ON LEARNING UNIVERSAL REPRESENTATIONS ACROSS LANGUAGES

Xiangpeng Wei[1,2], **Yue Hu**[1,2], **Rongxiang Weng**[3], **Luxi Xing**[1,2], **Heng Yu**[3], **Weihua Luo**[3]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{weixiangpeng,huyue,xingluxi}@iie.ac.cn
[3]Machine Intelligence Technology Lab, Alibaba Group, Hangzhou, China
{wengrx,yuheng.yh,weihua.luowh}@alibaba-inc.com

ABSTRACT

Recent studies have demonstrated the overwhelming advantage of cross-lingual pre-trained models (PTMs), such as multilingual BERT and XLM, on cross-lingual NLP tasks. However, existing approaches essentially capture the co-occurrence among tokens through involving the masked language model (MLM) objective with token-level cross entropy. In this work, we extend these approaches to learn sentence-level representations, and show the effectiveness on cross-lingual understanding and generation. We propose **Hi**erarchical **Con**tras**t**ive **L**earning (HICTL) to (1) learn universal representations for parallel sentences distributed in one or multiple languages and (2) distinguish the semantically-related words from a shared cross-lingual vocabulary for each sentence. We conduct evaluations on three benchmarks: language understanding tasks (QQP, QNLI, SST-2, MRPC, STS-B and MNLI) in the GLUE benchmark, cross-lingual natural language inference (XNLI) and machine translation. Experimental results show that the HICTL obtains an absolute gain of 1.0%/2.2% accuracy on GLUE/XNLI as well as achieves substantial improvements of +1.7$\sim$+3.6 BLEU on both the high-resource and low-resource English$\rightarrow$X translation tasks over strong baselines. We will release the source codes as soon as possible.

1 INTRODUCTION

Pre-trained models (PTMs) like ELMo (Peters et al., 2018), GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) have shown remarkable success of effectively transferring knowledge learned from large-scale unlabeled data to downstream NLP tasks, such as text classification (Socher et al., 2013) and natural language inference (Bowman et al., 2015; Williams et al., 2018), with limited or no training data. To extend such *pretraining-finetuning* paradigm to multiple languages, some endeavors such as multilingual BERT (Devlin et al., 2019) and XLM (Conneau & Lample, 2019) have been made for learning cross-lingual representation. More recently, Conneau et al. (2020) present XLM-R to study the effects of training unsupervised cross-lingual representations at a huge scale and demonstrate promising progresses on cross-lingual tasks.

However, all of these studies only perform masked language model (MLM) with token-level (i.e., *subword*) cross entropy, which limits PTMs to capture the co-occurrence among tokens and consequently fail to understand the whole sentence. It leads to two major shortcomings for current cross-lingual PTMs, i.e., *the acquisition of sentence-level representations* and *semantic alignments among parallel sentences in different languages*. Considering the former, Devlin et al. (2019) introduced the next sentence prediction (NSP) task to distinguish whether two input sentences are continuous segments from training corpus. However, this simple binary classification task is not enough to model sentence-level representations (Joshi et al., 2020; Yang et al., 2019; Liu et al., 2019; Lan et al., 2020; Conneau et al., 2020). For the latter, Huang et al. (2019) defined the cross-lingual paraphrase classification task, which concatenates two sentences from different languages as input and classifies whether they are with the same meaning. This task learns patterns of sentence-pairs well, but fails to distinguish the exact meaning of each sentence.

In response to these problems, we propose to strengthen PTMs through learning universal representations among semantically-equivalent sentences distributed in different languages. We introduce a novel **Hi**erarchical **Con**t**rastive** **L**earning (HICTL) framework to learn language invariant sentence representations via self-supervised non-parametric instance discrimination. Specifically, we use a BERT-style model to encode two sentences separately, and the representation of the first token (e.g., `[CLS]` in BERT) will be treated as the sentence representation. Then, we conduct instance-wise comparison at both sentence-level and word-level, which are complementary to each other. For the former, we maximize the similarity between two parallel sentences while minimize which among non-parallel ones. For the latter, we maintain a bag-of-words for each sentence-pair, each word in which is considered as a positive sample while the rest words in vocabulary are negative ones. To reduce the space of negative samples, we conduct negative sampling for word-level contrastive learning. With the HICTL framework, the PTMs are encouraged to learn language agnostic representation, thereby bridging the semantic discrepancy among cross-lingual sentences.[1]

The HICTL is conducted on the basis of XLM-$\text{R}_{\textbf{Base}}$ (Conneau et al., 2020) and experiments are performed on three benchmarks: language understanding tasks (QQP, QNLI, SST-2, MRPC, MNLI and STS-B) in the GLUE benchmark, cross-lingual natural language inference (XNLI) and machine translation. Extensive empirical evidences demonstrate that our approach can achieve consistent improvements over baselines on various tasks of both cross-lingual language understanding and generation. In more detail, our HICTL obtains absolute gains of 1.0%/1.1% accuracy on GLUE/XNLI over XLM-$\text{R}_{\textbf{Base}}$ and 2.2% accuracy on XNLI over XLM. For machine translation, our HICTL achieves substantial improvements over baselines on both low-resource (IWSLT English$\rightarrow$X) and high-resource (WMT English$\rightarrow$X) translation tasks.

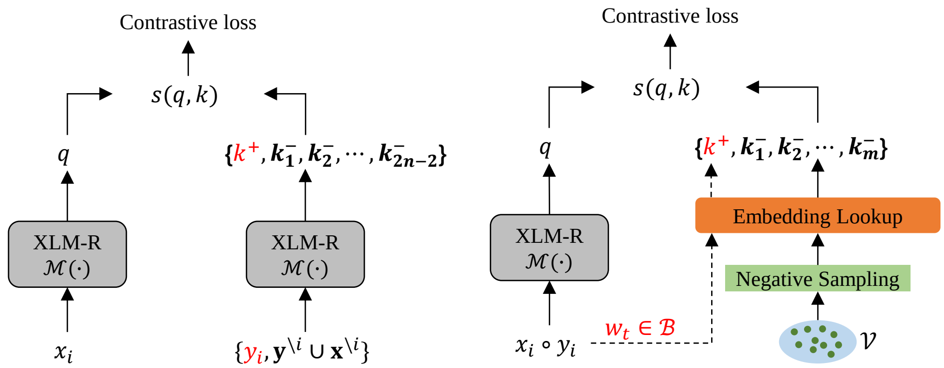
2 Related Work

Pre-trained Language Models. Recently, substantial work has shown that pre-trained models (PTMs) (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019) on the large corpus are benecial for downstream NLP tasks, like in GLUE (Wang et al., 2018) and XNLI (Conneau et al., 2018). The application scheme is to ne-tune the pre-trained model using the limited labeled data of specific target tasks. For cross-lingual pre-training, both Devlin et al. (2019) and Conneau & Lample (2019) trained a transformer-based model on multilingual Wikipedia which covers various languages, while XLM-R (Conneau et al., 2020) studied the effects of training unsupervised cross-lingual representations at a very large scale.

For sequence-to-sequence pre-training, UniLM (Dong et al., 2019) ne-tuned BERT with an ensemble of masks, which employs a shared Transformer network and utilizing specific self-attention mask to control what context the prediction conditions on. Song et al. (2019) extended BERT-style models by jointly training the encoder-decoder framework. XLNet (Yang et al., 2019) trained by predicting masked tokens auto-regressively in a permuted order, which allows predictions to condition on both left and right context. Raffel et al. (2019) unified every NLP problem as a text-to-text problem and pre-trained a denoising sequence-to-sequence model at scale. Concurrently, BART (Lewis et al., 2020) pre-trained a denoising sequence-to-sequence model, in which spans are masked from the input but the complete output is auto-regressively predicted.

Previous work have explored using pre-trained models to improve text generation, such as pre-training both the encoder and decoder on several languages (Song et al., 2019; Conneau & Lample, 2019; Raffel et al., 2019) or using pre-trained models to initialize encoders (Edunov et al., 2019; Zhang et al., 2019). Zhu et al. (2020) proposed a BERT-fused NMT model, in which the representations from BERT are treated as context and fed it into all layers of both the encoder and decoder, rather than served as input embeddings only. Zhong et al. (2020) formulated the extractive summarization task as a semantic text matching problem and proposed a Siamese-BERT architecture to compute the similarity between the source document and the candidate summary, which leverages the pre-trained BERT in a Siamese network structure. Our approach also belongs to the contextual pre-training so it could been applied to various downstream NLU and NLG tasks.

[1]The concurrent work (Feng et al., 2020; Chi et al., 2020) also conduct contrastive learning to produce similar representations across languages, but they only consider the sentence-level contrast. HICTL differs in learning to predict semantically-related words for each sentence additionally, which is particularly beneficial for cross-lingual text generation.



(a) Sentence-Level CTL (b) Word-Level CTL

Figure 1: Illustration of **Hi**erarchical **C**ontrastive **L**earning (HICTL).

Contrastive Learning. Contrastive learning (CTL) (Saunshi et al., 2019) aims at maximizing the similarity between the encoded query q and its matched key k^+ while keeping randomly sampled keys $\{k_0^-, k_1^-, k_2^-, ...\}$ faraway from it. With similarity measured by a score function $s(q, k)$, a form of a contrastive loss function, called InfoNCE (Oord et al., 2018), is considered in this paper:

$$\mathcal{L}_{ctl} = -\log \frac{\exp(s(q, k^+))}{\exp(s(q, k^+)) + \sum_i \exp(s(q, k_i^-))}, \tag{1}$$

where the score function $s(q, k)$ is essentially implemented as the cosine similarity $\frac{q^T k}{\|q\| \cdot \|k\|}$. q and k are often encoded by a learnable neural encoder, such as BERT (Devlin et al., 2019) or ResNet (He et al., 2016). k^+ and k^- are typically called positive and negative samples. In addition to the form illustrated in Eq. (1), contrastive losses can also be based on other forms, such as margin-based loses (Hadsell et al., 2006) and variants of NCE losses (Mnih & Kavukcuoglu, 2013).

Contrastive learning is at the core of several recent work on unsupervised or self-supervised learning from computer vision (Wu et al., 2018; Oord et al., 2018; Ye et al., 2019; Tian et al., 2019; He et al., 2019; Chen et al., 2020) to natural language processing (Mikolov et al., 2013; Mnih & Kavukcuoglu, 2013; Devlin et al., 2019; Clark et al., 2020). Kong et al. (2020) improved language representation learning by maximizing the mutual information between a masked sentence representation and local n-gram spans. Clark et al. (2020) utilized a discriminator to predict whether a token is replaced by a generator given its surrounding context. Iter et al. (2020) proposed to pre-train language models with contrastive sentence objectives that predicts the surrounding sentences given an anchor sentence. In this paper, we propose HICTL to encourage parallel cross-lingual sentences to have the identical semantic representation and distinguish whether a word is contained in them as well, which can naturally improve the capability of cross-lingual understanding and generation for PTMs.

3 METHODOLOGY

3.1 HIERARCHICAL CONTRASTIVE LEARNING

We propose hierarchical contrastive learning (HICTL), a novel comparison learning framework that unifies cross-lingual sentences as well as related words. HICTL can learn from both non-parallel and parallel multilingual data, and the overall architecture of HICTL is illustrated in Figure 1. We represent a *training batch* of the original sentences as $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ and its aligned counterpart is denoted as $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, where n is the batch size. For each pair $\langle x_i, y_i \rangle$, y_i is either the translation in the other language of x_i when using parallel data or the perturbation through reordering tokens in x_i when only monolingual data is available. $\mathbf{x}^{\backslash i}$ is denoted as a modified version of $\mathbf{x}$ where the i-th instance is removed.

Sentence-Level CTL. As illustrated in Figure 1a, we apply the XLM-R as the encoder to represent sentences into hidden representations. The first token of every sequence is always a special token (e.g., `[CLS]`), and the final hidden state corresponding to this token is used as the aggregate sentence representation for pre-training, that is, $r_x = f \circ g(\mathcal{M}(x))$ where $g(\cdot)$ is the aggregate function and $f(\cdot)$ is a linear projection, $\circ$ denotes the composition of operations. To obtain universal representation among semantically-equivalent sentences, we encourage r_{x_i} (the query, denoted as q) to be as similar as possible to r_{y_i} (the positive sample, denoted as k^+) but dissimilar to all other instances (i.e., $\mathbf{y}^{\setminus i} \cup \mathbf{x}^{\setminus i}$, considered as a series of negative samples, denoted as $\{k_1^-, k_2^-, ..., k_{2n-2}^-\}$) in a training batch. Formally, the sentence-level contrastive loss for x_i is defined as

$$\mathcal{L}_{sctl}(x_i) = -\log \frac{\exp \circ s(q, k^+)}{\exp \circ s(q, k^+) + \sum_{j=1}^{|\mathbf{y}^{\setminus i} \cup \mathbf{x}^{\setminus i}|} \exp \circ s(q, k_j^-)}. \tag{2}$$

Symmetrically, we also expect r_{y_i} (the query, denoted as $\hat{q}$) to be as similar as possible to r_{x_i} (the positive sample, denoted as $\hat{k}^+$) but dissimilar to all other instances in the same training batch, thus,

$$\mathcal{L}_{sctl}(y_i) = -\log \frac{\exp \circ s(\hat{q}, \hat{k}^+)}{\exp \circ s(\hat{q}, \hat{k}^+) + \sum_{j=1}^{|\mathbf{y}^{\setminus i} \cup \mathbf{x}^{\setminus i}|} \exp \circ s(\hat{q}, \hat{k}_j^-)}. \tag{3}$$

The sentence-level contrastive loss over the training batch can be formulated as

$$\mathcal{L}_S = \frac{1}{n} \sum_{i=1}^{n} \big\{ \mathcal{L}_{sctl}(x_i) + \mathcal{L}_{sctl}(y_i) \big\}. \tag{4}$$

Word-Level CTL. The motivations of introducing the word-level contrastive learning are in two folds. First, a sentence can be in several correct literals expressions and most of them share the similar bag-of-words (Ma et al., 2018). Thus, it is beneficial for sentence understanding by distinguishing its bag-of-words from the vocabulary. Second, there is a natural gap between the word embeddings of different languages. Intuitively, predicting the related words in other languages for each sentence can bridge the representations of words in different languages. As shown in Figure 1b, we concatenate the sentence pair $\langle x_i, y_i \rangle$ as $x_i \circ y_i$: `[CLS]` x_i `[SEP]` y_i `[SEP]` and the bag-of-words of which is denoted as $\mathcal{B}$. For word-level contrastive learning, the final state of the first token is treated as the query ($\tilde{q}$), each word $w_t \in \mathcal{B}$ is considered as the positive sample and all the other words ($\mathcal{V} \backslash \mathcal{B}$, i.e., the words in $\mathcal{V}$ that are not in $\mathcal{B}$ where $\mathcal{V}$ indicates the overall vocabulary of all languages) are negative samples. As the vocabulary usually with large space, we propose to only use a subset $\mathcal{S} \subset \mathcal{V} \backslash \mathcal{B}$ sampled according to the normalized similarities between $\tilde{q}$ and the embeddings of the words. As a result, the subset $\mathcal{S}$ naturally contains the hard negative samples which are beneficial for learning high quality representations (Ye et al., 2019). Specifically, the word-level contrastive loss for $\langle x_i, y_i \rangle$ is defined as

$$\mathcal{L}_{wctl}(x_i, y_i) = -\frac{1}{|\mathcal{B}|} \sum_{t=1}^{|\mathcal{B}|} \log \frac{\exp \circ s(\tilde{q}, e(w_t))}{\exp \circ s(\tilde{q}, e(w_t)) + \sum_{w_j \in \mathcal{S}} \exp \circ s(\tilde{q}, e(w_j))}. \tag{5}$$

where $e(\cdot)$ is the embedding lookup function and $|\mathcal{B}|$ is the number of unique words in the concatenated sequence $x_i \circ y_i$. The overall word-level conrastive loss can be formulated as:

$$\mathcal{L}_W = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{wctl}(x_i, y_i). \tag{6}$$

Multi-Task Pre-training. Both MLM and translation language model (TLM) are combined with HICTL by default, as the prior work (Conneau & Lample, 2019) have verified the effectiveness of them in XLM. In summary, the model can be optimized by minimizing the entire training loss:

$$\mathcal{L} = \mathcal{L}_{LM} + \mathcal{L}_S + \mathcal{L}_W, \tag{7}$$

where $\mathcal{L}_{LM}$ is implemented as either the TLM when using parallel data or the MLM when only monolingual data is available to recover the original words of masked positions given the contexts.

3.2 CROSS-LINGUAL FINE-TUNING

Sentence Classification. The representations produced by HICTL can be used in several ways for sentence classification tasks whether they involve single text or text pairs. The `[CLS]` representation of (1) single-sentence in sentiment analysis, (2) sentence pairs in paraphrasing and entailment is fed into an extra output-layer for classification.

Table 1: Statistics (#millions) of the training data used in pre-training.

LANG.	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh
MONO.	3.8	1.5	17.4	1.3	43.2	11.3	15.5	0.6	12.6	0.2	0.8	1.8	0.5	3.8	5.5
PARA.	9.8	0.6	9.3	4.0	-	11.4	13.2	1.6	11.7	0.2	3.3	0.5	0.7	3.5	9.6

Machine Translation. We also explore using HICTL to improve machine translation. In the previous work, Conneau & Lample (2019) has shown that the pre-trained encoders can provide a better initialization of supervised and unsupervised neural machine translation (NMT) systems. Liu et al. (2020) has shown that NMT models can be improved by incorporating pre-trained sequence-to-sequence models on various language pairs but highest-resource settings. As illustrated in Figure 2, we use the model pre-trained by HICTL as the encoder, and add a new set of decoder parameters that are learned from scratch. To prevent pre-trained weights from being washed out by supervised training, we train the encoder-decoder model in two steps. At the first step, we freeze the pre-trained encoder and only update the decoder. At the second step, we train all model parameters for a relatively small number of iterations. In both cases, we compute the similarities between the `[CLS]` representation of the encoder and all target words in advance. Then we aggregate them with the logits before softmax of each decoder step through an element-wise additive operation. The encoder-decoder model is optimized by maximizing the log-likelihood of bitext at both steps.

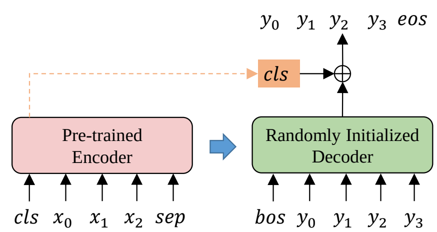


Figure 2: Fine-tuning on NMT task.

4 EXPERIMENTS

We consider three evaluation benchmarks: the GLUE benchmark (QQP, QNLI, SST-2, MRPC, MNLI and STS-B), cross-lingual natural language inference (XNLI) and machine translation (low-resource tasks: IWSLT English$\leftrightarrow$\{German, French, Chinese\}, high-resource tasks: WMT'14 English$\rightarrow$\{German, French\} and WMT'18 English$\rightarrow$\{Chinese, Russian\}). Next, we first describe the data and training details, and then compare the HICTL with the existing state-of-the-art models.

4.1 DATA AND MODEL

Our model is pre-trained on 15 languages, including English (en), French (fr), Spanish (es), German (de), Greek (el), Bulgarian (bg), Russian (ru), Turkish (tr), Arabic (ar), Vietnamese (vi), Thai (th), Chinese (zh), Hindi (hi), Swahili (sw) and Urdu (ur). For monolingual data, we use the Wikipedia from these languages. For bilingual data, we use the same (*English-to-X*) MT dataset as (Conneau & Lample, 2019), which are collected from MultiUN (Eisele & Yu, 2010) for French, Spanish, Arabic and Chinese, the IIT Bombay corpus (Kunchukuttan et al., 2018) for Hindi, the OpenSubtitles 2018 for Turkish, Vietnamese and Thai, the EUbookshop corpus for German, Greek and Bulgarian, Tanzil for both Urdu and Swahili, and GlobalVoices for Swahili. Table 1 lists the statistics.

We adopt the TRANSFORMER-ENCODER (Vaswani et al., 2017) as the backbone with 12 identical layers, 768 hidden units, 12 heads and GeLU activation (Hendrycks & Gimpel, 2016). To reduce pre-training complexity, we initialize our model from XLM-R (Conneau et al., 2020) with the `Base` setting. During pre-training, a training batch for HICTL covers 15 languages with equal probability, each instance with two sentences as input and the max sequence length is set to 128. We use Adam optimizer to train the model, and learning rate starts from $2.5e-5$ with invert linear decay. We run the pre-training experiments on 8 V100 GPUs with update frequency of 16, batch size 1024. The number of negative samples $m = 512$ for word-level contrastive learning.

4.2 EXPERIMENTAL EVALUATION

Cross-lingual Natural Language Inference (XNLI) The XNLI (Conneau et al., 2018) dataset extends the development and test sets of the Multi-Genre Natural Language Inference (MultiNLI)

Table 2: Results on Cross-lingual Natural Language Inference (XNLI). We report the accuracy on each of the 15 XNLI languages and the average accuracy of our HICTL as well as five baselines: BiLSTM (Conneau et al., 2018), mBERT (Devlin et al., 2019), XLM (Conneau & Lample, 2019), Unicoder (Huang et al., 2019) and XLM-R (Conneau et al., 2020).

MODEL	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Evaluation of cross-lingual sentence encoders (CROSS-LINGUAL TEST)																
BiLSTM	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
mBERT	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
XLM	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Unicoder	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
XLM-$R_{\mathbf{Base}}$	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
HICTL	**86.3**	**80.5**	**81.3**	**79.5**	**78.9**	**80.6**	**79.0**	**75.4**	**74.8**	**77.4**	**75.7**	**77.6**	**73.1**	**69.9**	**69.7**	**77.3**
Machine translate at test (TRANSLATE-TEST)																
BiLSTM	73.7	70.4	70.7	68.7	69.1	70.4	67.8	66.3	66.8	66.5	64.4	68.3	64.2	61.8	59.3	67.2
mBERT	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Unicoder	85.1	80.1	80.3	78.2	77.5	78.0	76.2	73.3	73.9	72.8	71.6	74.1	70.3	65.2	66.3	74.9
HICTL	**85.9**	**81.6**	**81.9**	**79.1**	**80.8**	**80.2**	**79.7**	**78.4**	**76.8**	**77.6**	**76.2**	**76.7**	**73.2**	**69.4**	**72.6**	**78.0**
Machine translate at training (TRANSLATE-TRAIN)																
BiLSTM	73.7	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6	65.4
mBERT	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Unicoder	85.1	80.0	81.1	79.9	77.7	80.2	77.9	75.3	76.7	76.4	75.2	79.4	71.8	71.8	64.5	76.9
HICTL	**85.7**	**81.3**	**82.1**	**80.2**	**81.4**	**81.0**	**80.5**	**79.7**	**77.4**	**78.2**	**77.5**	**80.2**	**75.4**	**73.5**	**72.9**	**79.1**
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																
XLM	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
Unicoder	85.6	81.1	82.3	80.9	79.5	81.4	79.7	76.8	78.2	77.9	77.1	80.5	73.4	73.8	69.6	78.5
XLM-$R_{\mathbf{Base}}$	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
HICTL	**86.5**	**82.3**	**83.2**	**80.8**	**81.6**	**82.2**	**81.3**	**80.5**	**78.1**	**80.4**	**78.6**	**80.7**	**76.7**	**73.8**	**73.9**	**80.0**

corpus (Williams et al., 2018) to 15 languages, and comes with a ground-truth English training set. The training set has been machine-translated to the remaining 14 languages[2], providing synthetic training data for these languages as well. We evaluate our model on cross-lingual transfer from English to other languages (denoted as CROSS-LINGUAL TEST). It means that the pre-trained model is fine-tuned on English MultiNLI, and then evaluated on the foreign language XNLI test. In addition, we also consider three machine translation baselines: (1) TRANSLATE-TEST: the developments and test sets are machine-translated to English and a single English model is used (2) TRANSLATE-TRAIN: we fine-tune a multiligual model using the training set machine-translated from English for each language (3) TRANSLATE-TRAIN-ALL: we fine-tune a multilingual model on the concatenation of all training sets from TRANSLATE-TRAIN. For fine-tuning stage, we use the same optimizer as pre-training and learning rate set as $2.5e-5$. We set the batch size to 32.

Table 2 shows XNLI results, by comparing our HICTL model with five baselines: Conneau et al. (2018) uses BiLSTM as sentence encoder and constraints bilingual sentence pairs have similar embedding. Multilingual BERT (mBERT for short) (Devlin et al., 2019) and XLM (Conneau & Lample, 2019) extend PTMs to cross-lingual pre-training. Unicoder (Huang et al., 2019) propose an universal language encoder and pre-train it using multiple tasks. XLM-R (Conneau et al., 2020) is to study the effects of training unsupervised cross-lingual representations at a very large scale.

HICTL achieves remarkable results in all fine-tuning settings. On cross-lingual transfer (CROSS-LINGUAL TEST), HICTL obtains 77.3% accuracy, outperforming the state-of-the-art XLM-$R_{\mathbf{Base}}$, Unicoder and XLM models by 1.1%, 1.9% and 2.2% average accuracy. Compared to mBERT, HICTL obtains substantial gains of 11.4%, 12.7% and 13.8% on Urdu, Arabic and Chinese respectively. Using the multilingual pre-training of TRANSLATE-TRAIN-ALL, HICTL further improves performance and reaches 80.0% accuracy, outperforming XLM-$R_{\mathbf{Base}}$ and Unicoder by 0.9% and 1.5% average accuracy respectively.

GLUE We evaluate the English performance of our model on the GLUE benchmark (Wang et al., 2019a) which gathers multiple classication tasks, including single-sentence tasks (e.g., SST-2 (Socher et al., 2013)), similarity and paraphrase tasks (MRPC (Dolan & Brockett, 2005), STS-B (Cer et al., 2017), QQP[3]), as well as inference tasks (e.g., MNLI (Williams et al., 2018),

[2] `https://dl.fbaipublicfiles.com/XNLI/XNLI-MT-1.0.zip`

[3] `https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs`

Table 3: **GLUE dev results.** Results with † and ‡ are from (Liu et al., 2019) and our in-house implementation respectively. We compare the performance of HICTL to BERT$_{\mathbf{Large}}$, XLNet$_{\mathbf{Large}}$ (Yang et al., 2019), RoBERTa (Liu et al., 2019) and XLM-R$_{\mathbf{Large}}$ on the English GLUE benchmark.

MODEL	#PARAMS	QQP	QNLI	SST-2	MRPC	MNLI	STS-B	Avg
BERT$^{\dagger}_{\mathbf{Large}}$	340M	91.3	92.3	93.2	88.0	86.6	90.0	90.2
XLM-R$_{\mathbf{Large}}$	550M	**92.3**	93.8	95.0	89.5	88.9	91.2	91.8
XLNet$^{\dagger}_{\mathbf{Large}}$	-	91.8	93.9	95.6	89.2	89.8	91.8	92.0
RoBERTa†	355M	92.2	**94.7**	**96.4**	**90.9**	**90.2**	**92.4**	**92.8**
XLM-R$^{\ddagger}_{\mathbf{Base}}$	270M	91.1	93.0	94.1	89.3	87.5	90.4	90.9
HICTL	270M	**92.1**	**93.8**	**95.4**	**90.0**	**89.2**	**91.1**	**91.9**

QNLI (Rajpurkar et al., 2016)). We use XLM-R$_{\mathbf{Base}}$ as baseline and also make comparisons with the state-of-the-art BERT$_{\mathbf{Large}}$, XLNet$_{\mathbf{Large}}$, RoBERTa and XLM-R$_{\mathbf{Large}}$.

Table 3 shows the performance of various models on the GLUE benchmark. We can observe that HICTL achieves significantly improvements, by 1.0% accuracy on average, over XLM-R$_{\mathbf{Base}}$ on all tasks. Note that HICTL also outperforms XLM-R$_{\mathbf{Large}}$ and reaches performance on par with XLNet$_{\mathbf{Large}}$ with less parameters. These results demonstrate the effectiveness of learning universal representations at both sentence and word level for many languages, which maintains strong capabilities on per-language downstream tasks as well.

Machine Translation The main idea of HICTL is to summarize cross-lingual parallel sentences into a shared representation that we term as semantic embedding, using which semantically related words can be distinguished from others. Thus it is natural to apply this global embedding to text generation. To that end, we fine-tune the pre-trained HICTL on machine translation tasks with both low-resource and high-resource settings. For the low-resource scenario, we choose IWSLT'14 English↔German (En↔De)[4], IWSLT'17 English→French (En→Fr) and English→Chinese (En→Zh) translation[5]. There are 160k, 183k, 236k, 235k bilingual sentence pairs for En↔De, En→Fr and En→Zh tasks. For the rich-resource scenario, we work on WMT'14 En→{De, Fr} and WMT'18 En→{Zh, Ru}. For WMT'14 En→{De, Fr}, the corpus sizes are 4.5M and 36M respectively, and we concatenate newstest2012 and newstest2013 as the validation set and use newstest2014 as the test set. For WMT'18 En→{Zh, Ru}, there are 24M and 8M sentence pairs respectively, we select the best models on newstest2017 and report results on newstest2018.

During fine-tuning, we use the pre-trained model to initialize the encoder and introduce a randomly initialized decoder. Previous work have verified the use of deep encoders and shallow decoders to improve translation speed (Kim et al., 2019; Kasai et al., 2020) and accuracy (Miceli Barone et al., 2017; Wang et al., 2019b). Thus we develop a shallower decoder with 4 (768 hidden units, 12 heads) identical layers to reduce the computation overhead. The number of hidden units and heads are same as the encoder, i.e. 768 and 12 respectively. At the first fine-tune step, we concatenate the datasets of all language pairs in either low-resource or high-resource setting to optimize the decoder only until convergence. Then we tune the whole encoder-decoder model using per-language corpus at the second step. The initial learning rate is $2e-5$ and `inverse_sqrt` learning rate (Vaswani et al., 2017) scheduler is also adopted. For WMT14 En→De, we use beam search with width 4 and length penalty 0.6 for inference. For other tasks, we use width 5 and length penalty 1.0. We use `multi-bleu.perl` to evaluate IWSLT'14 En↔De and WMT tasks, but `sacreBLEU` for the remaining tasks, for fair comparison with previous work.

Results are reported in Table 4. We implemented standard Transformer (apply the `base` and `big` setting for IWSLT and WMT tasks respectively) as baseline. The proposed HICTL can improve the BLEU scores of the eight tasks by 1.93, 1.76, 2.7, 2.2, 1.43, 1.25, 1.87 and 1.67. As the *task-adaptive pre-training* (TAPT) (Gururangan et al., 2020) can be applied to our HICTL with minimal modifications, thus we introduce a second phase of pre-training for HICTL on IWSLT or WMT parallel corpora (denoted as HICTL*), which can obtain additional gains of 0.7 BLEU on average.

[4]We split 7k sentence pairs from the training dataset for validation and concatenate dev2010, dev2012, tst2010, tst2011, tst2012 as the test set.

[5]`https://wit3.fbk.eu/mt.php?release=2017-01-ted-test`

Table 4: **BLEU scores [%].** We conduct experiments with both low-resource and high-resource settings. Two bert-fused NMT models (Yang et al., 2020; Zhu et al., 2020) are considered as our baselines. Following (Gururangan et al., 2020), we also adopt pre-trained language models to downstream tasks by introducing a second phase of pre-training for HICTL on IWSLT or WMT parallel corpora, which denoted as **HICTL***. Results with ‡ are from our in-house implementation.

MODEL	LOW-RESOURCE				HIGH-RESOURCE			
	IWSLT'14		IWSLT'17		WMT'14		WMT'18	
	En→De	De→En	En→Fr	En→Zh	En→De	En→Fr	En→Zh	En→Ru
Vaswani et al. (2017)	-	-	-	-	28.4	41.0	-	-
Yang et al. (2020)	-	-	-	-	30.1	42.3	-	-
Zhu et al. (2020)	30.45	36.11	38.7	28.2	30.75	**43.78**	-	-
TRANSFORMER‡	28.64	34.51	35.8	26.5	28.86	41.62	34.22	30.26
HICTL	30.57	36.27	38.5	28.7	30.29	42.87	36.09	31.93
HICTL*	**31.36**	**37.12**	**39.4**	**29.5**	**30.86**	43.31	**36.64**	**32.57**

Our approach also outperforms the BERT-fused model (Yang et al., 2020), a method treats BERT as an extra context and fuses the representations extracted from BERT with each encoder and decoder layer. Note we achieve new state-of-the-art results on IWSLT'14 En→De, IWSLT'17 En→{Fr, Zh} translations. These improvements show that mapping different languages into an universal representation space by using our HICTL to bridge the semantic discrepancy among cross-lingual sentence, is beneficial for both low-resource and high-resource translations.

5 CONCLUSION

We have demonstrated that pre-trained language models (PTMs) trained to learn commonsense knowledge from large-scale unlabeled data highly benefit from hierarchical contrastive learning (HICTL), both in terms of cross-lingual language understanding and generation. Learning universal representations at both word-level and sentence-level bridges the semantic discrepancy across languages. As a result, our HICTL sets a new level of performance among cross-lingual PTMs, improving on the state of the art by a large margin. We have also presented that by combing our method with task-adaptive pre-training, the better results cab be obtained. Even rich-resource languages also have been improved.

ACKNOWLEDGMENTS

We would like to thank Jing Yu for the instructive suggestions and invaluable help. Yue Hu is the corresponding author. This work is supported by the National Key Research and Development Programs under Grant No. 2017YFB0803301 and No. 2016YFB0801003, and work done at Alibaba Group.

REFERENCES

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://www.aclweb.org/anthology/D15-1075`.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055, 2017. URL `http://arxiv.org/abs/1708.00055`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of Machine Learning and Systems 2020*, pp. 10719–10729, 2020.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *CoRR*, abs/2007.07834, 2020. URL https://arxiv.org/abs/2007.07834.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Proc. of NIPS 2019*, pp. 7059–7069, 2019. URL http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.pdf.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://www.aclweb.org/anthology/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://www.aclweb.org/anthology/I05-5002.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32, NeurIPS 2019*, pp. 13063–13075. Curran Associates, Inc., 2019.

Sergey Edunov, Alexei Baevski, and Michael Auli. Pre-trained language model representations for language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4052–4059, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1409. URL https://www.aclweb.org/anthology/N19-1409.

Andreas Eisele and Chen Yu. Multiun: A multilingual corpus from united nation documents. In *International Conference on Language Resources & Evaluation*, 2010.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852, 2020. URL https://arxiv.org/abs/2007.01852.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360. Association for Computational Linguistics, 2020. URL https://www.aclweb.org/anthology/2020.acl-main.740.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, pp. 1735–1742. IEEE Computer Society, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 770–778. IEEE Computer Society, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL `http://arxiv.org/abs/1606.08415`.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2485–2494, Hong Kong, China, November 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D19-1252`.

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4859–4870, Online, 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.acl-main.439`.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77, 2020. URL `https://transacl.org/ojs/index.php/tacl/article/view/1853`.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A. Smith. Deep encoder, shallow decoder: Reevaluating the speed-quality trade-off in machine translation. *CoRR*, abs/2006.10369, 2020. URL `https://arxiv.org/abs/2006.10369`.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pp. 280–288, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5632. URL `https://www.aclweb.org/anthology/D19-5632`.

Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. A mutual information maximization perspective of language representation learning. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT bombay english-hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*. European Language Resources Association (ELRA), 2018.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=H1eA7AEtvS`.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880. Association for Computational Linguistics, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210, 2020.

Shuming Ma, Xu Sun, Yizhong Wang, and Junyang Lin. Bag-of-words as target for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 332–338, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2053. URL `https://www.aclweb.org/anthology/P18-2053`.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pp. 99–107, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4710. URL `https://www.aclweb.org/anthology/W17-4710`.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, 2013.

Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems 26*, pp. 2265–2273, 2013.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL `http://arxiv.org/abs/1807.03748`.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://www.aclweb.org/anthology/N18-1202`.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://www.aclweb.org/anthology/D16-1264`.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5628–5637, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/saunshi19a.html`.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D13-1170`.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97, pp. 5926–5936. PMLR, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019. URL `http://arxiv.org/abs/1906.05849`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30, NIPS 2017*, pp. 5998–6008. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://www.aclweb.org/anthology/W18-5446`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019a. URL `https://openreview.net/forum?id=rJ4km2R5t7`.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1810–1822, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1176. URL `https://www.aclweb.org/anthology/P19-1176`.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL `https://www.aclweb.org/anthology/N18-1101`.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 3733–3742. IEEE Computer Society, 2018.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. Towards making the most of BERT in neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pp. 9378–9385. AAAI Press, 2020. URL `https://aaai.org/ojs/index.php/AAAI/article/view/6479`.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32, NeurIPS 2019*, pp. 5753–5763. Curran Associates, Inc., 2019.

Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pp. 6210–6219. Computer Vision Foundation / IEEE, 2019.

Xingxing Zhang, Furu Wei, and Ming Zhou. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5059–5069, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1499. URL `https://www.aclweb.org/anthology/P19-1499`.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6197–6208. Association for Computational Linguistics, 2020. URL https://www.aclweb.org/anthology/2020.acl-main.552.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating BERT into neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Hyl7ygStwB.