ENHANCING CROSS-LINGUAL TRANSFER BY MANIFOLD MIXUP

Huiyun Yang\textsuperscript{1}, Huadong Chen\textsuperscript{1}, Hao Zhou\textsuperscript{*}\textsuperscript{1}, Lei Li\textsuperscript{2}
\textsuperscript{1}ByteDance AI Lab
\textsuperscript{2}University of California, Santa Barbara
\{yanghuiyun.11, chenuadong.howard, zhouhao.nlp\}@bytedance.com
leili@cs.ucsb.edu

\textbf{ABSTRACT}

Based on large-scale pre-trained multilingual representations, recent cross-lingual transfer methods have achieved impressive transfer performances. However, the performance of target languages still lags far behind the source language. In this paper, our analyses indicate such a performance gap is strongly associated with the cross-lingual representation discrepancy. To achieve better cross-lingual transfer performance, we propose the cross-lingual manifold mixup (X-MIXUP) method, which adaptively calibrates the representation discrepancy and gives compromised representations for target languages. Experiments on the XTREME benchmark show X-MIXUP achieves 1.8\% performance gains on multiple text understanding tasks, compared with strong baselines, and reduces the cross-lingual representation discrepancy significantly.

1 INTRODUCTION

Many natural language processing tasks have shown exciting progress utilizing deep neural models. However, these deep models often heavily rely on sufficient annotation data, which is not the case in the multilingual setting. The fact is that most of the annotation data are collected for popular languages like English and Spanish (Ponti et al., 2019; Joshi et al., 2020), while many long-tail languages could hardly obtain enough annotations for supervised training. As a result, cross-lingual transfer learning (Prettenhofer & Stein, 2011; Wan et al., 2011; Ruder et al., 2019) is crucial, transferring knowledge from the annotation-rich source language to low-resource or zero-resource target languages. In this paper, we focus on the zero-resource setting, where labeled data are only available in the source language.

Recently, multilingual pre-trained models (Conneau & Lample, 2019; Conneau et al., 2020a; Xue et al., 2020) offer an effective way for cross-lingual transfer, which yield a universal embedding space across various languages. Such universal representations make it possible to transfer knowledge from the source language to target languages through the embedding space, significantly improving the transfer learning performance (Chen et al., 2019; Zhou et al., 2019; Keung et al., 2019; Fang et al., 2020). Besides, Conneau et al. (2018) proposes translate-train, a simple yet effective cross-lingual data augmentation method, which constructs pseudo-training data for each target language via machine translation. Although these works have achieved impressive improvements in cross-lingual transfer (Hu et al., 2020; Ruder et al., 2021), significant performance gaps between the source language and target languages still remain (see Table 1). Hu et al. (2020) refers to the gap as the cross-lingual transfer gap, a difference between the performance on the source and target languages.

To investigate how the cross-lingual transfer gap emerges, we perform relevant analyses, demonstrating that transfer performance correlates well with the cross-lingual representation discrepancy (see Section 3 for details). Here the cross-lingual representation discrepancy means the degree of difference between the source and target language representations in the universal embedding space. As shown in Figure 1(a) in translate-train, the representation distribution of Spanish almost overlaps with English, while Arabic shows a certain representation discrepancy compared with English and

\* Corresponding author.
Figure 1: Representation visualization of four languages: English (en), Spanish (es), Arabic (ar) and Swahili (sw) based on XLM-R. We plot the sentence representation of the XNLI test set, which is parallel across 15 languages. We average hidden states of the last layer to get sentence representations and implement the dimensionality reduction by PCA. Obviously, the cross-lingual representation discrepancies are large in translate-train, but X-MIXUP reduces the discrepancy significantly.

Swahili performs larger discrepancy, where translate-train achieves 88.6 accuracy on English, 85.7 on Spanish, 82.2 on Arabic and 77.0 on Swahili. Intuitively, a larger representation discrepancy could lead to a worse cross-lingual transfer performance.

In this paper, we propose the Cross-Lingual Manifold Mixup (X-MIXUP) approach to fill the cross-lingual transfer gap. Based on our analyses, reducing the cross-lingual representation discrepancy is a promising way to narrow the transfer gap. Given the cross-lingual representation discrepancy is hard to remove, X-MIXUP directly faces the issue and explicitly accommodates the representation discrepancy in the neural networks, by mixing the representation of the source and target languages during training and inference. With X-MIXUP, the model itself can learn how to escape the discrepancy, which adaptively calibrates the representation discrepancy and gives compromised representations for target languages to achieve better cross-lingual transfer performance. X-MIXUP is motivated by robust deep learning (Vincent et al., 2008), while X-MIXUP adopts the mixup (Zhang et al., 2018) idea to handle the cross-lingual discrepancy.

Specifically, X-MIXUP is designed upon the translate-train approach, faced with the exposure bias (Ranzato et al., 2016) problem and data noise problem. During training, the source sequence is a real sentence and the target sequence is a translated one, while situations are opposite during inference. Besides, the translated text often introduces some noises due to imperfect machine translation systems. To address them, we further impose the Scheduled Sampling (Bengio et al., 2015) and Mixup Ratio in X-MIXUP to handle the distribution shift problem and data noise problem, respectively.

We verify X-MIXUP on the cross-lingual understanding benchmark XTREME (Hu et al., 2020), which includes several understanding tasks and covers 40 languages from diverse language families. Experimental results show X-MIXUP achieves 1.8% performance gains across different tasks and languages, comparing with strong baselines. It also reduces the cross-lingual representation discrepancy significantly, as Figure 1(b) shows.

2 RELATED WORK

Multilingual Representation Learning Recent studies have demonstrated the superiority of large-scale pre-trained multilingual representations on downstream tasks. Multilingual BERT (mBERT; Devlin et al., 2019) is the first work to extend the monolingual pre-training to the multilingual setting. Then, several extensions achieve better cross-lingual performances by introducing more monolingual or parallel data and new pre-training tasks, such as Unicoder (Huang et al., 2019), XLM-R (Conneau et al., 2020a), ALM (Yang et al., 2020), MMTE (Siddhant et al., 2020), InfoXLM (Chi et al., 2020), HICTL (Wei et al., 2020), ERNIE-M (Ouyang et al., 2020), mT5 (Xue et al., 2020), nmT5 (Kale et al., 2021), AMBER (Hu et al., 2021) and VECO (Luo et al., 2021). They have been the standard backbones of current cross-lingual transfer methods.
Table 1: Cross-lingual transfer performances of POS and NER tasks on languages with different data resources or different language families, where there are only labeled training data in English. The data resource refers to the resource of each language utilized in the pre-training process. For the language family, English belongs to the Germanic languages, so we divide languages into two types: Germanic one and others. Results show high-resource languages outperform low-resource ones significantly and languages dissimilar to the source language tend to perform worse.

<table>
<thead>
<tr>
<th>Language Type</th>
<th>Source</th>
<th>Language Resource</th>
<th>Language Family</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>High-resource</td>
<td>Low-resource</td>
</tr>
<tr>
<td></td>
<td></td>
<td>en</td>
<td>es, it, pt</td>
</tr>
<tr>
<td>mBERT</td>
<td>POS</td>
<td>95.5</td>
<td>86.9</td>
</tr>
<tr>
<td></td>
<td>NER</td>
<td>85.2</td>
<td>84.8</td>
</tr>
<tr>
<td>XLM-R</td>
<td>POS</td>
<td>96.1</td>
<td>89.4</td>
</tr>
<tr>
<td></td>
<td>NER</td>
<td>84.7</td>
<td>81.3</td>
</tr>
</tbody>
</table>

Cross-lingual Transfer Learning

Cross-lingual transfer learning [Prettenhofer & Stein, 2011; Wan et al., 2011; Ruder et al., 2019] aims to transfer knowledge learned from source languages to target languages. According to the type of transfer learning [Pan & Yang, 2010], previous cross-lingual transfer methods can be divided into three categories: instance transfer, parameter transfer, and feature transfer. The cross-lingual transferability improves a lot when engaged with the instance transfer by translation (i.e. translate-train, translate-test) or other cross-lingual data augmentation methods [Singh et al., 2019; Bornea et al., 2020; Qin et al., 2020; Zheng et al., 2021]. Chen et al. [2019] and Zhou et al. [2019] focus on the parameter transfer to learn a share-private model architecture. Besides, other works implement the feature transfer to learn the language-invariant features by adversarial networks [Keung et al., 2019; Chen et al., 2019] or re-alignment [Libovický et al., 2020; Zhao et al., 2020]. X-MIXUP utilizes both the instance transfer and feature transfer, which is based on the translate-train data augmentation approach and implements the feature transfer by cross-lingual manifold mixup.

Mixup and Its Variants

Mixup [Zhang et al., 2018] proposes to train models on the linear interpolation at both the input level and label level, which is effective to improve the model robustness and generalization. Generally, the interpolated pair is selected randomly. Manifold mixup [Verma et al., 2019] performs the interpolation in the latent space by conducting the linear combinations of hidden states. Previous mixup methods [Chen et al., 2020; Jindal et al., 2020] focus on the monolingual setting. However, X-Mixup focuses on the cross-lingual setting and faces many new challenges (see Section 4 for details). Besides, in contrast to previous mixup methods, X-MIXUP mixes the parallel pairs, which share the same semantics across different languages. As a result, the choice of parallel pairs for interpolation can build a smart connection between the source and target languages.

3 ANALYSES OF THE CROSS-LINGUAL TRANSFER PERFORMANCE

In this section, we concentrate on the cross-lingual transfer performance and find it is strongly associated with the cross-lingual representation discrepancy. Firstly, we observe the cross-lingual transfer performance on different target languages and propose an assumption. Then we conduct qualitative and quantitative analyses to verify it.

Although previous studies [Hu et al., 2020; Ruder et al., 2021] have shown impressive improvements on cross-lingual transfer, the cross-lingual transfer gap is still pretty large, more than 16 points in Hu et al. [2020]. Furthermore, results in Table 1 show the performance of low-resource languages and dissimilar languages fall far behind other languages in cross-lingual transfer tasks.

Compared with English, the representations of other languages, especially low-resource languages, are not well-trained [Lauscher et al., 2020; Wu & Dredze, 2020], because high-resource languages dominate the representation learning process, which results in the cross-lingual representation discrepancy. Besides, dissimilar languages often show differences in language characteristics (like vocabulary, word order), which also leads to the representation discrepancy. As a result, we assume...
Table 2: Spearman’s rank correlation $\rho$ between the CKA score and cross-lingual transfer performance on two XTREME tasks, where $^\dagger$ denotes training on the source language, and $^\ddagger$ denotes the translate-train approach. $^*$ denotes the p-value is lower than 0.05. Results indicate the correlation is solid.

<table>
<thead>
<tr>
<th>Task</th>
<th>XNLI$^\dagger$</th>
<th>XNLI$^\ddagger$</th>
<th>PAWS-X$^\dagger$</th>
<th>PAWS-X$^\ddagger$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>0.76*</td>
<td>0.69*</td>
<td>0.90*</td>
<td>0.93*</td>
</tr>
</tbody>
</table>

that the cross-lingual transfer performance is closely related to the representation discrepancy between the source language and target languages.

Following Conneau et al. [2020b], we utilize the linear centered kernel alignment (CKA; Kornblith et al., 2019) score to indicate the cross-lingual representation discrepancy

$$\text{CKA}(X, Y) = \frac{||Y^TX||^2_F}{||X^TX||^2_F||Y^TY||^2_F},$$

where $X$ and $Y$ are parallel sequences from the source and target languages, respectively. A higher CKA score denotes a smaller cross-lingual representation discrepancy.

To verify our assumption, we perform qualitative and quantitative analyses on the relationship between the CKA score and cross-lingual transfer performance. Figure 3 in Appendix B indicates a higher CKA score tends to induce better cross-lingual transfer performance. We also calculate the Spearman’s rank correlation between the CKA score and the transfer performance in Table 2, which shows a strong correlation between them. Both the trend and correlation score confirm the cross-lingual transfer performance is highly related to the cross-lingual representation discrepancy.

4 Methodology: X-Mixup

Based on the aforementioned analyses, we believe that reducing the cross-lingual representation discrepancy is the key to filling the cross-lingual transfer gap. In this section, we propose X-Mixup to explicitly reduce the representation discrepancy by implementing the manifold mixup between the source language and target language. With X-Mixup, the model can adaptively calibrate the representation discrepancy and give compromised representations for target languages. This section will first introduce the overall architecture of X-Mixup and its details. After that, the training objectives and inference process will be shown.

4.1 Overall Architecture

Figure 2 illustrates the overall architecture of X-Mixup. Sequences from the source and target languages are first encoded separately. Then within the encoder, X-Mixup implements the manifold mixup between the paired sequences (original sequence and its translation) within a specific layer, where Mixup Ratio controls the degree of mixup and Scheduled Sampling schedules the data sampling process during training.

Notations We use $S$ to denote the source language and $T$ to denote the target language. $h_l^i$ denotes the hidden states of a sequence in layer $l$. $D$ denotes the real text data collection and $\tilde{D}$ denotes the translation data collection. For downstream understanding tasks, there are annotation data in the source language $D_{S\text{Train}} = (X_{S\text{Train}}, Y_{S\text{Train}})$ and raw test data in the target language $D_{T\text{Test}} = (X_{T\text{Test}})$. Through translate-train, we can get pseudo-training data in the target language $D_{T\text{Train}} = (X_{T\text{Train}}, Y_{T\text{Train}})$. Similarly, through translate-test, we can get pseudo-test data in the source language $D_{S\text{Test}} = (X_{S\text{Test}})$. During training, the Scheduled Sampling process uses translation data $D_{T\text{Train}} = (Y_{T\text{Train}})$ from the source language. Note that we use translation data ($X_{T\text{Train}}^2$ and $X_{S\text{Test}}^3$) and translate-train labels ($Y_{T\text{Train}}^3$) from the official XTREME repository, which is in the same setting as baselines.

Basic Model We use mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020a) as the backbone model. Within each layer, there are two sub-layers: the multi-head attention layer and the...
feed-forward layer followed by the residual connection and layer norm. We use the same multi-head attention layer (see details in Appendix A.1) as BERT (Devlin et al., 2019), where inputs are query, key, and value respectively. In layer $l+1$, the hidden states of the source sequence $x_S$ and target sequence $x_T$ are acquired by the multi-head attention

$$h_{l+1}^{l+1} = \text{MultiHead}(h_{l+1}^S, h_{l+1}^S, h_{l+1}^S), \quad h_{l+1}^{l+1} = \text{MultiHead}(h_{l+1}^T, h_{l+1}^T, h_{l+1}^T).$$  \hfill (2)

**Manifold Mixup** To reduce the cross-lingual representation discrepancy, a straightforward idea is to find compromised representations between the source and target languages. It’s difficult to find such representations because of varying degrees of differences across languages, like vocabulary and word order. However, manifold mixup provides an elegant way to get intermediate representations by conducting linear interpolation on hidden states.

To extract target-related information from the source hidden states, the target hidden states are used as the query, and the source hidden states are used as the key and value. This cross-attention process is computed as

$$h_{T|S}^{l+1} = \text{MultiHead}(h_{T|S}^{l+1}, h_{S|T}^{l+1}, h_{S|T}^{l+1}),$$  \hfill (3)

which shares parameters with the multi-head attention. The manifold mixup process mixes the target hidden states $h_{T|^T}^{l+1}$ and the source-aware target hidden states $h_{T|S}^{l+1}$ based on the mixup ratio $\lambda$

$$\tilde{h}_{T}^{l+1} = \text{LN}(\lambda h_{T|S}^{l+1} + (1 - \lambda)h_{T|T}^{l+1}),$$  \hfill (4)

where $\lambda$ is an instance-level parameter, ranging from 0 to 1, and indicates the degree of manifold mixup. LN denotes the layer norm operation.

**Mixup Ratio** The machine translation process may change the original semantics and introduce data noises in varying degrees (Castilho et al., 2017; Fomicheva et al., 2020). Thus we introduce the translation quality modeling in the mixup process to handle this problem. Following Fomicheva et al.

---

3In this section, the feed-forward layer is omitted for simplification.
we use the entropy of attention weights to measure the translation quality
\[ H(A) = -\frac{1}{I} \sum_i \sum_j A_{ij} \log A_{ij}, \]
where \( A_{ij} = \text{softmax}(h_i^T h_j^S \sqrt{n}). \)

\( I \) is the number of target tokens and \( J \) is the number of source tokens. Lower entropy implies better cross-lingual alignment and higher translation quality.

To introduce the translation quality modeling into the manifold mixup process, we compute the mixup ratio as
\[ \lambda = \lambda_0 \cdot \sigma(H(A) + H(A^T))W + b, \]
where \( \sigma \) is the sigmoid function, and \( W, b \) are trainable parameters. \( \lambda_0 \) is the max value of the mixup ratio, which is set to 0.5 in this paper. We consider two-way alignment in the translation quality modeling, i.e. \( H(A) \) and \( H(A^T) \).

Scheduled Sampling: The source sequences utilized in training and inference are drawn from different distributions. During training, the source sequence is a real text from \( D_{S}^{\text{Train}} \), while during inference, the source sequence is a translation from \( \tilde{D}_{S}^{\text{Test}} \). This discrepancy, commonly called the exposure bias, leads to a gap between training and inference.

Motivated by the scheduled sampling approach [Bengio et al., 2015] in NMT, we sample the source sequence dynamically during training. Specifically, the source sequence fed into the manifold mixup is either a real text from \( D_{S}^{\text{Train}} \) or translation from \( \tilde{D}_{S}^{\text{Train}} \) with a certain probability
\[ p \begin{cases} p < p^*, & x_s \in D_{S}^{\text{Train}}, \\ p > p^*, & x_s \in \tilde{D}_{S}^{\text{Train}}, \end{cases} \]
where \( p^* \) is decreasing during training to match the situation of inference. We utilize the inverse sigmoid decay [Bengio et al., 2015], which decreases \( p^* \) as a function of the index of mini-batch.

4.2 Final Training Objective

The training loss is composed of two parts: the task loss and the consistency loss
\[ \mathcal{L} = \mathcal{L}_{\text{task}} + \text{MSE}(r_S, r_T) + \text{KL}(p_S, p_T). \]

where \( \text{MSE}(\cdot) \) is Mean Squared Error and \( \text{KL}(\cdot, \cdot) \) is Kullback-Leibler divergence. \( r_s \) is the sequence representation and \( p_s \) is the predicted probability distribution of downstream tasks.

The task loss \( \mathcal{L}_{\text{task}} \) is the sum of the source language task loss \( \mathcal{L}_{S}^{\text{task}} \) and target language one \( \mathcal{L}_{T}^{\text{task}} \), weighted by the hyper-parameter \( \alpha \), which is utilized to balance the training process
\[ \mathcal{L}_{\text{task}} = \alpha \mathcal{L}_{S}^{\text{task}} + (1 - \alpha) \mathcal{L}_{T}^{\text{task}}. \]

For classification, structured prediction, and span extraction tasks, the task loss is the cross-entropy loss (see details in Appendix A.2). For structured prediction tasks, it is non-trivial to implement the token-level label mapping across different languages. Thus we use the label probability distribution, predicted by the source language task model, as the pseudo-label for training, where tokens and labels are corresponding.

The consistency loss is composed of two parts: the representation consistency loss and the prediction consistency loss. The first loss is a regularization term and provides a way to align representations across different languages [Ruder et al., 2019]. The second loss is to make better use of the supervision of downstream tasks. It only exists in the classification task, as the translation process does not change the label of this task, while in other tasks, it does.

4.3 Inference

During inference, the manifold mixup process is the same as training, except for the Scheduled Sampling process. Concretely, for the source language, only translation data are available in the
Table 3: Main results on the XTREME benchmark. † denotes using other data augmentation strategy in addition to machine translation. ‡ denotes results from Ruder et al. (2021), which is an updated version of Hu et al. (2020).

<table>
<thead>
<tr>
<th>Model</th>
<th>Pair Sentence</th>
<th>Structured Prediction</th>
<th>Question Answering</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>XNLI</td>
<td>PAWS-X</td>
<td>POS</td>
<td>NER</td>
</tr>
<tr>
<td>Based on XLM-R-large</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XLM-R (Hu et al., 2020)</td>
<td>79.2</td>
<td>86.4</td>
<td>73.8</td>
<td>65.4</td>
</tr>
<tr>
<td>Trans-train (Wei et al., 2020)</td>
<td>82.9</td>
<td>90.1</td>
<td>74.6</td>
<td>66.8</td>
</tr>
<tr>
<td>Filter (Fang et al., 2020)</td>
<td>83.9</td>
<td>91.4</td>
<td>76.2</td>
<td>67.7</td>
</tr>
<tr>
<td>XTUNE (Zheng et al., 2021)</td>
<td>84.8</td>
<td>91.6</td>
<td>79.3†</td>
<td>69.9†</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>85.3</td>
<td>91.8</td>
<td>78.4</td>
<td>69.0</td>
</tr>
<tr>
<td>Based on mBERT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mBERT (Hu et al., 2020)</td>
<td>65.4</td>
<td>81.9</td>
<td>71.5</td>
<td>62.2</td>
</tr>
<tr>
<td>Joint-Align (Zhao et al., 2020)</td>
<td>72.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Trans-train (Hu et al., 2020)</td>
<td>75.1</td>
<td>88.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>78.8</td>
<td>89.7</td>
<td>76.5</td>
<td>65.0</td>
</tr>
</tbody>
</table>

Inference stage, without real data, so we use \( x_s \in \tilde{D}_{\text{Test}} \). For classification tasks, we synthesize the predictions of both the source and target sequences by taking the mean of the predicted probability distributions as the final prediction. For structured prediction and QA tasks, we only consider the prediction of the target sequence.

5 Experiments

This section first introduces the cross-lingual understanding benchmark, XTREME. Then briefly introduces the configurations of downstream tasks and baselines. Finally, shows the main results of baselines and X-MIXUP on XTREME.

5.1 Tasks and Settings

**Tasks** In our experiments, we focus on three types of tasks in XTREME: (1) sentence pair classification task: XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019); (2) structured prediction task: POS (Nivre et al., 2018) and NER (Pan et al., 2017); (3) question answering task: XQuAD (Artetxe et al., 2020), MLQA (Lewis et al., 2020) and TyDiQA (Clark et al., 2020). The details of these datasets can refer to Hu et al. (2020). We utilize the translate-train and translate-test data from the XTREME repo \(^5\) which also provide the pseudo-label of translate-train data for classification tasks and question answering tasks. The rest translation data are from Google Translate \(^6\).

**Models** Experiments are based on two multilingual pre-trained models: mBERT and XLM-R. We use the pre-trained models of Huggingface Transformers \(^7\) as the backbone model.

**Hyper-parameters** We select XNLI, POS, and MLQA as representative tasks to search for the best hyper-parameters. The final model is selected based on the averaged performance of all languages on the dev set. We perform grid search over the balance training parameter \( \alpha \) and learning rate from \([0.2, 0.4, 0.6, 0.8]\) and \([3e-6, 5e-6, 2e-5, 3e-5]\). We also search for the best manifold mixup layer from \([1, 4, 8, 12, 16, 20, 24]\). In final results, we implement mixup in the first layer for classification tasks, 4th layer for structured prediction tasks. For QA tasks, we implement mixup in the 16th layer for large model, 8th layer for base model. Concrete details of experiments are presented in Appendix C.1.

5.2 Baselines

We conduct experiments on two strong multilingual pre-trained models to verify the generality of methods: (1) **mBERT** Multilingual BERT is a 12-layer transformer model pre-trained on the multilingual Wikipedia, News articles, Books, Open web text, and In-Progress KLab. We use \( \text{bert-base-multilingual-cased} \) for mBERT and \( \text{xlm-roberta-large} \) for XLM-R.

\(^5\)https://github.com/google-research/xtreme
\(^6\)https://translate.google.com/
\(^7\)We use \( \text{bert-base-multilingual-cased} \) for mBERT and \( \text{xlm-roberta-large} \) for XLM-R.
Table 4: Comparisons between X-MIXUP and XTUNE under the same setting: XLM-R-base model and machine translation data augmentation. Results of XTUNE are from Zheng et al. (2021) Table 4.

<table>
<thead>
<tr>
<th>Model</th>
<th>XNLI</th>
<th>POS</th>
<th>MLQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>XTUNE\textsubscript{R1} [Zheng et al., 2021]</td>
<td>79.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>XTUNE\textsubscript{R2} [Zheng et al., 2021]</td>
<td>78.9</td>
<td>76.6</td>
<td>68.7/51.1</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>80.4</td>
<td>77.8</td>
<td>71.2/53.1</td>
</tr>
</tbody>
</table>

Wikipedias of 104 languages. (2) **XLM-R** is a 24-layer transformer model pretrained on 2.5T data extracted from Common Crawl covering 100 languages. Based on them, these are some strong baselines: (1) **Trans-train** Abbreviation for Translate-train. The training set of the source language is machine-translated to each target language and then the model is trained on the concatenation of all training sets. (2) **Joint-Align** [Zhao et al., 2020] aligns the monolingual sub-spaces of the source and target language by minimizing the distances of embeddings for matched word pairs. (3) **Filter** [Fang et al., 2020] splices the representation of the target sequence and its translation in intermediate layers to extract multilingual knowledge. (4) **XTUNE** [Zheng et al., 2021] uses two types of consistency regularization based on four types of data augmentation.

### 5.3 Main Results

Results on the XTREME benchmark are shown in Table 3. Concrete results for each task are presented in Appendix C.2. Compared with strong baselines, X-MIXUP shows its superiority across different backbones and tasks, which indicates its generality. X-MIXUP outperforms Trans-train by 2.2% based on mBERT, and X-MIXUP outperforms Filter by 1.5% based on XLM-R. The superiority of X-MIXUP over Filter is that X-MIXUP gives a calibrated representation for target languages, not just the concatenation of two representations. Besides, X-MIXUP considers the noise of translation data and limits the noise propagation by introducing mixup ratio.

XTUNE achieves the best results on structured prediction tasks and the low-resource QA task TyDiQA (only 3.7k training data in English), but XTUNE uses three other data augmentation approaches in addition to machine translation. To make a fairer comparison, we conduct experiments under the same setting in Table 4, which indicates X-MIXUP outperforms XTUNE on three types of tasks with only machine translation data augmentation. Besides, X-MIXUP only needs one-stage training, while XTUNE implements a two-stage training algorithm. However, X-MIXUP and XTUNE are complementary, where the former focuses on finding better representations for target languages while the latter concentrates on the cross-lingual data augmentation and consistency regularization.

### 6 Analysis and Discussion

To better understand X-MIXUP and explore how X-MIXUP influences the cross-lingual transfer performance, we conduct analyses on several questions. Results show X-MIXUP achieves performance improvements across different languages and it also reduces the cross-lingual representation discrepancy obviously. Table 6 in Appendix B verifies the effectiveness of X-Mixup on both seen and unseen languages. Besides, ablation results show the cross-lingual manifold mixup training contributes a lot to cross-lingual transfer.

**Q1: How X-MIXUP influences the cross-lingual representation discrepancy?** Language centroid (Rosenberg & Hirschberg, 2007) is the mean of the representations within each language. We plot the language centroid of different methods (see Figure 4 in Appendix B), which indicates X-MIXUP brings closer language centroids significantly. We also calculate the CKA scores of the XNLI dataset (see Table 7 in Appendix B). Results show X-MIXUP reduces the cross-lingual representation discrepancy evenly across different target languages, improving the CKA score by 10.4% on average. In conclusion, both the language centroids visualization and the CKA score improvement indicate X-MIXUP reduces the cross-lingual representation discrepancy effectively.

---

8 In this section, we utilize the XLM-R-large model as the backbone model.
<table>
<thead>
<tr>
<th>Model</th>
<th>XNLI</th>
<th>POS</th>
<th>MLQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-MIXUP</td>
<td>85.3</td>
<td>78.4</td>
<td>76.5/58.1</td>
</tr>
<tr>
<td>w/o mixup</td>
<td>82.9</td>
<td>75.7</td>
<td>72.7/54.8</td>
</tr>
<tr>
<td>w/o mixup inference</td>
<td>84.0</td>
<td>77.6</td>
<td>75.6/57.3</td>
</tr>
<tr>
<td>w/o scheduled sampling</td>
<td>84.6</td>
<td>78.0</td>
<td>76.3/57.9</td>
</tr>
<tr>
<td>w/o consistency loss</td>
<td>84.2</td>
<td>78.0</td>
<td>76.5/58.0</td>
</tr>
<tr>
<td>$\lambda = \lambda_0$</td>
<td>84.1</td>
<td>77.8</td>
<td>75.8/57.5</td>
</tr>
</tbody>
</table>

(Q2) How X-MIXUP influences the cross-lingual transfer gap? We compare the cross-lingual transfer gap in Appendix B Table 8. Compared with Trans-train, X-MIXUP reduces the averaged gap by 39.8% and shows its superiority across three types of tasks. Compared with state-of-the-art methods, X-MIXUP achieves the smallest cross-lingual transfer gap on four out of seven datasets, which suggests the effectiveness of X-MIXUP on classification and QA tasks.

(Q3) What is the essential component of X-MIXUP? There are five major components of X-MIXUP: cross-lingual manifold mixup training, mixup inference, Mixup Ratio, Scheduled Sampling, and consistency loss. To better understand X-MIXUP, we implement ablation studies in Table 5. Comparisons between X-MIXUP and w/o mixup show the effectiveness of cross-lingual manifold mixup across different tasks, and even without mixup inference (translate-test data), the mixup training can also achieve 2.6% improvements on average. Besides, comparisons between X-MIXUP and $\lambda = \lambda_0$ show the effectiveness of introducing the translation quality modeling in the mixup process. Scheduled sampling achieves more performance improvements on the classification task, as the task shares labels across languages, and scheduled sampling can prevent the model from solely relying on the gold source sequence to make predictions. In addition, the consistency loss is also more effective on the classification task, because there is additional prediction consistency loss which can transfer the task capability from the source language to target languages. Detailed ablation results on the consistency loss are shown in Appendix B Table 9, which shows the KL consistency loss contributes more than the MSE consistency loss on the classification task.

(Q4) Which layer is the best to implement the manifold mixup? We implement the cross-lingual manifold mixup in different layers (see Figure 5 in Appendix B for details) and find different tasks prefer different mixup layers. Although different tasks have their own preferences, no matter which layer we mix, the cross-lingual transfer performance can be improved, except for mixing within a higher layer on classification tasks. The drop in classification task is mainly because the source and target sequences share the same task label. Performing mixup in a higher layer may make the model rely on the source sequence and ignore the target sequence. The structured prediction task is not sensitive to the mixup layer, mainly because this task relies on both the short and long dependence. For QA tasks, the cross-lingual transfer performance shows a trend from rise to decline as the mixup layer increases. The QA task needs higher-level understanding, but higher layers are more language-specific, where sequences from different languages have different gold answers.

7 Conclusion

This paper focuses on enhancing the cross-lingual transfer performance on understanding tasks. Considering the large cross-lingual transfer gap in recent works, this paper first analyses related factors and finds this gap is strongly associated with the cross-lingual representation discrepancy. Then X-MIXUP is proposed to alleviate the discrepancy, which gives compromised representations for target languages by implementing the manifold mixup between the source and target languages. Empirical evaluations on XTREME verify the effectiveness of X-MIXUP across different tasks and languages. Besides, both the visualization and quantitative analyses show X-MIXUP reduces the cross-lingual representation discrepancy effectively. Furthermore, X-Mixup can also be applied to the multilingual pre-training process by implementing the cross-lingual manifold mixup on parallel data. Findings on the relationship between the cross-lingual transfer performance and representation discrepancy shed light on a promising way to boost cross-lingual transfer for future research.
REFERENCES


LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University.


C. Wan, Rong Pan, and Jiefei Li. Bi-weighting domain adaptation for cross-language text classification. In IJCAI, 2011.


A Method Details

A.1 Multi-head Attention

In the multi-head attention layer, multiple attention heads are concatenated

\[
\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O,
\]

and each head is the scaled dot-product attention

\[
\text{head}_i = \text{Attention}(QW^q_i, KW^k_i, VW^v_i),
\]

\[
\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V,
\]

where \( W^O, W^q, W^k \) and \( W^v \) are trainable parameters.

A.2 Training Objective

For classification tasks (e.g. NLI), the task loss is the cross-entropy loss

\[
\mathcal{L}_{\text{task}} = -\sum_j^C y_j \log p_j,
\]

where \( C \) is the size of the label set.

For structured prediction tasks (e.g. POS) and span extraction tasks (e.g. QA), the task loss is also the cross-entropy loss

\[
\mathcal{L}_{\text{task}} = -\sum_i^n \sum_j^C y_{ij} \log p_{ij},
\]

where \( n \) is the sequence length.
B ANALYSIS RESULTS

Figure 3: Performances on PAWS-X and XNLI test set, where languages are sorted by decreasing CKA scores. The trend indicates the performance gets worse along with the CKA score decreasing.

Figure 4: Language centroids visualization of the POS test set, which indicates X-Mixup brings closer these centroids obviously.

Figure 5: Performances on implementing X-MIXUP (solid line) in different layers and Trans-train (dashed line) on three downstream tasks.
Table 6: Performances on XNLI test set, where Trans-train and X-MIXUP are trained on 8 seen languages and tested on both these seen languages and 7 unseen languages. \(\Delta\) is the performance difference between X-MIXUP and Trans-train. Results show X-MIXUP performs better than Trans-train by a large margin on both seen and unseen languages.

<table>
<thead>
<tr>
<th>Model</th>
<th>Seen Languages</th>
<th>Unseen Languages</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>en</td>
<td>bg</td>
<td>el</td>
</tr>
<tr>
<td>Trans-train</td>
<td>87.6</td>
<td>84.7</td>
<td>84.2</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>89.5</td>
<td>87.1</td>
<td>86.3</td>
</tr>
<tr>
<td>(\Delta)</td>
<td>+1.9</td>
<td>+2.4</td>
<td>+1.9</td>
</tr>
</tbody>
</table>

Table 7: CKA scores and performances on XNLI test set, where \(\Delta\) is the score or performance difference between X-MIXUP and Trans-train. Results show X-MIXUP improves the CKA scores evenly across different target languages and the performance improvements are diverse. There is no obvious correlation between the CKA score improvement and performance improvement.

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>es</th>
<th>de</th>
<th>vi</th>
<th>fr</th>
<th>bg</th>
<th>tr</th>
<th>el</th>
<th>ru</th>
<th>ar</th>
<th>hi</th>
<th>sw</th>
<th>ur</th>
<th>th</th>
<th>zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKA score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans-train</td>
<td>1.00</td>
<td>0.81</td>
<td>0.80</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
<td>0.76</td>
<td>0.75</td>
<td>0.75</td>
<td>0.74</td>
<td>0.74</td>
<td>0.73</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.77</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>1.00</td>
<td>0.88</td>
<td>0.87</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td>0.83</td>
<td>0.82</td>
<td>0.84</td>
<td>0.83</td>
<td>0.82</td>
<td>0.81</td>
<td>0.80</td>
<td>0.79</td>
<td>0.85</td>
</tr>
<tr>
<td>(\Delta)</td>
<td>0.00</td>
<td>0.07</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.08</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans-train</td>
<td>88.6</td>
<td>85.7</td>
<td>84.5</td>
<td>82.6</td>
<td>84.2</td>
<td>85.2</td>
<td>82.1</td>
<td>84.5</td>
<td>81.8</td>
<td>82.2</td>
<td>80.8</td>
<td>77.0</td>
<td>77.7</td>
<td>80.2</td>
<td>82.7</td>
<td>82.6</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>89.9</td>
<td>87.7</td>
<td>86.9</td>
<td>85.4</td>
<td>87.1</td>
<td>87.3</td>
<td>84.9</td>
<td>86.8</td>
<td>85.1</td>
<td>85.2</td>
<td>83.5</td>
<td>81.2</td>
<td>79.6</td>
<td>83.2</td>
<td>85.2</td>
<td>85.3</td>
</tr>
<tr>
<td>(\Delta)</td>
<td>+1.3</td>
<td>+2.0</td>
<td>+2.4</td>
<td>+2.8</td>
<td>+2.9</td>
<td>+2.1</td>
<td>+2.8</td>
<td>+2.3</td>
<td>+3.3</td>
<td>+3.0</td>
<td>+2.7</td>
<td>+4.2</td>
<td>+1.9</td>
<td>+3.0</td>
<td>+2.5</td>
<td>+2.7</td>
</tr>
</tbody>
</table>

Table 8: The cross-lingual transfer gap (lower is better) of different methods on the XTREME benchmark. For QA tasks, we only show EM scores. \(\dagger\) denotes results from Wei et al. (2020). Overall, X-MIXUP achieves the smallest cross-lingual transfer gap on four out of seven datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>XNLI</th>
<th>PAWS-X</th>
<th>POS</th>
<th>NER</th>
<th>XQuAD</th>
<th>MLQA</th>
<th>TyDiQA</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT (Hu et al., 2020)</td>
<td>16.5</td>
<td>14.1</td>
<td>25.5</td>
<td>23.6</td>
<td>25.0</td>
<td>27.5</td>
<td>22.2</td>
<td>22.1</td>
</tr>
<tr>
<td>XLM-R (Hu et al., 2020)</td>
<td>10.2</td>
<td>12.4</td>
<td>24.3</td>
<td>19.8</td>
<td>16.3</td>
<td>19.1</td>
<td>13.1</td>
<td>16.5</td>
</tr>
<tr>
<td>Trans-train (Hu et al., 2020)</td>
<td>7.3</td>
<td>9.0</td>
<td>22.4(\dagger)</td>
<td>20.5(\dagger)</td>
<td>17.6</td>
<td>22.2</td>
<td>24.2</td>
<td>17.6</td>
</tr>
<tr>
<td>Filter (Fang et al., 2020)</td>
<td>6.0</td>
<td>5.2</td>
<td>19.7</td>
<td>16.3</td>
<td>7.3</td>
<td>15.7</td>
<td>9.2</td>
<td>11.3</td>
</tr>
<tr>
<td>xTUNE (Zheng et al., 2021)</td>
<td>5.5</td>
<td>5.2</td>
<td>17.3</td>
<td>14.8</td>
<td>10.1</td>
<td>18.5</td>
<td>0.9</td>
<td>10.3</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td><strong>4.9</strong></td>
<td><strong>5.2</strong></td>
<td><strong>18.1</strong></td>
<td><strong>15.9</strong></td>
<td><strong>6.7</strong></td>
<td><strong>13.9</strong></td>
<td>9.6</td>
<td><strong>10.6</strong></td>
</tr>
</tbody>
</table>

Table 9: Ablation results on the consistency loss, which show the KL consistency loss contributes more than the MSE consistency loss on the classification task.

<table>
<thead>
<tr>
<th>Model</th>
<th>XNLI</th>
<th>POS</th>
<th>MLQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-MIXUP</td>
<td>85.3</td>
<td>78.4</td>
<td>76.5/58.1</td>
</tr>
<tr>
<td>w/o MSE consistency loss</td>
<td>84.6</td>
<td>78.0</td>
<td>76.5/58.0</td>
</tr>
<tr>
<td>w/o KL consistency loss</td>
<td>84.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>w/o both</td>
<td>84.2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
C EXPERIMENTAL DETAILS

C.1 HYPER-PARAMETERS

For all tasks, we fine-tune on 8 Nvidia V100-32GB GPU cards with the batch size 64. For XQuAD and MLQA, we fine-tune 2 epochs. For other tasks, we fine-tune 4 epochs. There is no dev set in XQuAD, so we use the dev set of MLQA for the model selection. Table 10 shows hyper-parameters used for X-MIXUP.

Table 10: Hyper-parameters used for X-MIXUP, where $\alpha$ is used for balanced training in Eq 8 and $p_k$ is the scheduled sampling decay rate.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Classification</th>
<th>Structured Prediction</th>
<th>QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.4</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>$p_k$</td>
<td>1000</td>
<td>1000</td>
<td>2000</td>
</tr>
</tbody>
</table>

C.2 DETAILED RESULTS

Detailed results of each tasks and languages are shown below. Results of mBERT, XLM, MMTE and XLM-R are from XTREME (Hu et al., 2020). Results of Filter is the best results of Fang et al. (2020).

Table 11: XNLI accuracy scores for each language.

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>de</th>
<th>bg</th>
<th>de</th>
<th>el</th>
<th>fr</th>
<th>hi</th>
<th>ru</th>
<th>sw</th>
<th>th</th>
<th>tr</th>
<th>ur</th>
<th>vi</th>
<th>zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>94.0</td>
<td>85.7</td>
<td>87.4</td>
<td>87.0</td>
<td>73.0</td>
<td>69.6</td>
<td>77.0</td>
<td>77.0</td>
<td>69.6</td>
<td>77.0</td>
<td>77.0</td>
<td>77.0</td>
<td>81.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XLM</td>
<td>94.0</td>
<td>85.9</td>
<td>88.3</td>
<td>87.4</td>
<td>69.3</td>
<td>64.8</td>
<td>76.5</td>
<td>80.9</td>
<td>76.5</td>
<td>80.9</td>
<td>76.5</td>
<td>80.9</td>
<td>81.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMTE</td>
<td>93.1</td>
<td>85.1</td>
<td>87.2</td>
<td>86.9</td>
<td>72.0</td>
<td>69.2</td>
<td>75.9</td>
<td>81.3</td>
<td>75.9</td>
<td>81.3</td>
<td>75.9</td>
<td>81.3</td>
<td>81.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XLM-R</td>
<td>94.7</td>
<td>89.7</td>
<td>90.1</td>
<td>90.4</td>
<td>78.7</td>
<td>79.0</td>
<td>82.3</td>
<td>86.4</td>
<td>82.3</td>
<td>86.4</td>
<td>82.3</td>
<td>86.4</td>
<td>86.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter</td>
<td>95.9</td>
<td>92.8</td>
<td>93.0</td>
<td>93.7</td>
<td>87.4</td>
<td>87.6</td>
<td>89.6</td>
<td>91.5</td>
<td>89.6</td>
<td>91.5</td>
<td>89.6</td>
<td>91.5</td>
<td>91.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XTUNE</td>
<td>96.1</td>
<td>92.6</td>
<td>93.1</td>
<td>93.9</td>
<td>87.8</td>
<td>89.0</td>
<td>88.8</td>
<td>91.6</td>
<td>88.8</td>
<td>91.6</td>
<td>88.8</td>
<td>91.6</td>
<td>91.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>96.3</td>
<td>93.2</td>
<td>93.6</td>
<td>94.6</td>
<td>87.3</td>
<td>88.2</td>
<td>89.5</td>
<td>91.8</td>
<td>89.5</td>
<td>91.8</td>
<td>89.5</td>
<td>91.8</td>
<td>91.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12: PAWS-X accuracy scores for each language.
Table 13: POS results (F1) for each language.

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>ar</th>
<th>af</th>
<th>bg</th>
<th>de</th>
<th>el</th>
<th>es</th>
<th>et</th>
<th>eu</th>
<th>fa</th>
<th>fi</th>
<th>fr</th>
<th>he</th>
<th>hi</th>
<th>hu</th>
<th>id</th>
<th>it</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>31.2</td>
<td>67.1</td>
<td>37.2</td>
<td>67.2</td>
<td>57.4</td>
<td>56.8</td>
<td>27.8</td>
<td>41.5</td>
<td>35.2</td>
<td>42.8</td>
<td>36.3</td>
<td>51.7</td>
<td>37.8</td>
<td>50.2</td>
<td>40.6</td>
<td>51.3</td>
<td>40.0</td>
<td></td>
</tr>
<tr>
<td>XLM-R</td>
<td>27.9</td>
<td>57.8</td>
<td>34.4</td>
<td>59.1</td>
<td>57.4</td>
<td>55.3</td>
<td>27.6</td>
<td>41.1</td>
<td>35.1</td>
<td>42.6</td>
<td>36.2</td>
<td>51.6</td>
<td>37.6</td>
<td>49.8</td>
<td>39.3</td>
<td>50.3</td>
<td>39.0</td>
<td></td>
</tr>
<tr>
<td>Filter</td>
<td>35.0</td>
<td>71.2</td>
<td>42.6</td>
<td>67.8</td>
<td>57.3</td>
<td>56.3</td>
<td>28.2</td>
<td>41.8</td>
<td>35.8</td>
<td>43.4</td>
<td>37.3</td>
<td>52.0</td>
<td>38.4</td>
<td>51.8</td>
<td>40.7</td>
<td>52.3</td>
<td>40.4</td>
<td></td>
</tr>
<tr>
<td>XTUNE</td>
<td>35.2</td>
<td>72.2</td>
<td>41.8</td>
<td>69.0</td>
<td>57.5</td>
<td>56.6</td>
<td>28.4</td>
<td>42.0</td>
<td>36.0</td>
<td>43.6</td>
<td>37.6</td>
<td>52.2</td>
<td>38.6</td>
<td>52.0</td>
<td>40.9</td>
<td>52.5</td>
<td>40.7</td>
<td></td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>34.9</td>
<td>71.0</td>
<td>41.3</td>
<td>68.4</td>
<td>57.2</td>
<td>56.6</td>
<td>27.9</td>
<td>41.6</td>
<td>35.5</td>
<td>43.1</td>
<td>37.0</td>
<td>51.8</td>
<td>38.1</td>
<td>51.6</td>
<td>40.0</td>
<td>51.9</td>
<td>40.1</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>ar</th>
<th>de</th>
<th>es</th>
<th>hi</th>
<th>ru</th>
<th>th</th>
<th>tr</th>
<th>vi</th>
<th>zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>77.5</td>
<td>50.1</td>
<td>35.7</td>
<td>53.8</td>
<td>52.1</td>
<td>50.3</td>
<td>26.1</td>
<td>40.3</td>
<td>31.2</td>
<td>42.1</td>
<td>38.9</td>
</tr>
<tr>
<td>XLM-R</td>
<td>77.0</td>
<td>50.0</td>
<td>35.3</td>
<td>53.9</td>
<td>52.0</td>
<td>50.2</td>
<td>26.0</td>
<td>40.2</td>
<td>31.2</td>
<td>42.0</td>
<td>38.9</td>
</tr>
<tr>
<td>Filter</td>
<td>77.0</td>
<td>50.0</td>
<td>35.3</td>
<td>53.9</td>
<td>52.0</td>
<td>50.2</td>
<td>26.0</td>
<td>40.2</td>
<td>31.2</td>
<td>42.0</td>
<td>38.9</td>
</tr>
<tr>
<td>XTUNE</td>
<td>77.1</td>
<td>50.0</td>
<td>35.4</td>
<td>53.9</td>
<td>52.0</td>
<td>50.2</td>
<td>26.0</td>
<td>40.2</td>
<td>31.2</td>
<td>42.0</td>
<td>38.9</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>77.1</td>
<td>50.0</td>
<td>35.4</td>
<td>53.9</td>
<td>52.0</td>
<td>50.2</td>
<td>26.0</td>
<td>40.2</td>
<td>31.2</td>
<td>42.0</td>
<td>38.9</td>
</tr>
</tbody>
</table>

Table 15: XQuAD results (F1/EM) for each language.

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>ar</th>
<th>de</th>
<th>es</th>
<th>hi</th>
<th>vi</th>
<th>zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>80.7</td>
<td>59.1</td>
<td>33.6</td>
<td>52.9</td>
<td>51.3</td>
<td>50.7</td>
<td>24.8</td>
<td>38.1</td>
</tr>
<tr>
<td>XLM-R</td>
<td>80.4</td>
<td>58.8</td>
<td>33.3</td>
<td>52.6</td>
<td>51.1</td>
<td>50.5</td>
<td>24.5</td>
<td>37.8</td>
</tr>
<tr>
<td>Filter</td>
<td>80.4</td>
<td>58.8</td>
<td>33.3</td>
<td>52.6</td>
<td>51.1</td>
<td>50.5</td>
<td>24.5</td>
<td>37.8</td>
</tr>
<tr>
<td>XTUNE</td>
<td>80.5</td>
<td>58.9</td>
<td>33.4</td>
<td>52.7</td>
<td>51.2</td>
<td>50.6</td>
<td>24.6</td>
<td>37.9</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>80.5</td>
<td>58.9</td>
<td>33.4</td>
<td>52.7</td>
<td>51.2</td>
<td>50.6</td>
<td>24.6</td>
<td>37.9</td>
</tr>
</tbody>
</table>

Table 16: MLQA results (F1/EM) for each language.

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>ar</th>
<th>de</th>
<th>es</th>
<th>hi</th>
<th>vi</th>
<th>zh</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>75.3</td>
<td>42.6</td>
<td>30.1</td>
<td>49.3</td>
<td>45.2</td>
<td>44.5</td>
<td>24.8</td>
<td>38.1</td>
</tr>
<tr>
<td>XLM-R</td>
<td>69.9</td>
<td>41.5</td>
<td>29.9</td>
<td>48.7</td>
<td>44.4</td>
<td>43.7</td>
<td>24.5</td>
<td>37.8</td>
</tr>
<tr>
<td>Filter</td>
<td>69.9</td>
<td>41.5</td>
<td>29.9</td>
<td>48.7</td>
<td>44.4</td>
<td>43.7</td>
<td>24.5</td>
<td>37.8</td>
</tr>
<tr>
<td>XTUNE</td>
<td>70.1</td>
<td>41.6</td>
<td>30.0</td>
<td>49.4</td>
<td>44.6</td>
<td>44.0</td>
<td>24.6</td>
<td>38.1</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>70.1</td>
<td>41.6</td>
<td>30.0</td>
<td>49.4</td>
<td>44.6</td>
<td>44.0</td>
<td>24.6</td>
<td>38.1</td>
</tr>
</tbody>
</table>

Table 17: TyDiQA-GolP results (F1) for each language.

<table>
<thead>
<tr>
<th>Model</th>
<th>en</th>
<th>ar</th>
<th>de</th>
<th>el</th>
<th>es</th>
<th>et</th>
<th>eu</th>
<th>fa</th>
<th>fi</th>
<th>fr</th>
<th>he</th>
<th>hi</th>
<th>hu</th>
<th>id</th>
<th>it</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>85.2</td>
<td>77.4</td>
<td>61.1</td>
<td>77.0</td>
<td>70.0</td>
<td>70.0</td>
<td>72.5</td>
<td>77.4</td>
<td>75.4</td>
<td>68.3</td>
<td>46.2</td>
<td>72.2</td>
<td>79.8</td>
<td>56.6</td>
<td>60.0</td>
<td>74.6</td>
</tr>
<tr>
<td>XLM-R</td>
<td>85.2</td>
<td>77.4</td>
<td>61.1</td>
<td>77.0</td>
<td>70.0</td>
<td>70.0</td>
<td>72.5</td>
<td>77.4</td>
<td>75.4</td>
<td>68.3</td>
<td>46.2</td>
<td>72.2</td>
<td>79.8</td>
<td>56.6</td>
<td>60.0</td>
<td>74.6</td>
</tr>
<tr>
<td>Filter</td>
<td>85.2</td>
<td>77.4</td>
<td>61.1</td>
<td>77.0</td>
<td>70.0</td>
<td>70.0</td>
<td>72.5</td>
<td>77.4</td>
<td>75.4</td>
<td>68.3</td>
<td>46.2</td>
<td>72.2</td>
<td>79.8</td>
<td>56.6</td>
<td>60.0</td>
<td>74.6</td>
</tr>
<tr>
<td>XTUNE</td>
<td>85.2</td>
<td>77.4</td>
<td>61.1</td>
<td>77.0</td>
<td>70.0</td>
<td>70.0</td>
<td>72.5</td>
<td>77.4</td>
<td>75.4</td>
<td>68.3</td>
<td>46.2</td>
<td>72.2</td>
<td>79.8</td>
<td>56.6</td>
<td>60.0</td>
<td>74.6</td>
</tr>
<tr>
<td>X-MIXUP</td>
<td>85.2</td>
<td>77.4</td>
<td>61.1</td>
<td>77.0</td>
<td>70.0</td>
<td>70.0</td>
<td>72.5</td>
<td>77.4</td>
<td>75.4</td>
<td>68.3</td>
<td>46.2</td>
<td>72.2</td>
<td>79.8</td>
<td>56.6</td>
<td>60.0</td>
<td>74.6</td>
</tr>
</tbody>
</table>

Table 19: Published as a conference paper at ICLR 2022