity/objectivity classification (SUBJ)
- **Paraphrase identification**
  - Microsoft Research Paraphrase Corpus (MRPC)
- **Entailment and semantic relatedness**
  - SICK-R, SICK-E datasets
- **Semantic textual similarity**
  - STS benchmark from 2012-2016
- **Sentence characteristics & syntax**
  - Top syntactic sequence (TSS)

Experiments -- improvement

- **Proposed model**
  - Small batch size
  - Perform all tasks in single machine parallel mode
  - Result in long training process

- **Our improvement**
  - Increase training batch size
  - Distribute different tasks on multiple workers
  - Significantly shorten training time

https://github.com/Kuo-T/10605Proj

Experiments -- results

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STSB	Δ
Transfer Approaches											
NMT En-Fr	64.7	70.1	84.9	81.5	-	82.8	-	-	-	-	-
CNN-LSTM	77.8	82.1	93.6	89.4	-	92.6	76.5	0.862	-	-	-
Skipthought + LN	79.4	83.1	93.7	89.3	82.9	88.4	-	0.858	79.5	70.2	-
Naïve Bayes SVM	79.4	81.8	93.2	86.3	83.1	-	-	-	-	-	-
Infersent (SNLI)	79.9	84.6	92.1	89.8	83.3	88.7	75.1	0.885	86.3	-	-
Infersent (AllNLI)	81.1	86.3	92.4	90.2	**84.6**	88.2	76.2	0.884	86.3	75.5	0
Our Models											
+STN	78.9	85.8	93.7	87.2	80.4	84.2	72.4	0.84	82.1	72.4	-2.56
+STN +Fr +De	80.3	85.1	93.5	90.1	83.3	92.6	77.1	0.864	84.8	77.1	0.01
+STN +Fr +De +NLI	81.2	86.4	93.4	90.8	84	93.2	76.6	0.884	87	79.1	0.99
+STN +Fr +De +NLI +L	81.7	87.3	**94.3**	90.8	84	**94.2**	77.1	0.887	87.1	78.2	1.33
+STN +Fr +De +NLI +L +STP	**82.8**	**87.9**	94.2	91	84.5	92.4	78.2	0.885	86.2	78.4	1.46
+STN +Fr +De +NLI +2L +STP	82.7	87.5	93.9	**91.1**	82.8	92.6	77.4	0.884	87.6	**79.2**	1.49
+STN +Fr +De +NLI +L +STP +Par	82.5	87.6	94.1	90.8	83.2	93	**78.6**	**0.888**	**87.8**	78.6	**1.51**

Evaluations of sentence representations on set of 10 tasks. ***Δ*** indicates average improvement over Infersent (AllNLI) across all 10 tasks. **Underlines** are used for each task to indicate both our best performing model as well as the best transferring model that isn't ours.

https://github.com/Kuo-T/10605Proj/tree/master/log_files

Experiments -- log files

nli_large +
nli_large_bothskip

```
--------------------------------------------
Table 1 of Our Paper :
--------------------------------------------
MR                 [Dev:83.8/Test:82.8]
CR                 [Dev:88.9/Test:87.9]
SUBJ               [Dev:94.5/Test:94.2]
MPQA               [Dev:91.4/Test:91.0]
SST2               [Dev:85.8/Test:84.4]
SST5               [Dev:46.4/Test:46.6]
TREC               [Dev:90.3/Test:92.4]
MRPC               [Dev:78.2/TestAcc:78.7/TestF1:84.3]
SICKRelatedness    [Dev:0.884/Test:0.884]
SICKEntailment     [Dev:86.2/Test:86.8]
STS12              [Pearson:0.607/Spearman:0.610]
STS13              [Pearson:0.547/Spearman:0.561]
STS14              [Pearson:0.658/Spearman:0.643]
STS15              [Pearson:0.742/Spearman:0.745]
STS16              [Pearson:0.664/Spearman:0.667]
STSBenchmark       [Dev:0.81219/Pearson:0.78417/Spearman:0.78702]
--------------------------------------------
```

nli_large_bothskip +
nli_large_bothskip_2layer

```
--------------------------------------------
Table 2 of Our Paper :
--------------------------------------------
MR                 [Dev:83.6/Test:82.7]
CR                 [Dev:89.0/Test:87.5]
SUBJ               [Dev:94.4/Test:93.9]
MPQA               [Dev:91.5/Test:91.1]
SST2               [Dev:86.9/Test:83.9]
SST5               [Dev:47.6/Test:45.9]
TREC               [Dev:89.6/Test:92.6]
MRPC               [Dev:78.2/TestAcc:76.8/TestF1:82.8]
SICKRelatedness    [Dev:0.888/Test:0.884]
SICKEntailment     [Dev:86.2/Test:87.7]
STS12              [Pearson:0.597/Spearman:0.604]
STS13              [Pearson:0.539/Spearman:0.555]
STS14              [Pearson:0.632/Spearman:0.619]
STS15              [Pearson:0.721/Spearman:0.723]
STS16              [Pearson:0.655/Spearman:0.658]
STSBenchmark       [Dev:0.80435/Pearson:0.79049/Spearman:0.79244]
--------------------------------------------
```

nli_large_bothskip_parse +
nli_large_bothskip_2layer

```
--------------------------------------------
Table 3 of Our Paper :
--------------------------------------------
MR                 [Dev:83.5/Test:82.8]
CR                 [Dev:88.9/Test:87.9]
SUBJ               [Dev:94.5/Test:94.1]
MPQA               [Dev:91.4/Test:91.0]
SST2               [Dev:85.0/Test:84.0]
SST5               [Dev:47.4/Test:46.2]
TREC               [Dev:89.0/Test:92.2]
MRPC               [Dev:78.6/TestAcc:77.4/TestF1:84.5]
SICKRelatedness    [Dev:0.894/Test:0.886]
SICKEntailment     [Dev:85.8/Test:87.5]
STS12              [Pearson:0.600/Spearman:0.602]
STS13              [Pearson:0.524/Spearman:0.535]
STS14              [Pearson:0.638/Spearman:0.625]
STS15              [Pearson:0.724/Spearman:0.724]
STS16              [Pearson:0.661/Spearman:0.666]
STSBenchmark       [Dev:0.81389/Pearson:0.78482/Spearman:0.78778]
--------------------------------------------
```

https://github.com/Kuo-T/10605Proj

Conclusion

- **Highlights**
  - Scale up training, Multi-task, Large scale dataset
  - Outperformed most prior works on most tasks
- **Concerns**
  - Fixed length representations may not be suitable for complex, long piece of text
  - Not clear whether the performance improvement comes from having more unlabeled data (even if it is trained with the same training objective) or having multiple training objectives

Reference

[1] Tang, Shuai, et al. "Rethinking skip-thought: A neighborhood based approach." *arXiv preprint arXiv:1706.03146* (2017)

[2] Gan, Zhe, et al. "Unsupervised learning of sentence representations using convolutional neural networks." *arXiv preprint arXiv:1611.07897* (2016)

[3] Dong, Daxiang, et al. "Multi-task learning for multiple language translation." *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1. 2015

[4] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014)

[5] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014)