Contrastive Learning for Low Resource Machine Translation

Anonymous ACL submission

Abstract

Representation learning plays a vital role in natural language processing tasks. More recent works study the geometry of the representation space for each layer of pre-trained language models. They find that the context representation of all words is not isotropic in any layer of the pre-trained language model. However, how contextual are the contextualized representations produced by transformer-based machine translation models? In this paper, we find that the contextualized representations of the same word in different contexts have a greater cosine similarity than those of two different words, but this self-similarity is low between the same words. This suggests that output of machine translation models produce more context-specific representations. In this work, we present a contrastive framework for machine translation, that adopts contrastive learning to train model in a supervised way. By making use of data augmentation, our supervised contrastive learning method solves the issue of low-resource machine translation representations learning. Experimental results on the IWSLT14 and WMT14 datasets show our method can outperform competitive baselines significantly.

1 Introduction

Recent Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017) have achieved huge success. Still, these representations remain poorly understood. For instance, just how contextual are the contextualized representations produced by models? Are there infinitely many context-specific representations for each word, or are words essentially assigned one of a finite number of word-sense representations?

More recent works (Ethayarajh, 2019; Peters et al., 2018; Kurita et al., 2019) answer this question by studying the geometry of the representation space for each layer of pre-trained language models like BERT (Devlin et al., 2018), and GPT-2 (Radford et al., 2019). They find that the contextualized representations of all words are not isotropic in any layer of the contextualizing model. This suggests that upper layers of contextualizing models produce more context-specific representations. However, some analysis find that contextualized embeddings at the output layer of these powerful language models tend to degenerate and occupy an anisotropic cone in the vector space, which is called the representation degeneration problem.

To better understand the representations, Wang and Isola (2020) identify two key properties alignment and uniformity. Which takes alignment between semantically-related positive pairs and uniformity of the whole representation space to measure the quality of learned representations. In this work we use cos similarity to measure alignment and uniformity. Through empirical analysis, we find that low resource machine translation models greatly improve uniformity. However, the alignment also degrades drastically. While representations of the same word in different contexts still have a greater cosine similarity than those of two different words, this self-similarity is low between the same words.

As an alternative, forcing the representation of similar token to be mapped into similar outputs may suggest the usage of contrastive learning. Contrastive learning (Tian et al., 2020b; Chen and He, 2020; Caron et al., 2021) is a training approach popular in the computer vision field, which aims to bring representations of similar class or instances closer in the representation space, and move them further from different ones. With the success of contrastive learning in the computer vision field, there is an increasing interest in applying this method to NLP tasks (Jiang et al., 2020; Kim et al., 2021; Lee et al., 2021; Gunel et al., 2021; Gao et al., 2021).

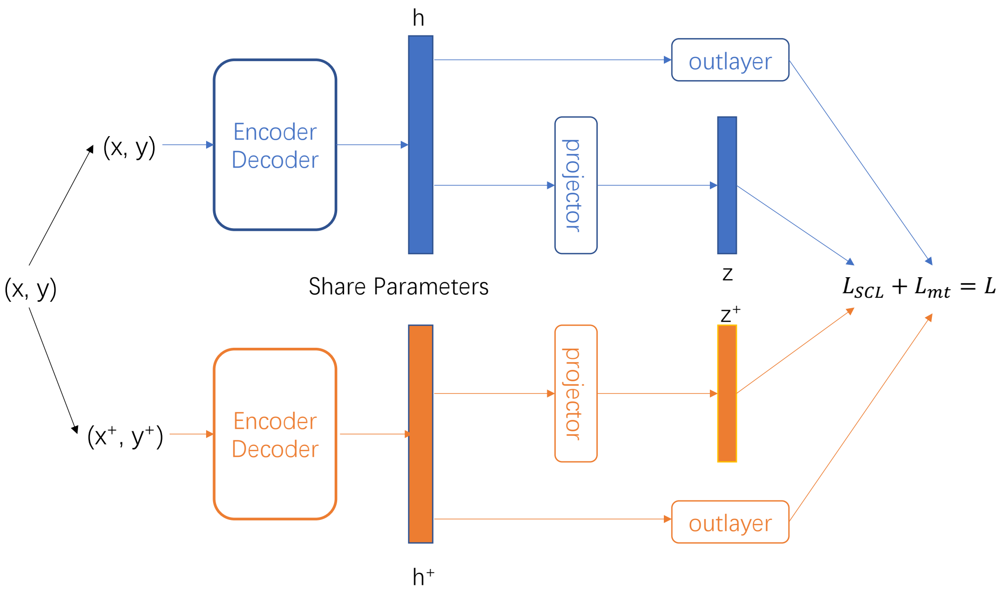

Figure 1: Basic architecture. Two augmented data of source and target sentence are processed by the same encoder-decoder network and a projector MLP. Then we apply contrastive loss to the representations z.

The common idea in these works is the following: pull together an anchor and a "positive" sample in embedding space, and push apart the anchor from many "negative" samples. Since no labels are available, a positive pair often consists of data augmentations of the sample, and negative pairs are formed by the anchor and randomly chosen samples from the mini-batch.

In this work, we propose a supervised contrastive learning (Khosla et al., 2021) with simple data augmentation. The representations of the same tokens are forced to be closer, while others from the mini-batch should be represented far from the anchor. We conducted experiments on the IWSLT14 WMT14 datasets and low data condition (1/5 of WMT14 training data), showing our method can outperform competitive baselines significantly.

2 Approach

2.1 Representation Similarity

We measure how contextual a word representation is using two different metrics: self-similarity and universal-similarity (Ethayarajh, 2019).

Let h be a token representation meanwhile h^+ means different contextual representations of the same token. The self similarity of token w is

$$\text{Self-Sim}(w) = \frac{1}{n^2 - n} \sum_{h} \sum_{h^+} \cos(h, h^+) \quad (1)$$

where cos denotes the cosine similarity. In other words, the self-similarity of a word w is the average cosine similarity between its contextualized representations across its n unique contexts. If token w does not contextualize the representations at all, then $\text{Self-Sim}(w) = 1$. The more contextualized the representations are for w, the lower we would expect its self-similarity to be.

Let h be a token representation meanwhile $h^{'}$ means different token representation by random sample. The universal similarity of token w is

$$\text{Un-Sim}(w) = \frac{1}{n^2 - n} \sum_{h} \sum_{h^{'}} \cos(h, h^{'}) \quad (2)$$

The universal representation similarity is the average cosine similarity between different tokens.

In the following sections, we will also use the two metrics to justify the inner workings of machine translation models.

2.2 Representation Learning Framework

Our approach is mainly inspired by SimCLR (Chen et al., 2020). As shown in Figure 1, there are four major components in our framework:

iwslt14	en-fr	fr-en	en-es	es-en	en-de	Avg
transformer	41.18	38.56	37.71	40.60	28.46	37.30
ours	42.23	40.68	38.96	41.60	29.82	38.66

Table 1: BLEU scores on IWSLT machine translation tasks.

- Data Augmentation. For each input sample, x, we generate two random augmentations, $x^+ = Aug(x)$, each of which represents a different view of the data and contains some subset of the information in the original sample.

- Encoder-Decoder Network, which maps inputs to representation vectors. Both augmented samples are separately input to the same network, resulting in a pair of representation vectors.

- Projector Network, which maps representation to a vector $z = Proj(h)$. We instantiate $Proj$ as either a multi-layer MLP. We normalize the output of this network to lie on the unit hypersphere, which enables using an inner product to measure cos similarity.

- A contrastive loss layer on top of the Framework. It maximizes the agreement between one representation and its corresponding version that is augmented from the same token while keeping it distant from other token representations in the same batch.

For each input sentence, we first pass it to the data augmentation module, in which two transformations $Aug1$ and $Aug2$ are applied to generate two versions of token embeddings: $e_i = Aug1(x)$, $e_j = Aug2(x)$. After that, both e_i and $e_j$will be encoded by multi-layer transformer-based encoder-decoder blocks and Projector Network produce the contextualized representations z_i and z_j. During each training step, we randomly sample N sentences to construct a mini-batch, resulting in $2N$ representations after augmentation. Each data point is trained to find out its counterpart among in-batch samples B:

$$\mathcal{L}_{\text{scl}} = \sum_{p \in P} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in B} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)} \tag{3}$$

Here, $z = Proj(EncDec(x, y_{<t}))$, the $\bullet$ symbol denotes the inner (dot) product, τ is a scalar temperature parameter. The index i is called the anchor, $P \equiv \{p \in B : \tilde{\boldsymbol{y}}_p = \tilde{\boldsymbol{y}}_i\}$is the set of indices of all positives in the mini batch.

3 Experiments

To show the effectiveness of our method, experiments are conducted on both low-resource and rich-resource translation tasks.

3.1 Settings

To compare with Vaswani et al. (2017), we conducted our experiments on different scale datasets. The datasets of low-resource scenario are from IWSLT competitions, which include IWSLT14 English-German (En-De), English-Spanish (En-Es) and English-French (En-Fr) translations. The rich-resource datasets come from the widely acknowledged WMT translation tasks, and we take the WMT14 English-German tasks. The IWSLT datasets contain about 170k training sentence pairs. The WMT data size is 4.5M, and validation and test data are from the corresponding newstest data.

We applied joint Byte-Pair Encoding (BPE) (Sennrich et al., 2015) with 32k merging operations on WMT data sets and 10k merging operations on IWSLT data sets. We used a dropout of 0.3 for all IWSLT experiments except for the Transformer-base setting on the WMT En-De task which was 0.1. The temprature in supervised contrastive loss is set as 0.1 for all translation tasks.

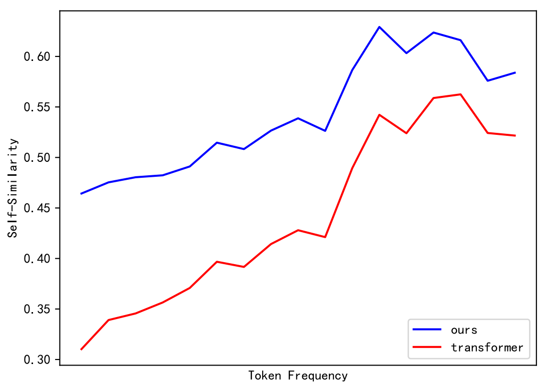

Figure 2: Self-Sim of our approach and transformer-base model. The X-axis is token frequency which drops gradually from left to right.

	BLEU	Self-Sim
full	27.62	0.421
low	22.88	0.393
+SCL	23.42	0.429

Table 2: BLEU scores and Self-Sim on WMT14.

BLEU	iwslt14 en-de
baseline	28.46
+ SCL	29.02
+ DA	29.20
+ SCL + DA	29.87

Table 3: Ablation study on IWSLT14 en-de dataset.

3.2 Analysis

The self-similarity of a word, is the average cosine similarity between its representations in different contexts. If the self-similarity is 1, then the representations are not context-specific at all; if the self-similarity is 0, that the representations are maximally context specific. In Figure 2, we plot the average self similarity of uniformly randomly sampled words, the higher the word frequency, the lower the self-similarity is on average. In other words, the higher the word frequency, the more context-specific the contextualized representations. But the lower the word frequency not have high self-similarity, implying that their contextualized representations are among the most context-specific.This is relatively surprising, given that these words are not polysemous. This finding suggests that the variety of contexts a word appears in, rather than its inherent polysemy, is what drives variation in its contextualized representations.

3.3 Main Results

We calculate the BLEU scores on these tasks for evaluation. The performances are shown in Table 1. We can see that our approach achieves more than 1.3 BLEU score improvements on IWSLT, which clearly shows the effectiveness of our method. In Table 2, we can see that the supervised contrastive learning enhances self-sim, and BLEU has also improved on WMT14 low data condition.

The efficacy of the data augmentation and the supervised positive sampling contrastive learning is evaluated. The variants are: the transformer baseline; DA, with additional word-dropout data augmentation; SCL, the contrastive learning using supervised positive sampling to optimize; and DA+SCL, trained with the addition of DA and SCL. The result is shown in Table 3.

From the result, it is clear that adding a contrastive objective can generally improve the recommendation performance compared with the baseline. Compared with DA+SCL, it can be concluded that the model-level dropout augmentation can provide a more semantically consistent unsupervised sample than the data-level augmentation. Furthermore, SCL relies on the target item to sample a semantically consistent supervised sample,which shows a large margin improvement over the unsupervised methods.

4 Related Work

Contrastive learning has become a very popular technique in unsupervised visual representation learning with solid performance. The main method is (Oord et al., 2018; He et al., 2020; Chen et al., 2020; Chen and He, 2020) encoding of different views of the same image as positive pairs. Contrastive learning also has been widely applied in language model pre-training task (Fang et al., 2020).

Recently, several approaches on contrastive learning for NMT have also been studied. Yang et al. (2019) proposed leveraging contrastive learning for reducing word omission errors. Pan et al. (2021) applied contrastive learning for multilingual MT. While these works have been conducted on sentence-level contrastive, we focus on extending contrastive learning on token-level NMT.

5 Conclusion and Future Work

In this work we propose a simple supervised contrastive framework for machine translation. We find that the variety of contexts a word appears in, rather than its inherent polysemy, is what drives variation in its contextualized representations. Meanwhile Our approach improves neural machine translation tasks with promising results. Future works should include a thorough study on better similarity measures and different data augmentation.

References

Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R. Manmatha, and Pietro Perona. 2020. Sequence-to-sequence contrastive learning for text recognition.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views.

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

Guy Bukchin, Eli Schwartz, Kate Saenko, Ori Shahar, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. 2021. Fine-grained angular contrastive learning with coarse labels.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2021. Unsupervised learning of visual features by contrasting cluster assignments.

Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. 2021. Wasserstein contrastive representation distillation.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xinlei Chen and Kaiming He. 2020. Exploring simple siamese representation learning.

Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. 2021. Tclr: Temporal contrastive learning for video representation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pretrained language model fine-tuning.

Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. 2021. Contrastive embedding for generalized zero-shot learning.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. 2020. Data-efficient image recognition with contrastive predictive coding.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. 2020. Robust pre-training by adversarial contrastive learning.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning.

Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2021. Self-guided contrastive learning for bert sentence representations.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, page 1–1.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. 2021. Semi-supervised action recognition with temporal contrastive learning.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020a. Contrastive representation distillation.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020b. What makes for good views for contrastive learning?

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Dong Wang, Ning Ding, Piji Li, and Hai-Tao Zheng. 2021a. Cline: Contrastive learning with semantic negative examples for natural language understanding.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2019. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*.

Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. 2021b. Improving weakly supervised visual grounding by contrastive knowledge distillation.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages.

Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. 2020. On mutual information in contrastive learning for visual representations.

Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. 2021. What should not be contrastive in contrastive learning.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. *arXiv preprint arXiv:2009.12061*.

Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Aijun Yang, Mingzhe Rong, and Xiaohua Wang. 2021. Complementary relation contrastive distillation.