VECO: VARIABLE ENCODER-DECODER PRE-TRAINING FOR CROSS-LINGUAL UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies about learning multilingual representations have achieved significant performance gains across a wide range of downstream cross-lingual tasks. They train either an encoder-only Transformer mainly for understanding tasks, or an encoder-decoder Transformer specifically for generation tasks, ignoring the correlation between the two tasks and frameworks. In contrast, this paper presents a variable encoder-decoder (VECO) pre-training approach to unify the two mainstreams in both model architectures and pre-training tasks. VECO splits the standard Transformer block into several sub-modules trained with both inner-sequence and cross-sequence masked language modeling, and correspondingly reorganizes certain sub-modules for understanding and generation tasks during inference. Such a workflow not only ensures to train the most streamlined parameters necessary for two kinds of tasks, but also enables them to boost each other via sharing common sub-modules. As a result, VECO delivers new state-of-the-art results on various cross-lingual understanding tasks of the XTREME benchmark covering text classification, sequence labeling, question answering, and sentence retrieval. For generation tasks, VECO also outperforms all existing cross-lingual models and state-of-the-art Transformer variants on WMT14 English-to-German and English-to-French translation datasets, with gains of up to 1∼2 BLEU.

1 INTRODUCTION

Driven by the striking success of pre-trained language models (Devlin et al., 2019), recent cross-lingual pre-training (Lample & Conneau, 2019; Liu et al., 2020b) has attracted an increasing of attention. It provides cross-lingual contextualized representations for the inputs of different languages, which significantly advances performance in both natural language understanding (NLU) and generation (NLG) tasks.

There are two mainstream architectures in current cross-lingual pre-training literature: *encoder-only* and *encoder-decoder*. The former like XLM (Lample & Conneau, 2019) focuses on conducting masked language modeling (MLM) with a single Transformer (Vaswani et al., 2017) encoder. This paradigm is naturally compatible with various NLU tasks, but tends to suffer from limited gains on cross-lingual generation tasks (e.g., machine translation) due to the lack of effective decoder initialization. In contrast, the latter like mBART (Liu et al., 2020b) attempts to pre-train the encoder-decoder Transformer via denoising auto-encoding tasks to provide complete initialization for downstream generation tasks. However, when applied in NLU scenarios, it usually requires more computation and memory to match the performance of the encoder-only models.

In light of the above pros and cons, this work presents Variable Encoder-deCOder (VECO) pre-training, which targets at providing pre-trained model initialization for both the *encoder-only* and *encoder-decoder* Transformer with the most streamlined parameters. We observe that Transformer encoder and decoder blocks have two common modules: `SelfAttention` and `FFN` (feed-forward network), with the main difference that the latter introduces an extra `CrossAttention` (attention across from the encoder to the decoder) module. Inspired by the lottery ticket hypothesis (Frankle & Carbin, 2018), we split the standard Transformer block into three independent modules $\{\texttt{SelfAttention}, \texttt{CrossAttention}, \texttt{FFN}\}$ to be collaboratively trained via two specific MLM

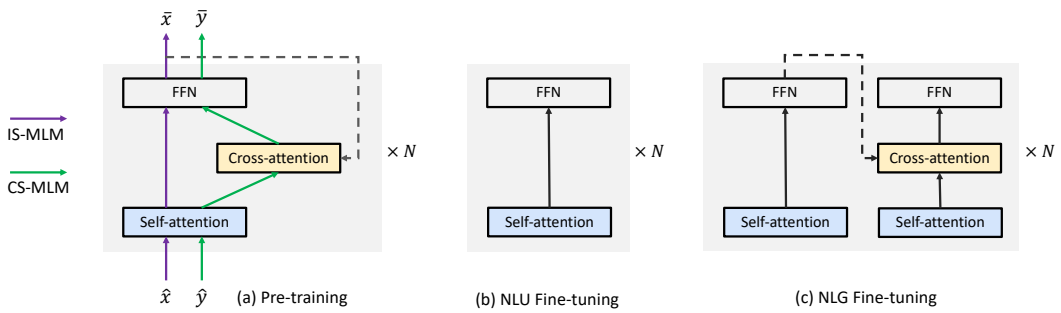

Figure 1: The overview of VECO. During pre-training, we feed two masked segments $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ into different modules to perform inner-sentence mask language modeling (IS-MLM) and cross-sentence mask language modeling (CS-MLM). More specifically, the masked segment $\hat{\boldsymbol{x}}$ can only attend to its context via self-attention to recover the original tokens $\bar{\boldsymbol{x}}$ (IS-MLM), while masked segment $\hat{\boldsymbol{y}}$ can attend to its preceding tokens via self-attention and the context $\hat{\boldsymbol{x}}$ via cross-attention to predict the original tokens $\bar{\boldsymbol{y}}$ (CS-MLM). For downstream NLU tasks, we throw out the cross-attention module and only fine-tune on the self-attention and FFN modules acted as an encoder. For NLG tasks, we keep all modules to initialize the corresponding encoder and decoders.

tasks. After that, we rebuild the desired complete architecture applicable for NLU or NLG with different specific combinations of these modules during fine-tuning.[1] Specifically, to be equipped with the ability of language understanding during pre-training, `SelfAttention` and `FFN` are assembled into a standard Transformer encoder for conducting inner-sequence masked language modeling (IS-MLM). In terms of generation, `SelfAttention`, `CrossAttention`, and `FFN` act together as the decoder in the standard sequence-to-sequence model, and are trained by the elaborately designed cross-sequence masked language modeling (CS-MLM) task. When applied to downstream fine-tuning, both `SelfAttention` and `FFN` modules constitute the Transformer encoder for contextual modeling in NLU or NLG, or cooperate with additional `CrossAttention` to provide the effective initialization of Transformer decoder. With such kind of workflow, VECO can be applied to both NLU and NLG tasks with the most streamlined parameters, which significantly reduces computational overhead and memory costs. Moreover, IS-MLM is specifically designed for understanding of individual sequences, while both understanding and generation tasks can benefit from CS-MLM. With such parameter sharing, VECO enables `SelfAttention` and `FFN` modules to be jointly trained by the two MLMs, which boosts both NLU and NLG performance.

We validate VECO on a variety of representative cross-lingual NLU and NLG benchmarks. For cross-lingual understanding tasks, we conduct experiments on the XTREME benchmark consisting of 9 cross-lingual tasks, including text classification, sequence labeling, question answering, and sentence retrieval. VECO ranks first at the XTREME leaderboard[2] at the submission deadline and obtains new state-of-the-art results on most of the tasks. For cross-lingual generation tasks, we validate VECO on the widely used WMT14 English-German and English-French machine translation benchmarks. VECO obtains 44.4 and 31.5 BLEU scores, consistently outperforming existing cross-lingual pre-training approaches and state-of-the-art Transformer variants by around 1∼2 BLEU.

2 VARIABLE ENCODER-DECODER PRE-TRAINING

2.1 BACKBONE NETWORK

The backbone network of VECO is composed of a stack of N identical layers. Each layer has three modules, consisting of a *required* self-attention module, an *optional* cross-attention module, and a *required* feed-forward linear module. Both self-attention and cross-attention modules are based on the multi-head attention (Vaswani et al., 2017):

$$\begin{aligned}\texttt{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \texttt{Concat}(\mathrm{head}_1,...,\mathrm{head}_\mathrm{h})\mathbf{W}^{\mathbf{O}} \\ \mathrm{head}_\mathrm{i} &= \texttt{Attention}(\mathbf{Q}\mathbf{W}^{\mathbf{Q}},\mathbf{K}\mathbf{W}^{\mathbf{K}},\mathbf{V}\mathbf{W}^{\mathbf{V}})\end{aligned} \tag{1}$$

[1]Thus the word *variable* means that the backbone Transformer varies during pre-training and fine-tuning.
[2]`https://sites.research.google/xtreme`

where $\mathbf{W^O}, \mathbf{W^Q}, \mathbf{W^K}$ and $\mathbf{W^V}$ are parameter matrices. $\texttt{Attention}(a, b, c)$ represents the attention operation with a as query, b as key, and c as value. We refer the readers to Vaswani et al. (2017) for more details.

The main difference of the self-attention module and cross-attention module is that $\mathbf{Q} = \mathbf{K} = \mathbf{V}$ holds in the self-attention module while only $\mathbf{K} = \mathbf{V}$ exists in the cross-attention module. We formalize these two modules as:

$$\begin{aligned} \texttt{SelfAttention}(\mathbf{x}; \theta_\mathbf{s}) &= \texttt{AddNorm}\big(\texttt{MultiHead}(\mathbf{Q} = \mathbf{x}, \mathbf{K} = \mathbf{x}, \mathbf{V} = \mathbf{x})\big) \\ \texttt{CrossAttention}(\mathbf{x}, \mathbf{y}; \theta_\mathbf{c}) &= \texttt{AddNorm}\big(\texttt{MultiHead}(\mathbf{Q} = \mathbf{x}, \mathbf{K} = \mathbf{y}, \mathbf{V} = \mathbf{y})\big) \end{aligned} \tag{2}$$

where $\theta_\mathbf{s}$ and $\theta_\mathbf{c}$ are the corresponding parameters, and $\texttt{AddNorm}$ denotes a residual connection (He et al., 2016) with a post layer normalization (Ba et al., 2016).

After that, a fully connected feed-forward network is applied to each element of input independently:

$$\texttt{FFN}(\mathbf{x}; \theta_\mathbf{f}) = \texttt{AddNorm}\big(\mathbf{W_2}\ \texttt{GeLU}\left(\mathbf{W_1 x}\right)\big) \tag{3}$$

where $\theta_\mathbf{f} = \{\mathbf{W_1}; \mathbf{W_2}\}$ are matrices of parameters.

2.2 Pre-training Objectives

In cross-lingual pre-training scenarios, we can utilize both monolingual and bilingual data widely used in previous works (Lample & Conneau, 2019; Chi et al., 2020b; Yang et al., 2020). We formalize both the two adjacent segments in the monolingual corpus and a pair of parallel sentence in the bilingual corpus as $(\boldsymbol{x}, \boldsymbol{y})$. We firstly adopt the same mask strategy like BERT (Devlin et al., 2019) to construct the masked input $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$. Then, the backbone Transformer takes the input to perform the inner-sequence and cross-sequence masked language modeling, which enables the model to be optimized jointly for cross-lingual language understanding and cross-sequence generation.

IS-MLM: Inner-Sequence Masked Language Modeling To be equipped with the ability of language understanding, we perform masked language modeling of a single-sequence on the self-attention and FFN modules, while skipping the cross-attention modules. As shown in Figure 1, the purple lines show the forward process of the IS-MLM task in each layer. To be specific, the embeddings of the masked input sequence $\hat{\boldsymbol{x}}$ are fed into the self-attention and FFN modules in each layer to get a contextual presentation $\mathbf{X}^{(i)}$:

$$\begin{aligned} \mathbf{H}^{(i)} &= \texttt{SelfAttention}(\mathbf{X}^{(i-1)}; \theta_s) \\ \mathbf{X}^{(i)} &= \texttt{FFN}(\mathbf{H}^{(i)}; \theta_f) \end{aligned} \tag{4}$$

The contextual presentations $\mathbf{X}^{(N)}$ of the last layer is used to recover the masked tokens $\bar{\boldsymbol{x}}$. Thus the training loss of inner-sequence masked language modeling can be formalized as:

$$\mathcal{L}_{\mathrm{IS-MLM}}(\boldsymbol{x}) = -\log P(\bar{\boldsymbol{x}} | \hat{\boldsymbol{x}}; \theta_s, \theta_f) \tag{5}$$

CS-MLM: Cross-Sequence Masked Language Modeling In order to fully train the cross-attention module that plays a primary role in semantic mapping between sentences in cross-lingual generation tasks (e.g., machine translation), we simulate a decoder by reusing the self-attention and FFN modules to corporate with the cross-attention module. As shown in Figure 1, the green lines depict the forward process of the CS-MLM task in each layer. Specifically, we first extract the contextual representation of $\hat{\boldsymbol{y}}$ via the $\texttt{SelfAttention}$ module[3]; then a cross-attention module is employed to model an interactive representation of $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$:

$$\begin{aligned} \mathbf{S}^{(i)} &= \texttt{SelfAttention}(\mathbf{Y}^{(i-1)}; \theta_s) \\ \mathbf{Z}^{(i)} &= \texttt{CrossAttention}(\mathbf{S}^{(i)}, \mathbf{X}^{(N)}; \theta_c) \\ \mathbf{Y}^{(i)} &= \texttt{FFN}(\mathbf{Z}^{(i)}; \theta_f) \end{aligned} \tag{6}$$

Finally, $\mathbf{Y}^{(N)}$, considering both the context of the semantic-related sequence $\hat{\boldsymbol{x}}$ and its left segments $\hat{\boldsymbol{y}}_{<t}$, is used to predict the masked tokens $\bar{\boldsymbol{y}}$:

$$\mathcal{L}_{\mathrm{CS-MLM}}(\boldsymbol{x}, \boldsymbol{y}) = -\log P(\bar{\boldsymbol{y}} | \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}_{<t}; \theta_s, \theta_c, \theta_f) \tag{7}$$

[3] A triangular attention mask matrix is used to only attend to the preceding tokens of each masked token.

Note that when optimizing the CS-MLM objective, we detach $\mathbf{X}^{(N)}$ in Eq. (6) from the computation graph to let the two objectives optimized in isolation. It also stabilizes the training of this "virtual" decoder model, since very deep encoder-decoders are typically hard to train.

Several monolingual pre-training models such as MASS (Song et al., 2019), BART (Lewis et al., 2019) and PALM (Bi et al., 2020) also present similar auto-regressive language modeling tasks. The core difference is that we share the self-attention and FFN modules in the encoder and decoder. We find that such parameter sharing can enhance the semantic mapping between different languages not only in the embedding-level (e.g, shared BPE vocabulary), but also in the module-level, thus strongly regularizing model during training (Xia et al., 2019).

To conclude, the total MLM loss for a training instance $(\boldsymbol{x}, \boldsymbol{y})$, by exchanging the $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ in Eq. (5) and Eq. (7), can be further formalized as:

$$\begin{aligned}\mathcal{L}_{\mathrm{MLM}}(\boldsymbol{x}, \boldsymbol{y}) &= \mathcal{L}_{\mathrm{IS-MLM}}(\boldsymbol{x}) + \mathcal{L}_{\mathrm{IS-MLM}}(\boldsymbol{y}) + \mathcal{L}_{\mathrm{CS-MLM}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{L}_{\mathrm{CS-MLM}}(\boldsymbol{y}, \boldsymbol{x}) \\ &= -\Big(\log P(\bar{\boldsymbol{x}}|\hat{\boldsymbol{x}}) + \log P(\bar{\boldsymbol{y}}|\hat{\boldsymbol{y}}) + \log P(\bar{\boldsymbol{y}}|\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}_{<t}) + \log P(\bar{\boldsymbol{x}}|\hat{\boldsymbol{y}}, \hat{\boldsymbol{x}}_{<t})\Big)\end{aligned} \tag{8}$$

2.3 Fine-tuning on Downstream NLU and NLG Tasks

VECO for Cross-lingual Natural Language Understanding Since the mainstream framework for NLU is an encoder-only Transformer, we only keep the self-attention and FFN modules while throwing out the cross-attention module in each layer (Figure 1(b)). Note that the cross-attention modules only occupy less than 20% of the total parameters, which is smaller than the discarded generator of ELECTRA (Clark et al., 2020).

VECO for Cross-lingual Natural Language Generation Considering the most popular backbone network for cross-lingual generation tasks like machine translation is the standard Transformer encoder-decoder model, we reorganize the VECO modules to act like that. As shown in Figure 1(c), the self-attention and FFN modules constitute the Transformer encoder for contextual modeling, and the three modules work together to act as a decoder for both inner-sequence and cross-sequence contextual modeling. Due to the training difficulty and inference speed of the deep network, we can choose a subset (e.g., 6 layers) of all layers to assemble the decoder. More in-depth analysis can be found in Table 4 and Section 5.2.

3 Pre-training Setup

Model Configuration We pre-train a 24-layer model with 1024 embedding/hidden size and 4096 feed-forward size ($\sim$ 662M parameters). We do not use language embeddings to allow our model to better deal with downstream tasks of unseen languages. We adopt the same 250K vocabulary that is also used by XLM-R (Conneau et al., 2019) and mBART (Liu et al., 2020b).

Pre-Training Datasets For monolingual training datasets, we reconstruct Common-Crawl Corpus used in XLM-R (Conneau et al., 2019). We up/down-sample the monolingual text like XLM from each language with a smoothing parameter $\alpha = 0.5$. For bilingual data, we perform collection from the OPUS website[4] like previous works (Lample & Conneau, 2019; Chi et al., 2020b). More details about the languages and statistics of our training corpus can be founded in Appendix A.

Optimization Settings For each training iteration, we alternately sample a batch of adjacent segments from monolingual corpus and a batch of parallel sentences from bilingual datasets to conduct a pair of masked input $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$. We firstly perform IS-MLM for both $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$. Then, we reuse their hidden states from the last layer to perform CS-MLM. Meanwhile, we also adopt the translation masked language modeling (TLM) proposed in XLM (Lample & Conneau, 2019) when the inputs are parallel bilingual sentences. Thus the overall training objective is the sum of three MLM objectives. During training, the model parameters except for cross-attention are initialized by XLM-R. We first freeze the parameters of XLM-R and only update the cross-attention parameters for faster convergence. Then, we jointly train the whole model. We pre-train our model with mixed-precision training using 64 Nvidia Telsa V100 32GB GPUs. More hyperparameters can be founded in Table 5.

[4] http://opus.nlpl.eu/

Model	Sentence Classification		Structured Prediction		Question Answering			Sentence Retrieval		Avg.
	XNLI	**PAWS**	**POS**	**NER**	**XQuAD**	**MLQA**	**TyDiQA**	**BUCC**	**Tatoeba**	
	Acc	Acc	F1	F1	F1/EM	F1/EM	F1/EM	F1	Acc	
***Cross-lingual Transfer**: Fine-tune model on English training set and test on all languages*										
mBERT†	65.4	81.9	70.3	62.2	64.5/49.4	61.4/44.2	59.7/43.0	56.7	38.7	59.6
XLM†	69.1	80.9	70.1	61.2	59.8/44.3	48.5/32.6	43.6/29.1	56.8	32.6	55.5
XLM-R†	79.2	86.4	72.6	65.4	**76.6/60.8**	71.6/**53.2**	65.1/45.0	66.0	57.3	68.1
VECO	**79.9**	**88.7**	**75.1**	**65.7**	77.3/61.8	**71.7/53.2**	**67.6/49.1**	**85.0**	**75.1**	**73.1**
***Translate-Train-All**: Fine-tune model on English training data and translated data of the target language*										
XLM-R‡	82.6	90.4	-	-	**80.2/65.9**	72.8/54.3	66.5/47.7	-	-	-
XLM-R*	82.8	90.2	72.6	65.4	80.0/65.8	73.0/54.3	74.5/58.3	80.2	75.2	74.4
VECO	**83.0**	**91.1**	**75.1**	**65.7**	79.9/66.3	**73.1/54.9**	**75.0/58.9**	**89.3**	**86.9**	**77.2**

Table 1: XTREME results on each dataset (as of Oct 02, 2020). Averaged results on the four categories can be found at leaderboard: `https://sites.research.google/xtreme`. "†" and "‡" indicates results from Hu et al. (2020) and Fang et al. (2020), respectively. "*" indicates the results obtained by our implementation. The detailed results for each language are in Appendix C.

4 EXPERIMENTS ON LANGUAGE UNDERSTANDING

4.1 EXPERIMENTAL SETUP

Downstream Tasks We conduct NLU evaluations on XTREME (Hu et al., 2020), a representative massively multilingual multi-task benchmark. It consists of various NLU tasks over 40 languages. XTREME tasks can be classified into four different categories: (1) sentence-pair classification: XNLI, PAWSX; (2) structured prediction: POS and NER; (3) question answering: XQuAD, MLQA and TyDiQA; (4) sentence retrieval: BUCC and Tatoeba. Tasks in the first three categories only provide golden training corpus in English and dev/test set in different target languages. For the two zero-shot sentence retrieval tasks, no training datasets are provided. We refer the reader to Hu et al. (2020) for additional details about the datasets.

Fine-tuning Setting Following previous works (Conneau et al., 2019; Hu et al., 2020), we consider two typical fine-tuning settings: (1) *Cross-lingual Transfer* which fine-tunes the pre-trained model only using English golden data and directly performs inference on the test data of different target languages; (2) *Translate-Train-All* which first automatically translates the English golden data to the remaining target languages and fine-tunes a multilingual model on the concatenation of all data. We use the machine-translated data released by XTREME, except for two sequence-labeling tasks (POS, NER) since the golden token labels in the target language are unavailable. We directly fine-tune our model on the XTREME official codebase[5]. See Appendix B for more training details and hyperparameters. To have a fair comparison with the strong baseline XLM-R (Conneau et al., 2019) under the translate-train-all setting, we also show the results of XLM-R using the same fine-tuning hyperparameters as VECO.

4.2 EXPERIMENTAL RESULTS

The detailed test results of 9 tasks on the XTREME benchmark are shown in Table 1. It demonstrates that the proposed VECO outperforms previous cross-lingual models on most of the datasets. Compared to XLM-R, it averagely scores 5.0 and 2.8 points higher under the cross-lingual transfer and translation-train-all settings, respectively. It is worth noting that, VECO delivers a large improvement on *zero-shot* sentence retrieval tasks (BUCC, Tatoeba). This phenomenon reflects that our model has a strong cross-lingual modeling ability, thus it can better mine parallel sentences in a multilingual corpus. The reasons are two-folds. First is the introduction of more bilingual data during pre-training, which is a direct and effective way to enhance the cross-lingual ability of the model. Second is the mutual improvement between two pre-training tasks and the superiority of model design.

[5] `https://github.com/google-research/xtreme`

Models	Datasets		Tasks			Languages									Avg.
	Mono.	Bili.	IS-MLM	CS-MLM	TLM	en	de	es	fr	ru	tr	th	vi	zh	
$\text{XLM}_{\text{SMALL}}$	✓		✓			76.6	58.0	63.5	62.0	56.1	54.8	54.4	57.6	55.5	59.8
$\text{VECO}_{\text{SMALL}}$	✓		✓	✓		77.1	60.9	64.7	62.5	56.1	55.3	54.7	57.5	56.6	60.6
$\text{XLM}_{\text{SMALL}}$	✓	✓	✓		✓	78.3	55.1	72.1	72.5	68.2	51.1	48.8	51.6	52.9	61.2
$\text{VECO}_{\text{SMALL}}$	✓	✓	✓	✓	✓	80.0	66.4	75.6	75.6	70.9	58.7	56.1	63.3	62.9	67.7

Table 2: XNLI accuracy scores for each language under the cross-lingual transfer setting. Both XLM and VECO are small-sized models trained from scratch on the same monolingual (Mono.) and bilingual (Bili.) corpus using the same hyperparameters. For this set of experiments, we only use a subset of the full training corpus and report the results of languages appeared in them.

To analyze whether bilingual data plays a leading role in improving performance, we conduct a set of more fair experiments. We train small-sized XLM and VECO models from scratch using the same subset of training data and hyperparameters (see Appendix A for details). Table 2 shows the results of XNLI, the most widely used cross-lingual evaluation dataset before XTREME released. We observe that, when using monolingual corpus only, VECO can outperform XLM by 0.8 points. It suggests that our models can benefit from adjacent sentences used by the CS-MLM task to be equipped with a stronger ability of contextual modeling. Moreover, when trained on both the monolingual and bilingual corpus, VECO can achieve a larger improvement compared to XLM. It reveals that VECO can better utilize the bilingual corpus, compared to only-optimized translation language modeling (TLM) in XLM.

5 EXPERIMENTS ON LANGUAGE GENERATION

5.1 EXPERIMENTAL SETUP

Datasets We choose machine translation (MT) task, a typical cross-lingual generation scenario. In order to illustrate the generality of our approach and have a fair comparison with the most recent state-of-the-art Transformer works (Liu et al., 2020a), we choose two most widely used datasets: WMT14 English→German (En-De) and English→French (En-Fr) translation. WMT14 En-De is a medium-resource dataset which provides 4.5M pairs for training and validation. We adopt standard newstest2014 as the test set. WMT14 En-Fr is a high-resource dataset which contains 36M pairs of parallel sentences. We use newstest2012+newstest2013 for validation and newstest2016 for test. We measure case-insensitive tokenized BLEU with `multi-bleu.perl` to have a fair comparison with previous Transformer variants and de-tokenized `SacreBLEU`[6] to avoid the influence of different tokenization between models (Post, 2018).

Fine-tuning Setting In theory, VECO can initialize a 24-layer encoder and 24-layer decoder Transformer model. However, since VECO follows the post-layernorm design of vanilla Transformer (Vaswani et al., 2017) like XLM-R, we find it is hard to train such a deep post-layernorm Transformer variant without careful parameter searching. This phenomenon is consistent with the findings in recent researches about deep transformer (Liu et al., 2020a; Bachlechner et al., 2020). Since these works also show that deeper encoders are more worthwhile than deeper decoders, thus the main results of VECO and XLM-R are based on the most normal 6-layer decoder with the full 24-layer encoder. The batch size is 64k and 256k for En-De and En-Fr respectively. The total training updates are set to 100k. The learning rate is 1e-4/2e-4, with linear warmup over the first 16k steps and linear decay. We run En-De and En-Fr MT experiments on 16 and 32 V100 GPUs respectively. We average the last 10 checkpoints and use beam search with a beam size of 5.

Baselines We consider two types of Transformer baselines: randomly initialized and cross-lingual models initialized. For random initialization, we take the original Transformer-big and the state-of-the-art Deep Transformer (Liu et al., 2020a) into consideration. Besides, we also reproduce a Transformer baseline which adopts the same fine-tuning hyperparameters as VECO but with random initialization. For cross-lingual pre-training methods, we include mBART, which shows impressive results on MT. We also conduct the WMT experiments for XLM-R, following the totally same fine-tuning settings as VECO.

[6]Hash: BLEU+case.mixed+lang.en-{de,fr}+numrefs.1+smooth.exp+test.wmt14/full+tok.13a+version.1.4.9

Model	#Layers		WMT14 En-Fr		WMT14 En-De	
	Enc	Dec	BLEU	sacreBLEU	BLEU	sacreBLEU
Randomly Initialize						
Transformer-Big (Vaswani et al., 2017)	6	6	41.0	-	28.4	-
Transformer-xBig (Ours impl.)	24	6	42.8	40.3	28.6	27.7
Deep-Transfromer (Liu et al., 2020a)	60	12	43.8	41.8	30.1	29.5
Cross-lingual Models Initialize						
XLM	6	6	-	-	28.9	-
XLM-R (Ours impl.)	24	6	43.7	41.1	30.8	29.9
mBART (Ours impl.)	12	12	43.2	41.0	30.0	29.1
VECO	24	6	**44.4**	**42.0**	**31.5**	**30.5**

Table 3: Results on WMT14 En-Fr and WMT14 En-De Machine Translation.

Method	#Layers		WMT14 En-De	
	Enc	Dec	BLEU	sacreBLEU
Randomly Initialize	24	3	28.5	27.6
	24	6	28.6	27.7
*Randomly Initialize + More Bilingual Data**				
	24	6	30.4	29.4
VECO Initialize	24	First-3	30.8	29.8
	24	Last-3	31.2	30.3
	24	First-6	31.1	30.1
	24	Last-6	**31.5**	**30.5**

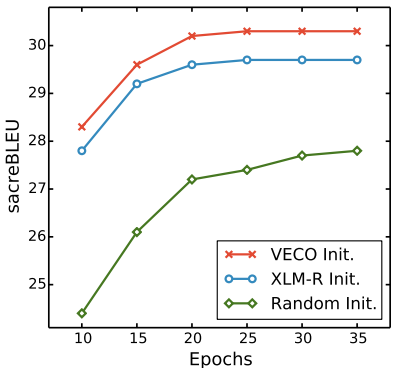

Table 4: BLEU scores (left) and learning curves (right) of different initialization methods.

5.2 Experimental Results

Table 3 shows the comparison between VECO and baselines. Compared to existing cross-lingual models, VECO can consistently outperform the best models about 1 BLEU. We can also observe that VECO can largely outperform randomly initialized same-sized Transformer-xBig model by 2 BLEU. And it even beats the state-of-the-art Deep-Transformer with a 60-layer encoder and 12-layer decoder on both datasets.

Figure 4 (left) contrasts two ways initializing a Transformer with $n(< 24)$ decoder layers via selecting: (1) the first n layers; (2) the last n layers from a 24-layer pre-trained VECO model. We consider $n = \{3, 6\}$ to conduct experiments. The predominant method adopted in this paper, the straightforward strategy of selecting the last n layers, exhibits better performance. It is possibly because the last several layers play a more important role in making predictions over the whole vocabulary. Moreover, we can find that there is 0.2$\sim$0.3 BLEU improvement when increasing the decoder layers from 3 to 6.[7] Regardless of the initialization method, the VECO-initialized model can gain consistent 1$\sim$2 BLEU improvement over the randomly-initialized model.

Moreover, Figure 4 (right) displays the sacreBLEU scores of same-sized (24-layer encoder and 6-layer decoder) models during training. We find that VECO-initialized model can get a surprising more than 28 sacreBLEU score just after 10 epochs, which is better than the final score of the randomly initialized model at 35 epochs. It reveals that VECO can provide a fairly good initialization for the machine translation model, which can converge quickly and further boost the results.

To investigate whether the exciting improvement in MT mainly comes from a) the use of parallel corpus during pre-training or b) the superiority of the designed model and pre-training tasks, we also conduct a more comparable experiment. We first train an out-of-domain Transformer-xBig model

[7]However, we observe that only marginal improvement can be gained when further increasing the decoder layers to 12 or 24 in our primary experiments, which is also in line with the findings in Liu et al. (2020a).

using the whole En-De parallel data ($\sim$ 68M) used in VECO pre-training, and then continue-train the model on the in-domain WMT14 En-De training dataset. Results are shown in Table 4 (left) marked with *. Under this set of a totally fair comparison, VECO still maintains a lead of 1.1 BLEU score. This directly confirms that the improvement on MT is not only due to the use of bilingual data. More importantly, reasonable pre-training tasks and model design ensure *better* use of bilingual and large-scale unlabeled multilingual corpus.

6 Related Work

Encoder-only Cross-lingual Pre-training mBERT (Devlin et al., 2019) is the first work to pre-train a Transformer encoder over multiple languages. There have been several extensions that use the same encoder-only backbone as mBERT, with the main difference is the introduction of more training corpus and pre-training tasks. XLM (Lample & Conneau, 2019) utilizes both monolingual and bilingual corpus to perform mask language modeling. XLM-R (Conneau et al., 2019) extends to be built on RoBERTa (Liu et al., 2019) using larger monolingual training data. Unicoder (Huang et al., 2019), ALM (Yang et al., 2020), and InfoXLM (Chi et al., 2020b) propose new pre-training tasks to better utilize the bilingual data. These works deliver impressive performance on cross-lingual understanding tasks, while only marginal improvement has been gained on cross-lingual generation tasks like machine translation, especially on high-resource languages.

Encoder-Decoder Cross-lingual Pre-training BART (Lewis et al., 2019) pre-trains denoising auto-encoder built with a standard Transformer-based encoder-decoder architecture. And mBART (Liu et al., 2020b) extends it to the multilingual setting, demonstrating significant gains in low/medium-resource machine translation, but with a decrease in high resource languages. Chi et al. (2020a) first trains an encoder via MLM and then frozen the encoder to train the decoder only via two generative tasks. A similar approach is also proposed in Liang et al. (2020) that extends Unicoder (Huang et al., 2019) to generation tasks, with the main difference exists in the joint training of encoder-decoder. All these cross-lingual models emphasize training a dedicated model for NLG, thus these didn't improve or even hurt the NLU capabilities of the encoder. Or they require more computation and memory to match the performance of the encoder-only models.

Unified Language Representation for NLU and NLG BART (Lewis et al., 2019) and UNILM (Dong et al., 2019) also endeavor to build a unified model for NLU and NLG tasks, with the core idea that allows the model to see both unidirectional and bidirectional context. UNILM pre-trains a Transformer encoder with an ensemble of attention masks, while BART pre-trains a Transformer encoder-decoder model with arbitrary noising functions. Our work differs from them in several ways. Firstly, they only focus on the monolingual (English) domain, while we dedicate to the more challenging multilingual domain. Secondly, since multiple languages cannot be completely mapped to the same space at every layer only via self-attention like UNILM, the cross-attention module is important to model the cross-lingual mapping. Thus we drop the way to pre-train an encoder-only model under the multilingual scenario. Thirdly, previous work (Xia et al., 2019) have shown that sharing the encoder and decoder of the Transformer can strengthen the semantic correlation between languages and regularize the encoder-decoder with high capacity on machine translation. Due to the above reasons, VECO varies during training to train specific modules via a bidirectional task (IS-MLM) and unidirectional task (CS-MLM).

7 Conclusion

We present VECO, a variable cross-lingual pre-training model, targeting at initializing both NLU preferred encoder-only and NLG specialized encoder-decoder Transformer. We analyze three core modules in standard Transformer and propose two masked language modeling tasks to train the reasonable combinations of them. The two tasks jointly optimize for inner-sequence understanding and cross-sequence generation, enabling them to boost each other via strong regularization from module sharing. On this account, VECO achieves consistent improvements on various language understanding and generation tasks compared to existing cross-lingual encoder-only and encoder-decoder approaches, opening up new ways of thinking about pre-trained backbone architecture.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/pdf/1607.06450.pdf.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. *arXiv preprint arXiv:2003.04887*, 2020. URL https://arxiv.org/pdf/2003.04887.pdf.

Bin Bi, Chenliang Li, Chen Wu, Ming Yan, and Wei Wang. PALM: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. *arXiv preprint arXiv:2004.07159*, 2020. URL https://arxiv.org/pdf/2004.07159.pdf.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. Cross-lingual natural language generation via pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020a. URL https://arxiv.org/pdf/1909.10481.pdf.

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. *arXiv preprint arXiv:2007.07834*, 2020b. URL https://arxiv.org/pdf/2007.07834.pdf.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020. URL https://openreview.net/pdf?id=r1xMH1BtvB.

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. *arXiv preprint arXiv:1911.02116*, 2019. URL https://arxiv.org/pdf/1911.02116v1.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. URL https://www.aclweb.org/anthology/N19-1423.pdf.

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*, 2019. URL https://arxiv.org/pdf/1905.03197.pdf.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. FILTER: An enhanced fusion method for cross-lingual language understanding. *arXiv preprint arXiv:2009.05166*, 2020. URL https://arxiv.org/pdf/2009.05166.pdf.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Training pruned neural networks. *CoRR*, abs/1803.03635, 2018. URL https://arxiv.org/pdf/1803.03635.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. URL https://arxiv.org/pdf/1512.03385.pdf.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*, 2020. URL https://arxiv.org/pdf/2003.11080.pdf.

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. *arXiv preprint arXiv:1909.00964*, 2019. URL https://arxiv.org/pdf/1909.00964.pdf.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. URL https://arxiv.org/pdf/1808.06226.pdf.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019. URL https://arxiv.org/pdf/1901.07291.pdf.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. URL https://arxiv.org/pdf/1910.13461.pdf.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*, 2020. URL https://arxiv.org/pdf/2004.01401.pdf.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. Very deep transformers for neural machine translation. *arXiv preprint arXiv:2008.07772*, 2020a. URL https://arxiv.org/pdf/2008.07772.pdf.

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 pre-training approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/pdf/1907.11692.pdf.

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. *arXiv preprint arXiv:2001.08210*, 2020b. URL https://arxiv.org/pdf/1910.13461.pdf.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation*, 2018. URL https://arxiv.org/pdf/1804.08771.pdf.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked sequence to sequence pre-training for language generation. 2019. URL https://arxiv.org/pdf/1905.02450.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzman, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019. URL https://arxiv.org/pdf/1911.00359.pdf.

Yingce Xia, Tianyu He, Xu Tan, Fei Tian, Di He, and Tao Qin. Tied transformers: Neural machine translation with shared encoder and decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019. URL https://www.aaai.org/ojs/index.php/AAAI/article/view/4487/4365.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. Alternating language modeling for cross-lingual pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6480/6336.

A Pre-Training Details

For monolingual data, following XLM-R (Conneau et al., 2019), we build a clean Common-Crawl Corpus using an open-source tool CCNet (Wenzek et al., 2019). There are 1TB monolingual data in 50 languages before up/down-sampling. We collect bilingual corpus from the OPUS

Pre-training Hyperparameters	**Large**	**Small**
Number of layers	24	6
Hidden Size	1024	768
FFN inner hidden size	4096	3072
Attention heads	16	12
Attention head size	64	64
Embedding Size	1024	768
Mask percent (monolingual/ bilingual)	15%/25%	15%/25%
Learning Rate Decay	Linear	Linear
Warmup steps	12k	12k
Learning Rate	2e-4	3e-4
Adam ϵ	1e-6	1e-6
Adam β_1	0.9	0.9
Adam β_2	0.98	0.999
Attention Dropout	0.1	0.1
Dropout	0.1	0.1
Weight Decay	0.01	0.01
Max Sequence Length (monolingual/bilingual)	512/128	512/128
Batch Size (monolingual/bilingual)	1024/4096	1024/4096
Train Steps	240k	240k
Total Parameters	662M	247M

Table 5: The pre-training hyperparameters.

website[8], including MultiUN, UNPC, Bombay, EU-bookshop, OpenSubtitles2018, Tanzil, GlobalVoices, ParaCrawl, MultiParaCrawl, DGT, Tilde, Europarl, Wikipedia, ECB, TED2013, News-Commentary, Ubuntu, Books, UN, infopankki-v1, EUconst, and Bianet. In total, there are 1TB bilingual training corpus before pre-processing. We then apply subword tokenization directly on raw text data using Sentence Piece Model (Kudo & Richardson, 2018) without any additional pre-processing. We use the whole corpus to train VECO, while using a subset ($\sim 1/4$) that contains 33 languages to train $\text{XLM}_{\text{SMALL}}$ and $\text{VECO}_{\text{SMALL}}$.

The full set of pre-training hyperparameters for small-sized and large-sized VECO (default) are listed in Table 5.

B NLU FINE-TUNING DETAILS

We consider two typical fine-tuning settings:

- *Cross-lingual Transfer*: For all tasks except sentence retrieve tasks (BUCC and Tatoeba), we select the model with the best average result over all the languages on the dev sets, by searching the learning rate over [1e-5,2e-5,3e-5], training epoch over [3,5,10], and batch size over [16,32,64]. As for sentence retrieve tasks without fine-tuning on any parallel sentences at all, we use the average word embeddings in the middle layer to extract the sentence representations.
- *Translate-Train-All*: We found that the large-sized training datasets can benefit from a smaller learning rate. Therefore, for translate-train-all, we select the model on the dev sets, by searching the learning rate over [3e-6,5e-6,1e-5]. For XNLI, PAWS-X, XQuAD and MLQA, we fine-tune our model directly on the translation data provided by the official XTREME repo[9]. Following the participants (FILTER and Anonymous1) on the XTREME leaderboard, we use several methods to further improve the scores. For TyDiQA, we start fine-tuning based on the best XQuAD translate-train-all model. For the sentence retrieve tasks, BUCC and Tatoeba, we use the averaged representation in the middle layer of the

[8] `http://opus.nlpl.eu/`

[9] `https://github.com/google-research/xtreme`

best XNLI model. Note that we only use these tricks under the translate-train-all setting, wishing to have a possible fair comparison on the XTREME leaderboard. However, we still use less fine-tuned data compared to concurrent works FILTER and Anonymous1. Specifically, FILTER, a model-agnostic fine-tune method, uses more translation data on POS and NER tasks. Anonymous1 uses other labeled data outside of XTREME.

The full set of fine-tuning hyperparameters is listed in Table 6.

Learning Rate	[1e-5,2e-5,3e-5] for *Cross-lingual Transfer*
	[3e-6,5e-6,1e-5] for *Translate-Train-All*
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Batch Size	[16,32,64]
Train Epochs	[3,5,10]

Table 6: The fine-tuning hyperparameters

C Detailed Results on XTREME

The detailed results of each XTREME task under the cross-lingual transfer and translate-train-all settings on all languages are listed in the following tables.

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	**Avg.**
Cross-lingual Transfer																
mBERT†	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9	67.8	49.7	54.1	60.9	57.2	69.3	67.8	65.4
MMTE†	79.6	64.9	70.4	68.2	67.3	71.6	69.5	63.5	66.2	61.9	66.2	63.6	60.0	69.7	69.2	67.5
XLM†	82.8	66.0	71.9	72.7	70.4	75.5	74.3	62.5	69.9	58.1	65.5	66.4	59.8	70.7	70.2	69.1
XLM-R†	**88.7**	77.2	83.0	82.5	80.8	83.7	82.2	75.6	79.1	71.2	**77.4**	78.0	**71.7**	79.3	78.2	79.2
VECO	88.2	**79.2**	**83.1**	**82.9**	**81.2**	**84.2**	**82.8**	**76.2**	**80.3**	**74.3**	77.0	**78.4**	71.3	**80.4**	**79.1**	**79.9**
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
XLM-R†	88.6	82.2	85.2	84.5	84.5	85.7	84.2	**80.8**	81.8	77.0	80.2	82.1	**77.7**	82.6	82.7	82.6
VECO	**88.9**	**82.4**	**86.0**	**84.7**	**85.3**	**86.2**	**85.8**	80.1	**83.0**	77.2	**80.9**	**82.8**	75.3	**83.1**	**83.0**	**83.0**

Table 7: XNLI accuracy scores for each language. "†" and ‡ results are provided by Hu et al. (2020) and Conneau et al. (2019), respectively.

Model	en	de	es	fr	ja	ko	zh	**Avg.**
Cross-lingual Transfer								
mBERT	94.0	85.7	87.4	87.0	73.0	69.6	77.0	81.9
XLM	94.0	85.9	88.3	87.4	69.3	64.8	76.5	80.9
MMTE	93.1	85.1	87.2	86.9	72.0	69.2	75.9	81.3
XLM-R	94.7	89.7	90.1	90.4	78.7	79.0	82.3	86.4
VECO	**96.2**	**91.3**	**91.4**	**92.0**	**81.8**	**82.9**	**85.1**	**88.7**
Translate-Train-All								
VECO	96.4	93.0	93.0	93.5	87.2	86.8	87.9	91.1

Table 8: PAWS-X accuracy scores for each language.

Model	de	fr	ru	zh	**Avg.**
Cross-lingual Transfer					
mBERT	62.5	62.6	51.8	50.0	56.7
XLM	56.3	63.9	60.6	46.6	56.8
MMTE	67.9	63.9	54.3	53.3	59.8
XLM-R	67.5	66.5	73.5	56.7	66.0
VECO	**89.6**	**84.6**	**87.4**	**78.5**	**85.0**
Translate-Train-All					
VECO	93.0	88.7	89.9	85.7	89.3

Table 9: BUCC results (F1 scores).

Model	en	ar	de	el	es	hi	ru	th	tr	vi	zh	**Avg.**
Cross-lingual Transfer												
mBERT	83.5 / 72.2	61.5 / 45.1	70.6 / 54.0	62.6 / 44.9	75.5 / 56.9	59.2 / 46.0	71.3 / 53.3	42.7 / 33.5	55.4 / 40.1	69.5 / 49.6	58.0 / 48.3	64.5 / 49.4
XLM	74.2 / 62.1	61.4 / 44.7	66.0 / 49.7	57.5 / 39.1	68.2 / 49.8	56.6 / 40.3	65.3 / 48.2	35.4 / 24.5	57.9 / 41.2	65.8 / 47.6	49.7 / 39.7	59.8 / 44.3
MMTE	80.1 / 68.1	63.2 / 46.2	68.8 / 50.3	61.3 / 35.9	72.4 / 52.5	61.3 / 47.2	68.4 / 45.2	48.4 / 35.9	58.1 / 40.9	70.9 / 50.1	55.8 / 36.4	64.4 / 46.2
XLM-R	86.5 / 75.7	68.6 / 49.0	**80.4 / 63.4**	**79.8 / 61.7**	**82.0 / 63.9**	**76.7 / 59.7**	**80.1 / 64.3**	**74.2 / 62.8**	**75.9 / 59.3**	79.1 / 59.0	59.3 / 50.0	76.6 / 60.8
VECO	**87.6 / 76.5**	**73.6 / 56.1**	79.8 / 62.2	79.6 / 61.6	81.2 / 61.6	74.7 / 57.6	78.7 / 62.1	72.8 / 60.6	75.1 / 58.3	**79.0 / 59.8**	**69.2 / 59.2**	**77.3 / 61.8**
Translate-Train-All												
VECO	88.3/77.9	76.9/61.1	80.5/64.6	81.5/64.1	84.2/66.8	78.8/62.5	80.2/66.1	77.0/70.4	77.8/62.2	82.5/63.7	71.6/69.4	79.9/66.3

Table 10: XQuAD results (F1 / EM) for each language.

Model	en	ar	de	es	hi	vi	zh	**Avg.**
Cross-lingual Transfer								
mBERT	80.2 / 67.0	52.3 / 34.6	59.0 / 43.8	67.4 / 49.2	50.2 / 35.3	61.2 / 40.7	59.6 / 38.6	61.4 / 44.2
XLM	68.6 / 55.2	42.5 / 25.2	50.8 / 37.2	54.7 / 37.9	34.4 / 21.1	48.3 / 30.2	40.5 / 21.9	48.5 / 32.6
MMTE	78.5 / –	56.1 / –	58.4 / –	64.9 / –	46.2 / –	59.4 / –	58.3 / –	60.3 / 41.4
XLM-R	83.5 / 70.6	**66.6 / 47.1**	**70.1 / 54.9**	**74.1 / 56.6**	**70.6 / 53.1**	**74.0 / 52.9**	62.1 / 37.0	71.6 / 53.2
VECO	**83.6 / 70.5**	65.0 / 44.6	69.8 / 54.6	73.8 / 55.6	69.1 / 51.4	73.1 / 51.8	**67.3 / 43.6**	**71.7 / 53.2**
Translate-Train-All								
VECO	84.1/71.3	67.8/47.1	70.7/55.8	74.6/56.6	71.1/53.4	74.8/54.4	68.8/45.8	73.1/54.9

Table 11: MLQA results (F1 / EM) for each language.

Model	en	ar	bn	fi	id	ko	ru	sw	te	**Avg.**
Cross-lingual Transfer										
mBERT	**75.3 / 63.6**	62.2 / 42.8	49.3 / 32.7	59.7 / 45.3	64.8 / 45.8	**58.8 / 50.0**	60.0 / 38.8	57.5 / 37.9	49.6 / 38.4	59.7 / 43.9
XLM	66.9 / 53.9	59.4 / 41.2	27.2 / 15.0	58.2 / 41.4	62.5 / 45.8	14.2 / 5.1	49.2 / 30.7	39.4 / 21.6	15.5 / 6.9	43.6 / 29.1
MMTE	62.9 / 49.8	63.1 / 39.2	55.8 / 41.9	53.9 / 42.1	60.9 / 47.6	49.9 / 42.6	58.9 / 37.9	63.1 / 47.2	54.2 / 45.8	58.1 / 43.8
XLM-R	71.5 / 56.8	67.6 / 40.4	**64.0 / 47.8**	70.5 / 53.2	**77.4 / 61.9**	31.9 / 10.9	**67.0 / 42.1**	**66.1 / 48.1**	70.1 / 43.6	65.1 / 45.0
VECO	71.3 / 58.2	**73.1 / 52.8**	58.9 / 42.5	**70.9 / 55.1**	77.2 / 60.0	54.2 / 39.9	66.1 / 37.6	65.8 / 45.7	**70.6 / 50.7**	**67.6 / 49.1**
Translate-Train-All										
VECO	77.2/64.8	77.0/57.5	72.2/56.6	76.6/59.3	80.0/64.4	63.4/52.2	72.8/50.5	79.4/67.1	76.0/58.0	75.0/58.9

Table 12: TyDiQA-GolP results (F1 / EM) for each language.

Model	af	ar	bg	de	el	en	es	et	eu	fa	fi	fr	he	hi	hu	id	it
Cross-lingual Transfer																	
mBERT	86.6	56.2	85.0	85.2	81.1	95.5	86.9	79.1	60.7	66.7	78.9	84.2	56.2	67.2	78.3	71.0	88.4
XLM	88.5	63.1	85.0	85.8	84.3	95.4	85.8	78.3	62.8	64.7	78.4	82.8	65.9	66.2	77.3	70.2	87.4
MMTE	86.2	65.9	87.2	85.8	77.7	**96.6**	85.8	81.6	61.9	67.3	81.1	84.3	57.3	76.4	78.1	**73.5**	89.2
XLM-R	**89.8**	**67.5**	**88.1**	88.5	86.3	96.1	88.3	86.5	72.5	70.6	85.8	87.2	**68.3**	**76.4**	82.6	72.4	89.4
VECO	88.3	67.4	87.4	**88.5**	**86.7**	95.9	**89.0**	**87.8**	**75.1**	**70.9**	**86.2**	**88.9**	67.5	76.2	**82.9**	72.9	**89.9**
	ja	kk	ko	mr	nl	pt	ru	ta	te	th	tl	tr	ur	vi	yo	zh	**Avg.**
mBERT	**49.2**	70.5	49.6	69.4	88.6	86.2	85.5	59.0	75.9	41.7	81.4	68.5	57.0	53.2	**55.7**	61.6	71.5
XLM	49.0	70.2	50.1	68.7	88.1	84.9	86.5	59.8	76.8	**55.2**	76.3	66.4	61.2	52.4	20.5	65.4	71.3
MMTE	48.6	70.5	**59.3**	74.4	83.2	86.1	88.1	63.7	81.9	43.1	80.3	71.8	61.1	56.2	51.9	**68.1**	73.5
XLM-R	15.9	78.1	53.9	80.8	89.5	87.6	89.5	**65.2**	**86.6**	47.2	92.2	76.3	**70.3**	56.8	24.6	25.7	73.8
VECO	31.4	**79.3**	53.1	**84.3**	**89.8**	**88.3**	**90.2**	64.3	85.8	48.0	**93.7**	**77.2**	69.2	**58.1**	26.2	39.4	**75.1**

Table 13: POS results (Accuracy) for each language.

Model	en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv
Cross-lingual Transfer																				
mBERT	**85.2**	77.4	41.1	77.0	70.0	78.0	72.5	77.4	75.4	66.3	46.2	77.2	79.6	56.6	65.0	76.4	**53.5**	81.5	29.0	**66.4**
XLM	82.6	74.9	44.8	76.7	70.0	78.1	73.5	74.8	74.8	62.3	49.2	79.6	78.5	**57.7**	66.1	76.5	53.1	80.7	23.6	63.0
MMTE	77.9	74.9	41.8	75.1	64.9	71.9	68.3	71.8	74.9	62.6	45.6	75.2	73.9	54.2	66.2	73.8	47.9	74.1	**31.2**	63.9
XLM-R	84.7	**78.9**	**53.0**	81.4	**78.8**	78.8	**79.5**	**79.6**	79.1	60.9	61.9	79.2	**80.5**	56.8	**73.0**	79.8	53.0	81.3	23.2	62.5
VECO	83.8	77.5	48.2	**83.9**	77.2	**79.4**	79.3	75.4	**80.4**	**68.3**	**68.2**	**80.6**	80.1	55.0	71.0	**80.9**	52.9	**81.7**	19.4	63.2
	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh
mBERT	64.6	45.8	59.6	52.3	58.2	**72.7**	45.2	81.8	80.8	64.0	67.5	50.7	48.5	3.6	71.7	71.8	36.9	71.8	**44.9**	**42.7**
XLM	67.7	**57.2**	26.3	59.4	62.4	69.6	47.6	81.2	77.9	63.5	68.4	53.6	49.6	0.3	**78.6**	71.0	43.0	70.1	26.5	32.4
MMTE	60.9	43.9	58.2	44.8	58.5	68.3	42.9	74.8	72.9	58.2	66.3	48.1	46.9	**3.9**	64.1	61.9	37.2	68.1	32.1	28.9
XLMR	**71.6**	56.2	**60.0**	**67.8**	**68.1**	57.1	54.3	**84.0**	81.9	69.1	70.5	59.5	55.8	1.3	73.2	76.1	56.4	**79.4**	33.6	33.1
VECO	67.1	51.2	59.9	63.4	65.0	70.0	**56.1**	83.4	**83.1**	**71.3**	**70.5**	**60.5**	**56.2**	1.4	71.3	**80.4**	**69.3**	76.0	37.4	29.1

Table 14: NER results (F1) for each language.

Model	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja
Cross-lingual Transfer																		
BERT	42.7	25.8	49.3	17	77.2	29.8	68.7	29.3	25.5	46.1	39	66.3	41.9	34.8	38.7	54.6	58.4	42
XLM	43.2	18.2	40	13.5	66.2	25.6	58.4	24.8	17.1	32.2	32.2	54.5	32.1	26.5	30.1	45.9	56.5	40
XLMR	**58.2**	47.5	71.6	43	88.8	61.8	75.7	52.2	35.8	70.5	71.6	73.7	66.4	72.2	65.4	77	68.3	60.6
VECO	48.2	**70.9**	**86.7**	**57.7**	**97.5**	**81.5**	**94.8**	**89.7**	**62.9**	**82.1**	**87.9**	**88.8**	**74.7**	**80.7**	**87.6**	**89.6**	**89.2**	**83.2**
Translate-Train-All																		
VECO	80.9	85.1	91.3	78.1	98.5	89.5	97.4	94.8	79.8	93.1	95.4	93.7	85.8	94.2	93.8	93.0	92.2	92.8
	jv	ka	kk	ko	ml	mr	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh
Cross-lingual Transfer																		
BERT	17.6	20.5	27.1	38.5	19.8	20.9	68	69.9	61.2	11.5	14.3	16.2	13.7	16	34.8	31.6	62	71.6
XLM	**22.4**	22.9	17.9	25.5	20.1	13.9	59.6	63.9	44.8	12.6	20.2	12.4	31.8	14.8	26.2	18.1	47.1	42.2
XLMR	14.1	52.1	48.5	61.4	65.4	56.8	80.8	82.2	74.1	20.3	26.4	35.9	29.4	36.7	65.7	24.3	74.7	68.3
VECO	17.6	**58.5**	**53.9**	**75.3**	**80.1**	**64.2**	**94.4**	**92.8**	**88.6**	**37.4**	**61.9**	**65.8**	**84.5**	**52.5**	**89.3**	**64.3**	**85.8**	**82.7**
Translate-Train-All																		
VECO	35.1	83.0	74.1	88.7	94.8	82.5	95.9	94.6	92.2	69.7	82.4	91.0	94.7	73.0	95.2	63.8	95.1	93.9

Table 15: Tatoeba results (Accuracy) for each language