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

Kuo Tian
Carnegie Mellon University
Pittsburgh, PA 15213
kuot@andrew.cmu.edu

Zehao Guan
Carnegie Mellon University
Pittsburgh, PA 15213
zehaog@andrew.cmu.edu

Zeyu Peng
Carnegie Mellon University
Pittsburgh, PA 15213
zeyupeng@andrew.cmu.edu

Abstract

Recent natural language processing (NLP) tasks, such as reading comprehension and sequence labeling, have benefited from the distributed vector representations of words trained on large amount of unlabeled text data. Nevertheless, learning a sequence of words as a sentence, is still a big challenge. In this project, we present a simple, effective multi-task learning framework for sentence representations and then combine the inductive biases of diverse training objectives in a single model. After being trained over 100 million sentences, our model outperforms other previous methods across weakly related tasks. This model also shows great improvement in the context of transfer learning and low-resource settings.

1 Introduction

To generalize across diverse set of tasks, it is significant to build representations that encode several aspects of sentences. Neural approaches to tasks such as skip-thoughts, machine translation, natural language inference, and constituency parsing likely have different inductive biases. Our work exploits this in the context of a simple one-to-many multi-task learning (MTL) framework, wherein a single recurrent sentence encoder is shared across multiple tasks. We hypothesize that sentence representations learned by training on a reasonably large number of weakly related tasks will generalize better to novel tasks unseen during training, since this process encodes the inductive biases of multiple models.

The primary contribution of our work is to combine the benefits of diverse sentence-representation learning objectives into a single multi-task framework. To the best of our knowledge, this is the first large-scale reusable sentence representation model obtained by combining a set of training objectives with the level of diversity explored here. While our work aims at learning fixed-length distributed sentence representations, it is not always practical to assume that the entire "meaning" of a sentence can be encoded into a fixed-length vector. We demonstrate through extensive experimentation that representations learned in this way lead to improved performance across a diverse set of novel tasks not used in the learning of our representations. Such representations facilitate low-resource learning as exhibited by significant improvements to model performance for new tasks in the low labelled data regime - achieving comparable performance to a few models trained from scratch using only 6% of the available training set on the Quora duplicate question dataset.

2 Related Work

So far several papers have demonstrated that neural machine translation systems and sequence to sequence parsers appear to capture morphology and certain syntactic properties. Our work is most similar to that of Luong et al. (2015)[1], who train a many-to-many sequence- to-sequence model on a diverse set of weakly related tasks that includes machine translation, constituency parsing, image captioning, sequence autoencoding, and intra-sentence skip-thoughts. There are two key differences between that work and our own. First, like McCann et al. (2017)[2], their use of an attention mechanism prevents learning a fixed-length vector representation for a sen- tence. Second, their work aims for improvements on the same tasks on which the model is trained, as opposed to learning re-usable sentence representations that transfer elsewhere. We further present a fine-grained analysis of how different tasks contribute to the encoding of dif- ferent information signals in our representations following work by Shi et al. (2016)[3] and Adi et al. (2016)[4].

3 Sequence-to-Sequence Learning

3.1 Sequence-to-Sequence model

The tasks that we consider for multi-task learning are formulated as sequence-to-sequence problems[5]. The sequence-to-sequence model is a specific case of encoder-decoder models dealing with sequential inputs and outputs. Basically, the input x and output y are sequences $x_1, x_2, ..., x_m$ and $y_1, y_2, .., y_n$. The encoder produces a fixed length vector representation $\mathbf{h_x}$ of the input, which the decoder then conditions on to generate an output. The decoder breaks down the joint probability of outputs to a product of conditional probabilities via the chain rule:

$$\mathcal{P}(y|x) \quad = \quad \prod_{i=1}^{n} \mathcal{P}(y_i|y_{<i}, \mathbf{h_x})$$

The encoder and decoder can be parameterized as RNN variants such as Long Short-term Memory (LSTMs) or Gated Recurrent Units (GRUs)[6]. Our model used the latter one as the units of the sequence-to-sequence model. The hidden representation h_x is the last hidden state of the encoder. Our goal in this work is to get a compressed hidden representation of sentences which can be used to perform various tasks. The sequence-to-sequence model gives a single, fixed-length and distributed representation. We use GRU for encoder and decoder in interest of computational speed. The encoder is a bidirectional GRU while the decoder is a unidirectional conditional GRU whose parameterization is as follows:

$$\begin{aligned}
\mathbf{r_t} &= \sigma(\mathbf{W_r} x_t + \mathbf{U_r h_{t-1}} + \mathbf{C_r h_x}) \\
\mathbf{z_t} &= \sigma(\mathbf{W_z} x_t + \mathbf{U_z h_{t-1}} + \mathbf{C_z h_x}) \\
\hat{\mathbf{h_t}} &= tanh(\mathbf{W_d} x_t + \mathbf{U_d}(\mathbf{r_t h_{t-1}}) + \mathbf{C_d h_x}) \\
\mathbf{h_{t+1}} &= (1 - \mathbf{z_t})\mathbf{h_{t-1}} + \mathbf{z_t}\hat{\mathbf{h_t}}
\end{aligned}$$

3.2 Multi-task Sequence-to-Sequence Learning

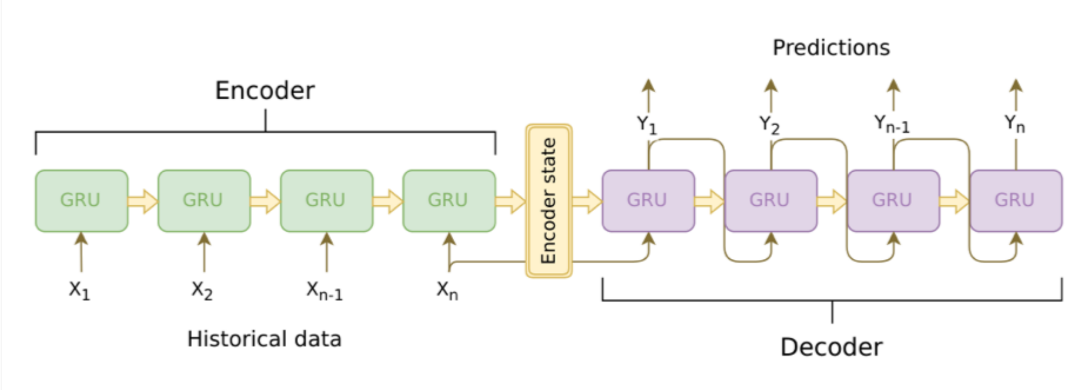


Figure 1: Architecture of our model

We implement the multi-task learning on our sequence-to-sequence model based on the idea of combining inductive biases from different training objectives. The same encoder processes the input sentences from different tasks into a compressed summary h_x which is then used to condition a task-specific GRU to produce the output sentence[7]. Figure 1 shows the structure of the sequence-to-sequence model used in this work.

4 Multi-task Training Setup

4.1 Training tasks

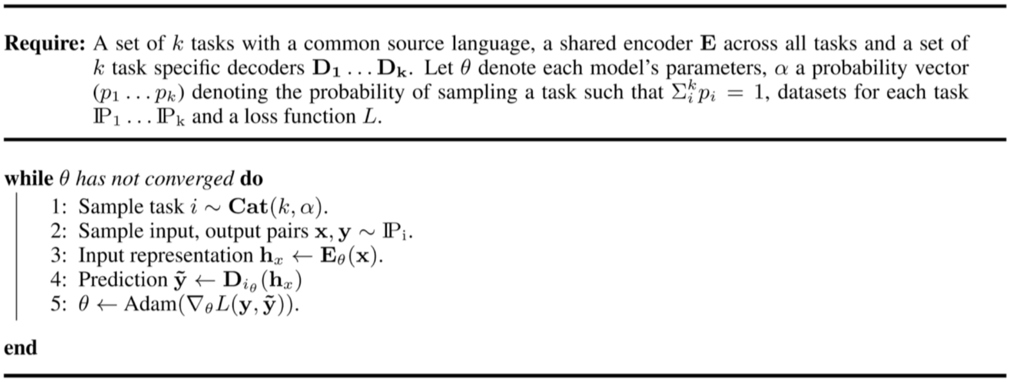
Our motivation for multi-task training lies in two aspects. On one hand, learning multiple related tasks jointly results in good generalization as measured by the number of training examples required per task. On the other hand, inductive bias learned on sufficiently training tasks are likely to be good for learning novel tasks drawn from the same environment.

As described in section 3, the goal of our model is using the sequence-to-sequence model to learn a hidden representation which can be applied to complete other tasks. Specifically, most of the tasks are language-related tasks and have available large amount dataset. In our work, we trained our model on 4 tasks.

- **Skip-thought vectors**
  Skip-thought vectors[8] task's learning objective is to simultaneously predict next and previous sentences from the current sentence. The encoder for the current sentence and decoder for the previous (**STP**) and next sentence (**STN**) are parameterized as different RNNs.[9]
- **Neural Machine Translation**
  This can be formulated as a sequence-to-sequence learning problem where the input is a sentence in source language and the output is its corresponding translation in target sentence.[10]
- **Constituency Parsing**
  The input to encoder is the sentence itself and the decoder produces its linearized parse tree.[11]
- **Natural Language Inference**
  Natural language inference is a 3-way classification problem. Given a premise and a hypothesis sentence, the objective is to classify their relationship as either entailment, contradiction, or neutral.[12]

4.2 Multi-task algorithm

The original multi-task learning algorithm in the paper is shown in Figure 2. We conduct the sampling on different tasks and train our model, calculate the loss between our results and the real outputs,[13] then using Adam criterion to update the model parameters. We find that the training process can be accelerated if the tasks can be assigned to different worker machines, which is carried out in the experiment phase.[14]



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 $\theta$ has not converged do
    Sample task $i \sim \mathbf{Cat}(k, \alpha)$.
    Sample input, output pairs $\mathbf{x}, \mathbf{y} \sim \mathbb{P}_\mathrm{i}$.
    Input representation $\mathbf{h}_x \leftarrow \mathbf{E}_\theta(\mathbf{x})$.
    Prediction $\tilde{\mathbf{y}} \leftarrow \mathbf{D}_{i_\theta}(\mathbf{h}_x)$
    $\theta \leftarrow \mathrm{Adam}(\nabla_\theta L(\mathbf{y}, \tilde{\mathbf{y}}))$.
end
```

Figure 2: Multi-task training algorithm

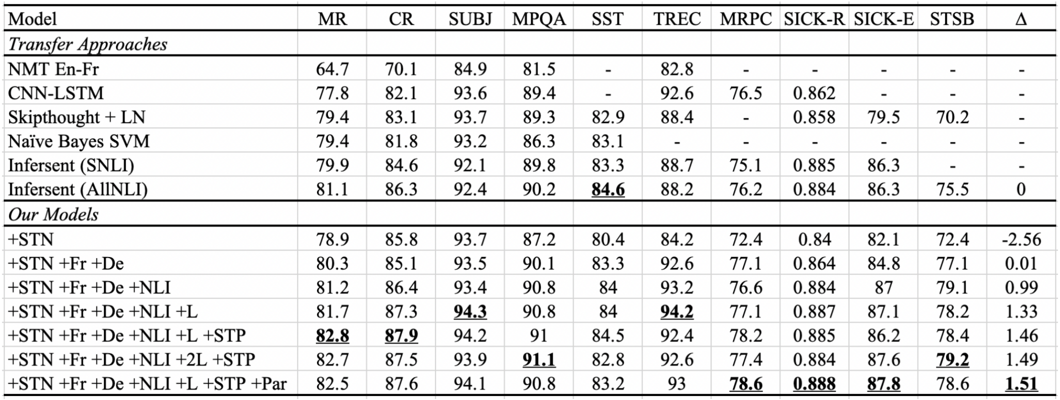
5 Experiments and Results

5.1 Datasets

We train out model on 5 different datasets corresponding to the learning tasks mentioned in section 2. For skip-thought vectors task, we use the BookCorpus dataset. For neural machine translation, we use a corpus of around 4.5 million English-German sentence pairs from WMT15 and 40 million English-French sentence pairs from WMT14. For the constituency parsing task, the dataset mentioned in the paper is not available so we didn't train our model on this task. We basically load the open-source pre-trained parameters of this task. We perform the NLI task on SNLI and MultiNLI.

5.2 Evaluation

We train and test our model on 4 8-GPU V100 GCP instances, and assign 1 server as the master machine, 3 other machines conduct the 3 learing tasks in parallel. Then we use SentEval toolkit of Facebook to perform the 10 evaluation tasks. We obtain the results about evaluation of sentence representations on a set of 10 tasks are listed in the following table.



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**

Figure 3: Evaluation of sentence representatons

In this table, Δ indicates the average improvement over Infersent (AllNLI) across all 10 tasks. Bold numbers indicate the best performing transfer model on a given task. Underlines are used for each task to indicate both our best performing model as well as the best performing transfer model that isn't ours.

After multi-task learning, our experimental results show that by increasing training batch size, and distribute the multiple tasks on multiple workers, we can significantly decrease the training time, and get average improvement in comparison to the baseline model.

6 Conclusion

By training on large scale dataset and with various task objectives, we present a multi-task framework for learning general-purpose fixed-length sentence representations. We demonstrate that the learned representations yield competitive or superior results to previous general-purpose sentence representation methods. We also observe that this approach produces good word embeddings.

Nevertheless, there are several things in this paper that need further clarification and experiments. First of all, there is an impressive number of experiments had been done, but the results are a bit mixed, and it is not always clear that adding more tasks help. Even though this paper addresses an important problem of learning general purpose sentence representations, there is no definitive conclusion. For example, it is 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. The method in the paper needs long time to run because of the lack of parallelism. Another weakness of this paper is that fixed length representations may not be suitable for complex, long pieces of text (often, sentences), such representations may be useful for several tasks.

7 References

[1] Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. Evaluation of word vector representations by subspace alignment. 2015.

[2]Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. arXiv preprint arXiv:1511.06114, 2015.

[3]Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. arXiv preprint arXiv:1708.00107, 2017.

[4]Xing Shi, Inkit Padhi, and Kevin Knight. Does string-based neural mt learn source syntax? In EMNLP, pp. 1526–1534, 2016.

[5]Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. arXiv preprint arXiv:1608.04207, 2016.

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

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

[8] 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.

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

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

[11] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. arXiv preprint arXiv:1608.04207, 2016.

[12] Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. arXiv preprint arXiv:1602.03483, 2016.

[13] Nikola Mrksic , Ivan Vulic , Diarmuid O Seaghdha, Ira Leviant, Roi Reichart, Milica Gasic , Anna Korhonen, and Steve Young. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. arXiv preprint arXiv:1706.00374, 2017.

[14] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. arXiv preprint arXiv:1702.03814, 2017.

8 Appendix

Paper link: https://arxiv.org/abs/1804.00079

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