DiTTO : A Feature Representation Imitation Approach for Improving Cross-Lingual Transfer

Shanu Kumar **Abbaraju Soujanya** **Sandipan Dandapat**
Sunayana Sitaram **Monojit Choudhury**
Microsoft Corporation, India
{shankum,asoujanya,sadandap,susitara,monojitc}@microsoft.com

Abstract

Zero-shot cross-lingual transfer is promising, however has been shown to be sub-optimal, with inferior transfer performance across low-resource languages. In this work, we envision languages as domains for improving zero-shot transfer by jointly reducing the feature incongruity between the source and the target language and increasing the generalization capabilities of pre-trained multilingual transformers. We show that our approach, *DiTTO* , significantly outperforms the standard zero-shot fine-tuning method on multiple datasets across all languages using solely unlabeled instances in the target language. Empirical results show that jointly reducing feature incongruity for multiple target languages is vital for successful cross-lingual transfer. Moreover, our model enables better cross-lingual transfer than standard fine-tuning methods, even in the few-shot setting.

1 Introduction

Due to the emergence of pre-trained Massively Multilingual Transformers (MMTs) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2020), zero-shot cross-lingual transfer (Hu et al., 2020; Ruder et al., 2021; Lauscher et al., 2020; Ansell et al., 2021; Pfeiffer et al., 2022) has received significant attention in the NLP community. This approach originated due to the skew in resource distribution in languages (Joshi et al., 2020), with most languages of the world having a scarcity of labeled data. Zero-shot transfer involves fine-tuning the MMT with task-specific data in one or more source languages, followed by evaluation on target languages whose labeled instances are not used during fine-tuning. Accurate zero-shot transfer is crucially important for MMTs to be useful for low-resource languages.

The performance of MMTs drops in the following two cases - when the source and target languages exhibit dissimilar typological features, or when the size of pre-training data in the target language is limited (Lauscher et al., 2020; Ebrahimi et al., 2022). Two common techniques to improve zero-shot performance include few-shot cross-lingual transfer (Lauscher et al., 2020; Kumar et al., 2022) and the translate-train approach (Ruder et al., 2021; Ahuja et al., 2022). Several studies have been conducted comparing these approaches, of which (Ahuja et al., 2022) concludes that if the cost of machine translation is greater than zero, the optimal and lowest-cost performance is achieved with at least some manually labeled data (i.e. the few-shot method). Since annotating data is expensive for many languages (Dandapat et al., 2009; Sabou et al., 2012; Fort, 2016), we investigate improving cross-lingual zero-shot transfer using only unlabelled data in this paper.

Zero-shot Cross-lingual Transfer has been identified as an under-specified optimization problem (Wu et al., 2022). A majority of the solutions reports a high performance on the source language but fluctuating performance on target languages. Wu et al. (2022) use linear interpolation to prove that it is possible to obtain a subset of solutions which have optimal performance on both source and target languages. Furthermore, they also conclude that current optimization techniques cannot converge to this smaller subset of optimal solutions without the availability of labeled target language data. Aghajanyan et al. (2020) and Liu et al. (2021) have observed similar behavior in the zero-shot setup and hypothesize that sub-optimal zero-shot performance may be due to the degradation of generalizable representations of MMTs during the fine-tuning stage. This leads to the model trained on the source language not being able to generalize well to the target languages. MMTs have also been shown to be over-parameterized (Smith and Le, 2018; Kolesnikov et al., 2020; Zhang et al., 2021), which leads to memorizing the training data (source language) and achieving poor generalization during

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 385–406

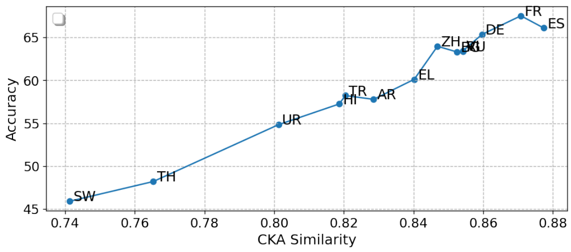


Figure 1: Relation between the zero-shot performance using mBERT, and CKA similarity between the source (EN) and various target languages in XNLI dataset.

cross-lingual transfer.

Similar to Deshpande et al. (2022), in our experiments, we also observe that once MMTs are fine-tuned on source languages, there is an incongruity between the features of the source and target languages, as shown in Figure 2. We speculate that the mismatch in the feature representation space causes problems in generalization. We also find that this mismatch strongly correlates with zero-shot performance as shown in Figure 1.

Furthermore, we hypothesize that this instability can be reduced either by finding solutions that can generalize well or learning to match the feature representations. Solutions (Zhang et al., 2018; Jiang et al., 2020) that have been used for improving generalization in other tasks can be considered, so that the model reaches to a better local minima. Sharpness-aware Optimization (SAM) (Foret et al., 2021) is one such technique that has been used to improve the generalization of language models (Bahri et al., 2022) and vision transformers (Chen et al., 2021) by smoothing the loss landscape for various adversarial tasks. SAM is used to generalize across domains, however, by treating languages as separate domains, we can apply SAM for generalizing across languages. While SAM looks promising, our experiments (cf. 7.3) showed that it does not guarantee optimal generalization at all times. We need to further reduce the incongruency between language features by aligning target language features to mimic the features of the source language. We propose ***DiTTO*** for improving cross-lingual transfer by source language **Di**rected adversarial **T**ransition of **T**arget language using sharpness aware **O**ptimization.

The key contributions of this work are: 1) Exhibiting the limitations of standard fine-tuning by unveiling the feature incongruity between source and target languages. 2) *DiTTO* enhances cross-lingual transfer by joint feature transformation of the multiple target languages to mimic the source. 3) *DiTTO* makes cross-lingual transfer cost-effective and efficient for distant (typologically different languages), resource-lean and unseen (not present in the pre-training data) languages. 4)*DiTTO* exhibits superior performance compared to augmenting the training data for either the source or the target language.

2 Related Work

Cross-Lingual Transfer: Since the inception of pre-trained MMTs, zero-shot learning has become popular for cross-lingual tasks. Recent works (Lauscher et al., 2020; Ebrahimi et al., 2022; Wu et al., 2022) have shown it to be sub-optimal for target languages which are either distant to the source language or have limited data during pre-training of the MMT. Some works (Wu and Dredze, 2020; Yu and Joty, 2021) have tried to improve the transfer using feature alignment from parallel data or bitexts (Zhang et al., 2020; Tiedemann, 2012) which is often expensive to obtain for many languages. To address this issue, *DiTTO* relies only on unlabeled data in the target languages. As pre-training size of the language affect transfer performance, adapter-based frameworks (Pfeiffer et al., 2020; Ansell et al., 2021) have been proposed for learning language and task representations for low-resource languages and languages that are unseen during pre-training. Though this framework is helpful for unseen languages, it provides limited gains for typologically dissimilar and high resource languages, and our method can easily be integrated with adaptors to further improve the transfer performance.

Improving Generalization: Deep neural networks such as MMTs are generally over-parameterized and fine-tuning leads to easy memorization of the labeled training data, does not always generalize well to other domains (Smith and Le, 2018; Kolesnikov et al., 2020; Zhang et al., 2021). Various methods have been proposed to improve the generalization like dropout (Srivastava et al., 2014), label smoothing (Müller et al., 2019), batch normalization (Ioffe and Szegedy, 2015), mixup (Zhang et al., 2018).

A few papers (Dziugaite and Roy, 2017; He et al., 2019; Jiang et al., 2020) have explored the connection between the flatness of minima and generalization gaps, showing flatter minima leads to better generalization. Recently, SAM has been proposed to find a smoother minima by minimizing the loss value and its sharpness. SAM has been shown to improve the generalization capabilities of vision transformers (Chen et al., 2021). Recently, Bahri et al. (2022) employed SAM in language models such as GPT-3 (Brown et al., 2020) and T5 (Raffel et al., 2020), showing significant improvements in generalization in English. In this work, we use SAM to improve the generalization across other languages. Another line of work (Aghajanyan et al., 2020; Liu et al., 2021) hypothesizes that inferior transfer is due to forgetting and degradation of feature representation from pre-trained MMTs when they are fine-tuned on the source language data. They propose to preserve the pre-trained features to improve the generalization using regularization and continual learning.

Unsupervised Domain Adaptation (UDA): Various studies have been proposed to reduce the domain shift to perform UDA by minimizing discrepancy distances such as Maximum Mean Discrepancy (MMD) (Long et al., 2015) and correlation alignment distance (Sun and Saenko, 2016). Adversarial-based feature alignment methods (Ganin and Lempitsky, 2015; Ganin et al., 2016; Long et al., 2018; Kurmi et al., 2019) have been one of the popular UDA methods where the domain discrepancy between the domains is reduced using an adversarial objective. In this work, we use Domain-Adversarial Neural Networks (DANN) (Ganin et al., 2016) for performing adversarial adaptation of languages.

3 Background

Training a Zero-Shot Model: In zero-shot cross-lingual transfer, we fine-tune an MMT on a source language and evaluate its performance on the target language, whose instances are not used during fine-tuning. To do this, we need a source language s and task-specific labeled dataset $\mathbb{L}_s = \{(x_i^s, y_i^s)\}_{i=1}^n$ with n examples. We use the provided MMT $\mathcal{M}$ as the encoder and fine-tune it along with the task-specific classifier $\mathcal{C}$ by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{train}}(\mathcal{M}, C) = \mathbb{E}_{(\mathbf{x}_i^s, \mathbf{y}_i^s) \sim \mathbb{L}_s} \mathcal{L}(C(\mathcal{M}(\mathbf{x}_i^s)), \mathbf{y}_i^s) \tag{1}$$

Sharpness-Aware Minimization (SAM): SAM seeks to find the parameter w such that even its neighborhood has seemingly similar low training loss $\mathcal{L}_{\text{train}}$ with minimal variation by optimizing the following objective:

$$\min_{w} \max_{||\epsilon||_2 \leq \rho} \mathcal{L}_{\text{train}}(w + \epsilon) \tag{2}$$

where ρ is the size of the neighborhood. Since, the exact solution of the inner maximization is hard to obtain, the authors of SAM propose a simple first order approximation:

$$\begin{aligned} \epsilon(\hat{w}) \approx \\ \arg\max_{||\epsilon||_2 \leq \rho} \mathcal{L}_{train}(w) + \epsilon^T \nabla_w \mathcal{L}_{\text{train}}(w) \\ = \rho \nabla_w \mathcal{L}_{\text{train}}(w) / ||\nabla_w \mathcal{L}_{\text{train}}(w)||_2 \end{aligned} \tag{3}$$

After computing $\hat{\epsilon}$, the parameter w is updated based on the the sharpness-aware gradient $\nabla_w \mathcal{L}_{\text{train}}(w)|_{w+\epsilon(\hat{w})}$.

Domain-Adversarial Neural Networks (DANN): DANN (Ganin et al., 2016) has been successful applied for many unsupervised domain adaptation tasks for minimizing the domain shift (Du et al., 2020; Long et al., 2018). DANN needs a labeled source domain dataset $\mathbb{L}_s = \{(x_i^s, y_i^s)\}_{i=1}^n$ with n examples and an unlabeled target domain dataset $\mathbb{U}_t = \{x_i^t\}_{i=1}^m$ with m examples. It consists of three modules: Encoder $\mathcal{E}$, Task-Specific Classifier $\mathcal{C}$, and Domain Discriminator $\mathcal{D}$. In a nutshell, DANN requires solving a two-player game where the first player is the Domain Discriminator $\mathcal{D}$, is trained to distinguish the target domain from the source domain, and the second player is the encoder $\mathcal{E}$, which is trained simultaneously to confuse the Discriminator $\mathcal{D}$ such that the encoder learns to generate domain invariant features. We minimize the task-specific classification loss $\mathcal{L}_{\mathcal{C}}$ using the source domain labeled dataset for optimizing the classifier $\mathcal{C}$ and encoder $\mathcal{E}$.

$$\mathcal{L}_{\mathcal{C}}(\mathcal{E}, \mathcal{C}) = \mathbb{E}_{(\mathbf{x}_i^s, \mathbf{y}_i^s) \sim \mathbb{L}_s} \mathcal{L}(\mathcal{C}(\mathcal{E}(\mathbf{x}_i^s)), \mathbf{y}_i^s) \tag{4}$$

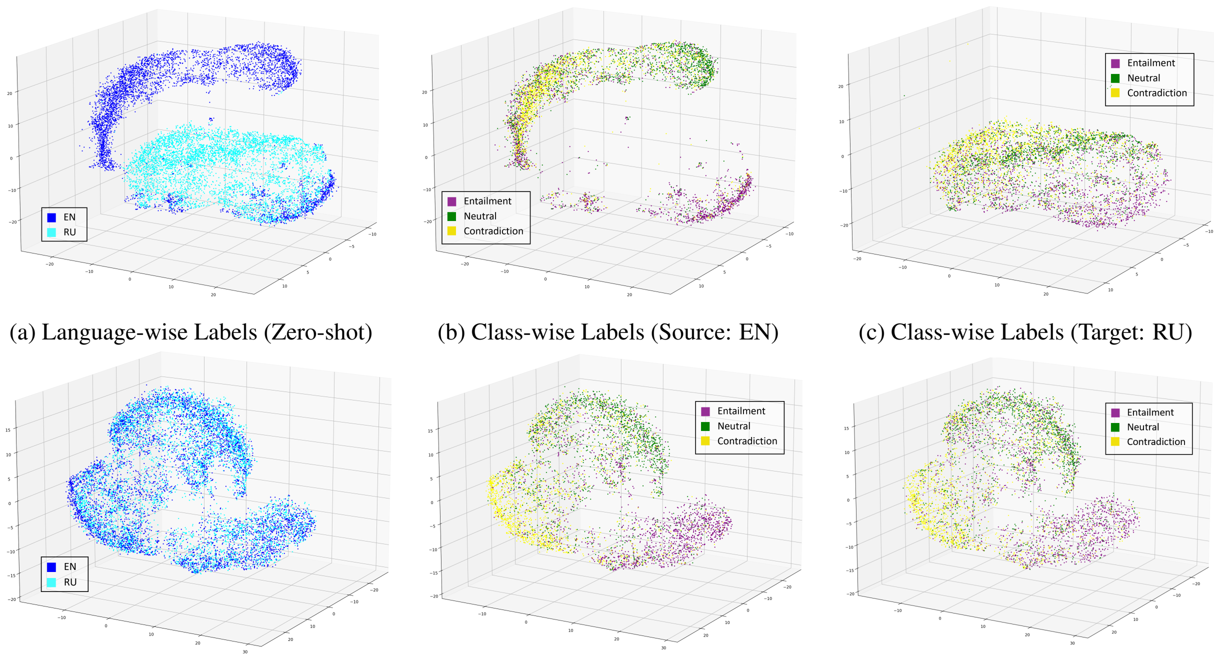
$\mathcal{D}$ is trained to predict the domains by minimizing domain classification loss:

$$\begin{aligned} \mathcal{L}_{\mathcal{D}}(\mathcal{E}, \mathcal{D}) = -\mathbb{E}_{\mathbf{x}_i^s \sim \mathbb{L}_s} \log[\mathcal{D}(\mathcal{E}(\mathbf{x}_i^s))] \\ -\mathbb{E}_{\mathbf{x}_j^t \sim \mathbb{U}_t} \log[1 - \mathcal{D}(\mathcal{E}(\mathbf{x}_j^t))] \end{aligned} \tag{5}$$

$\mathcal{L}_D$ is maximized for $\mathcal{E}$ so that $\mathcal{D}$ is not able to distinguish between the domains. The minimax optimization of DANN is defined as:

$$\begin{aligned} \min_{\mathcal{E}, C} \quad & \mathcal{L}_C(\mathcal{E}, C) - \lambda \mathcal{L}_{\mathcal{D}}(\mathcal{E}, \mathcal{D}) \\ \min_{\mathcal{D}} \quad & \mathcal{L}_{\mathcal{D}}(\mathcal{E}, \mathcal{D}) \end{aligned} \tag{6}$$

where λ is a hyper-parameter to control trade-off between classification and domain adversarial loss.



(a) Language-wise Labels (Zero-shot) (b) Class-wise Labels (Source: EN) (c) Class-wise Labels (Target: RU)

(d) Language-wise Labels (*DiTTO*) (e) Class-wise Labels (Source: EN) (f) Class-wise Labels (Target: RU)

Figure 2: 3D t-SNE visualization of the features from the last layer of fine-tuned mBERT on XNLI (S=1%).

		mBERT			XLM-R		
Dataset	$\|\mathbb{T}\|$	1%	10%	100%	1%	10%	100%
XNLI	14	10.3	12.3	15.5	8.3	10.3	11.3
MARC	5	14.8	18.1	20.3	4.5	8.8	9.9
AmNLI	10	24.8	32.9	41.2	29.9	39.2	45.1

Table 1: We have reported the mean of difference $\triangle$ between the zero-shot performance of all the target languages and source language for varying amount of the source language (EN) data used while fine-tuning. $|\mathbb{T}|$ is the number target languages available in the dataset.

4 Limitations of Zero-shot Learning

Inconsistent Cross-Lingual Transfer: We have reported the average difference (δ) in the zero-shot performance between the target and the source language in Table 1. We experiment with mBERT and XLM-R on XNLI (Conneau et al., 2018), AmNLI (Ebrahimi et al., 2022) and MARC (Keung et al., 2020) datasets to measure the average δ between zero-shot performance of the target and source language. Table 1 shows that XNLI and AmNLI having relatively higher δ due to diverse number of languages. We also notice that mBERT has a higher δ than XLM-R across all tasks except AmNLI, showing the importance of amount of pre-training size.

Feature Incongruity between Languages: We hypothesize that the inconsistent zero-shot performance is due to the mismatch in the feature representation space of the fine-tuned MMT on the source language. To verify that we visualize the target and source language feature representations learned using standard zero-shot training method using 3D t-SNE (Van der Maaten and Hinton, 2008) in Figure 2. In Figure a, there is clear distinction between the source (En) and target language (Ru) features. While in Figure 2b and 2c, the feature space for the entailment class is overlapping with the source language, but fairly distinct for the other two classes, this could be potential cause for inferior cross-lingual transfer.

We measure centered kernel alignment (CKA) (Kornblith et al., 2019) between the source and the target language feature representations to quantify the incongruity. In Figure 1. We have plotted the CKA similarity with the zero-shot performance across all the languages. The plot suggests that there is a strong correlation between CKA and zero-shot performance, with Pearson and Spearman correlation coefficients as 0.98 and 0.96, respectively establishing our hypothesis.

5 Unveiling *DiTTO*

Typological similarity and incongruency between feature representations lead us to envision different languages as domains. As discussed in the previous section, DANN is useful in minimizing the domain shift across domains using only unlabeled data in the target domain. We propose to perform adversarial adaptation of the target language features for transforming the same towards the source language feature distribution.

We have a set of target languages $\mathbb{T}$ with each target language t having dataset $\mathbb{U}_t = \{x_i^t\}_{i=1}^{\mathcal{T}}$ with $\mathcal{T}$ unlabeled examples and an unlabeled set $\mathbb{U}_s = \{x_i^s\}_{i=1}^{\mathcal{S}}$ with $\mathcal{S}$ examples in the source language. In DANN, there is one target domain, whereas in our case we have a set of target languages $\mathbb{T}$ and we hypothesize and empirically show that performing adaptation for each language separately may cause degradation in other target languages, as seen in Table 5. Hence, we propose *DiTTO* where we jointly perform adaptation across all target languages.

DiTTO consists of an MMT $\mathcal{M}$ for encoding the features, a task-specific classifier $\mathcal{C}$ and Language Discriminators $\mathcal{D}^L = \{\mathcal{D}_t^L\}_{i=1}^{|\mathbb{T}|}$. We train these modules using two losses: task-specific classification loss $\mathcal{L}_\mathcal{C}$, defined in the Equation (1) and language discrimination loss $\mathcal{L}_\mathrm{L}$ for distinguishing the target and source language.

As we have $|\mathbb{T}|$ discriminators, we randomly sample a target language t from a prior distribution $p(\mathbb{T})$ at each training step and train the discriminator $\{\mathcal{D}_t^L\}$ to accurately distinguish target t and source language using the following loss:

$$\begin{aligned}\mathcal{L}_L(\mathcal{M}, \mathbf{D}_t^L) = -\mathbb{E}_{\mathbf{x}_i^s \sim \mathbb{U}_s} \log[b_t^L(\mathcal{M}(\mathbf{x}_i^s))] \\ -\mathbb{E}_{\mathbf{x}_j^t \sim \mathbb{U}_t} \log[1 - \mathcal{D}_t^L(\mathcal{M}(\mathbf{x}_j^t))]\end{aligned} \tag{7}$$

We maximize the above loss $\mathcal{L}_L(\mathcal{M}, \mathcal{D}_t^L)$ for confusing the language discriminator $\mathcal{D}_t^L$ to transform the target features towards the source language.

In our initial experiments (reported in Table 4), we observed some instability due to adversarial adaptation (Mao et al., 2017; Xing et al., 2021). We propose optimizing the task-specific loss $\mathcal{L}_C$ using SAM so that it may generalize to the target languages, improving the stability during adversarial adaptation. We directly fine-tune the MMT $\mathcal{M}$ on the source language labeled dataset $\mathcal{D}_s^l$ by minimizing Equation (1) using SAM. Following DANN and SAM, the final optimization objective of *DiTTO* can be defined as:

$$\min_{\mathcal{M},C} \max_{||\epsilon||_2 \leq \rho} \mathcal{L}_C(\hat{\mathcal{M}}, \hat{C}) - \lambda \mathbb{E}_{t \sim p(\mathbb{T})} \mathcal{L}_L(\mathcal{M}, \mathcal{D}_t^L) \tag{8}$$

$$\min_{\mathcal{D}_L} \quad \mathbb{E}_{t \sim p(\mathbb{T})} \mathcal{L}_L(\mathcal{M}, \mathcal{D}_t^L) \tag{9}$$

where $\hat{\mathcal{M}}$, $\hat{C}$ are the updated parameters using ϵ.

6 Experimental Setup

6.1 Datasets

We evaluate our method on three benchmark datasets consisting of languages from various language families, to ensure better cross-lingual transfer evaluation. **XNLI** dataset (Conneau et al., 2018) consists of translated dataset in 14 languages from English. The task requires any model to predict whether the premise entails, contradicts, or neutral to the given hypothesis. AmericasNLI (**AmNLI**) dataset (Ebrahimi et al., 2022) is an extension of XNLI to 10 indigenous languages of the Americas, which are even unseen during pre-training of XLM-R and mBERT. Multilingual Amazon Review Corpus (**MARC**) dataset (Keung et al., 2020) is a large-scale dataset consisting of Amazon reviews for text classification in 6 languages. We use the review text and title to predict its star rating.

6.2 Baselines and *DiTTO* Variants

In the **Baseline** experiments, we fine-tune MMTs on labeled data of the source language using Equation (1). In the vanilla ***DiTTO*** setup, we use all the target languages available in the dataset. In the vanilla setup, we want to assign a higher probability to those target languages with a lower zero-shot performance from the Baseline method. We defined the prior distribution $p(\mathbb{T})$ of target languages as follows:

$$\Delta_t = \max(\mathcal{Z}(s) - \mathcal{Z}(t), 0) \tag{10}$$

$$p(t) = \delta_t + \sigma_{\Delta_t} \tag{11}$$

where, $\mathcal{Z}$ is the zero-shot performance from the Baseline method, Δ_t is the non-negative delta between the source and target language, and σ_δ is the standard deviation of the Δ_t across all the target languages.

DiTTO (UNF) is a variant of vanilla *DiTTO* in which we set the prior distribution $p(\mathbb{T})$ to be uniform across all the target languages. ***DiTTO* (t)** is a single target language variant of *DiTTO* where only one target language t is used during training. ***DiTTO*-LA** does not perform adaptation of the target languages, however optimization is done using SAM on the source language labeled data. ***DiTTO*-SAM** performs language adaptation without SAM.

6.3 Training Details

We conduct all of our experiments using mBERT (*bert-base-multilingual-cased*) and XLM-R (*xlm-roberta-base*). We use a batch size of 32 and a maximum sequence length of 128 across all the datasets. We fine-tune for $\{15, 20, 25\}$, $\{3, 5, 7\}$, $\{2, 3, 5\}$ epochs while using 1%, 10% and 100% of the source language data respectively. We use the AdamW (Loshchilov and Hutter, 2018) optimizer with linear scheduler and learning rate as 1e-5 for the encoder and classifier and 5e-5 for the discriminator. We set the λ hyper-parameter as 1 for all the experiments. We run experiments for each hyper-parameter and report the best average accuracy on three random seeds.

7 Results

In this section, we describe the results of several experiments to analyze the *DiTTO* method and compare its performance with the Baseline in the zero-shot setting. In order to justify the robustness of our method, we conduct experiments with the varying amount of source language data. In our experiments, EN is the default source language and we categorize target languages as follows: 1. *Distant*: languages that are typologically dissimilar to the source language 2. *Low-resource*: languages that have scarcity of data for pre-training 3. *Unseen*: languages that were not included in the pre-training data of MMT. Furthermore, we compare the techniques in the few-shot setting with few labeled examples in target languages. Then, we perform a thorough ablation study and analyze various variants of *DiTTO* . Finally, we show evidence in the form of congruity between the source and target language feature representations and t-SNE visualization in support of our hypothesis.

7.1 Zero-shot Transfer Results

Performance across datasets: In Table 2, we have reported the relative gains from *DiTTO* for zero-shot setting averaged across all the languages over the baseline method using 1%, 10% and 100% of the source language data. We observe the gains are positive (upto 23.05%) across all the training configurations. The gains are much higher for mBERT than XLM-R due to lower cross-lingual transfer in mBERT except the AmNLI dataset. The relative gains start to decrease with the increased amount of the source language data S on all the datasets except AmNLI, where the gains remains consistent for higher values of S (10% and 100%).

Performance across Seen Target Languages: We have reported the absolute gains of *DiTTO* in Figure 3 on XNLI using XLM-R We observe positive gains from *DiTTO* for all the target languages, with much larger gains especially on the low-resource and distant languages compared to the Baseline model. Similar to the earlier observation in Table 2, the gains starts to decrease across target languages as we increase the amount of the source language data.

	mBERT			XLM-R		
Dataset	1%	10%	100%	1%	10%	100%
XNLI	**23.05**	6.58	2.10	13.57	4.10	2.71
AmNLI	11.61	**19.72**	15.10	17.95	**19.87**	**19.09**
MARC	12.28	15.40	**19.03**	5.61	3.04	2.41

Table 2: Relative gains (in %) of *DiTTO* over Baseline. grey

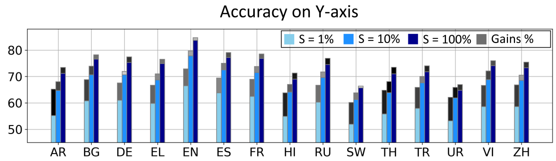


Figure 3: Absolute gains (darker shades of grey denotes higher gains) from *DiTTO* for XLM-R on XNLI dataset. Magnified view available in Figure 7 in Appendix.

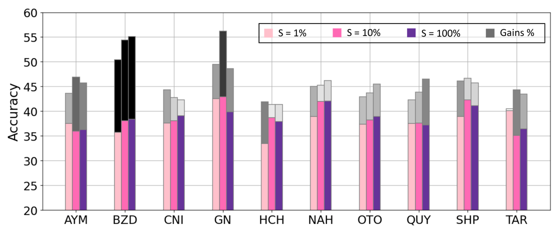


Figure 4: Absolute gains (darker shades of grey denotes higher gains) from *DiTTO* for XLM-R on AmNLI.

Performance across Unseen Target Languages: To measure the impact of *DiTTO* on unseen languages, we report the absolute gains from *DiTTO* on XLM-R on the AmNLI dataset in Figure 4. We have provided a similar analysis for mBERT in Figure 8 of the Appendix. The gains from *DiTTO* are consistent across all unseen languages. We observe that the gains are higher for languages with better Baseline performances, which is in contrast to trends on seen languages. For unseen languages, we do not observe the trend of diminishing gains with an increase in the source language data. If we compare the gains on AmNLI with the XNLI dataset, we notice *DiTTO* providing on average 1.7 times higher gains across all the configurations.

7.2 Few-shot Transfer Results

It can be argued that the gains from *DiTTO* in zero-shot setting can be achieved using few-shot cross-lingual transfer. Therefore, we conduct experiments in the few-shot setting by adding k labeled instances in each of the target languages to measure capabilities of *DiTTO* when some labeled data is available along with unlabeled data. In Figure 5, we have reported the accuracy and relative gains[1] using Baseline and *DiTTO* on MARC dataset. We have provided a similar analysis for AmNLI dataset in Figure 11 in the Appendix.

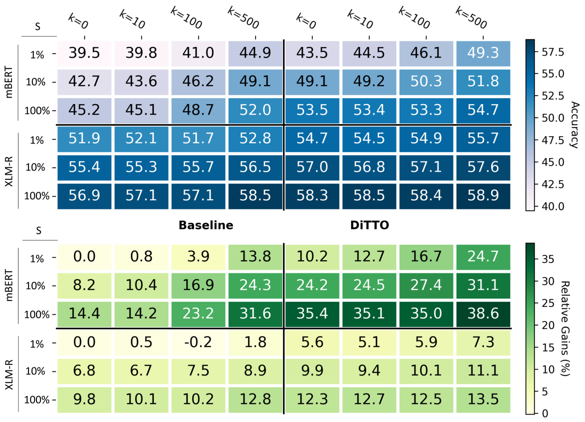


Figure 5: Accuracy/relative gains on MARC dataset.

The heat maps in Figure 5 show that while XLM-R has better accuracy than mBERT for both Baseline and *DiTTO* setup, but the gains (both absolute and relative) on mBERT for both methods are higher compared to XLM-R. We also notice that by increasing either the source or the target language data, performance for both Baseline and DiTTO increased and hence, we will compare the gains of *DiTTO* with Baseline in these two dimensions.

Impact of Target Language Labeled Data: We observe that when we fix the amount of source language data and increase the value of k, the gains from *DiTTO* are higher than Baseline. Also, the gains from *DiTTO* on k=0 is comparable with the gains of baseline on k=500. In AmNLI, the gains from *DiTTO* for lower values of k are quite high compared to baseline, while for higher values of k Baseline performance for XLM-R is comparable to DiTTO.

Impact of Source Language Labeled Data: We also noticed that by fixing the value of k and increasing the size of source language data S, there is an increase in gains for both methods on MARC. However, the increase in gains from *DiTTO* is much higher than Baseline. At the same time, on the AmNLI dataset consisting of unseen target languages, the gains is much smaller with the increase in S (cf. Appendix).

[1]The relative gain is calculated with respect to the accuracy obtained by the Baseline method on $S = 1\%$ and $k = 0$.

Chinese as Source Language: To measure the effectiveness of *DiTTO* across different source languages, we conduct zero-shot experiments considering Chinese (ZH) as the source language on the MARC dataset. We have reported the average accuracy across all the languages in Table 3. *DiTTO* provides consistent gains over the Baseline method across all the training configurations, comparatively higher gains than EN as the source language.

	mBERT			XLM-R		
Dataset	1%	10%	100%	1%	10%	100%
Baseline	32.88	39.48	42.68	45.83	50.86	51.38
DiTTO	**39.09**	**46.90**	**50.82**	**51.84**	**53.34**	**55.27**
RG(%)	**19.45**	**20.80**	**20.67**	**13.31**	5.34	8.12

Table 3: We have reported the zero-shot accuracy averaged across all languages with **ZH** as the source language data on MARC dataset. RG denotes the relative gains averaged across all the languages from using *DiTTO* over Baseline.

Performance and Cost Trade-off: *DiTTO* is seven times more cost-effective in terms of both source and target language data. We validate this by plotting the accuracy from both methods against the cost incurred while collecting the labeled data for fine-tuning. For detailed analysis refer to the section B in the Appendix.

7.3 Ablations and Variants Analysis

Ablation Study: Here we scrutinize the contributions from adaptation of target languages and optimization with SAM. We report the zero-shot relative gains in Table 4 by ablating each of these components. We observe that removing any component reduces the performance for most of the training configurations, indicating that both target language adaptation and optimization have a contribution in achieving better results. We also observe that removing SAM (*DiTTO* - SAM) leads to unstable performances on XNLI and AmNLI datasets with negative relative gains on AmNLI (S=1%) for both MMTs, and on XNLI (S=10%) for mBERT, showing instability caused in adversarial training (Mao et al., 2017; Xing et al., 2021). Removing target language adaptation (*DiTTO*-LA) reduces the relative gains by a significant margin, showing the importance of adaptation of target language features. It performs similar to *DiTTO* on XNLI (S=1%) dataset using mBERT, demonstrating just optimization using SAM can also improve cross-lingual transfer. The performance of *DiTTO* - SAM, it is often higher compared to *DiTTO* - LA, which indicates that Language Adaptation is a much more crucial for improving cross-lingual transfer.

	Method	mBERT 1%	mBERT 10%	mBERT 100%	XLM-R 1%	XLM-R 10%	XLM-R 100%
XNLI	*DiTTO*	**23.05**	**6.58**	**2.10**	**13.57**	**4.10**	**2.71**
	DiTTO - SAM	8.84	-0.36	1.80	6.04	1.81	2.02
	DiTTO - LA	22.25	3.89	2.02	7.43	2.81	1.74
MARC	*DiTTO*	**12.28**	**15.40**	**19.03**	**5.61**	**3.05**	**2.41**
	DiTTO - SAM	8.64	9.90	14.98	2.89	-0.13	2.27
	DiTTO - LA	5.5	1.54	2.20	4.02	0.35	-0.54
AmNLI	*DiTTO*	**11.61**	**19.72**	**15.10**	**17.95**	**19.87**	**19.09**
	DiTTO - SAM	-3.85	14.35	14.52	-11.88	14.81	15.89
	DiTTO - LA	7.21	5.17	-1.00	7.57	7.58	9.33

Table 4: Ablation Study: Zero-shot relative gains (in %) averaged across all the languages over Baseline.

Single vs Multiple Target Language Adaptation: In the base setup of *DiTTO* , we propose to perform an adaptation of all the target languages available in the dataset. We conduct zero-shot experiments with a single target language variant *DiTTO* (t) to validate our assumption. In Table 5, we observe that the single language variant provides similar gains as the vanilla *DiTTO* in the selected language t. However, often there is very little/no improvement observed in languages other than t. *DiTTO* (JA) and *DiTTO* (ZH) under-perform than Baseline for most of the languages.

Method	EN	DE	ES	FR	JA	ZH	AVG
Baseline	54.3	42.3	42.3	43.8	36.8	32.3	42.0
DiTTO (DE)	55.7	**48.2**	42.4	44.0	**36.5**	34.9	43.8
DiTTO (ES)	55.5	42.5	**45.9**	45.7	**36.4**	34.8	43.5
DiTTO (FR)	55.2	44.9	43.7	**46.4**	**36.5**	35.6	43.7
DiTTO (JA)	55.2	**40.9**	**41.3**	**42.5**	**38.1**	33.9	42.0
DiTTO (ZH)	**55.8**	**41.8**	**41.9**	**42.9**	**35.1**	**40.2**	43.0
DiTTO (UNF)	55.0	46.4	**46.2**	45.9	**38.5**	40.4	**45.4**
DiTTO	55.3	**47.0**	45.1	**46.2**	**38.6**	**40.7**	**45.5**

Table 5: Accuracy for single and multiple target language variants of *DiTTO* on MARC (S=1%, mBERT).

Target Language Prior Distribution: In *DiTTO* with multiple target language variant, the prior language distribution $p(\mathbb{T})$ is used to sample a target language for adaptation. To measure the importance of prior distribution, we experiment with two variants: (i) sampling based on the zero-shot performance of the Baseline method, which is used in the base setup of *DiTTO* and (ii) *DiTTO* (UNF) - with uniform sampling. Both the variants outperform Baseline with similar gains as shown in Table 5. In the vanilla *DiTTO* , where languages with lower zero-shot performance have a higher likelihood during sampling, provides better gains on these selected languages compared to the *DiTTO* (UNF).

Task-Adaptive Pre-training (TAPT): The Baseline method does not utilize the available unlabeled data in the target languages, whereas *DiTTO* uses the unlabeled data to improve the performance across all the target languages. Recently task-adaptive pre-training (TAPT) (Gururangan et al., 2020; Hossain et al., 2020; Caselli et al., 2021) using unlabeled task-specific data has been shown to improve the performance for pre-trained language models across multiple tasks. However, TAPT has yet to be evaluated in a multilingual setting.

To make a fair comparison, we have compared our proposed method with another baseline using unlabelled data, we shall refer this as *Baseline (TAPT)*. *TAPT* uses continued pre-training on the unlabeled target language data and fine-tuning is performed using the source language labelled dataset. We have reported the comparison between the new baseline method in Table 6. The *Baseline (TAPT)* method outperforms the Baseline method where unlabeled data is not used in the source language (EN); however, it regresses for all the target languages. We hypothesize that the TAPT method generally improves the performance of the language used during fine-tuning. Still, it suffers from similar issues which the Baseline method suffers, such as low feature congruity in the fine-tuned features between the languages. *DiTTO* , which does not suffer the feature incongruity issue, outperforms *Baseline (TAPT)* for all the languages.

Method	EN	DE	ES	FR	JA	ZH
Baseline	56.40	56.02	53.29	52.15	49.77	48.05
Baseline (TAPT)	**57.80**	55.58	52.14	51.70	49.60	45.71
DiTTO	**61.28**	**59.06**	**54.87**	**55.50**	**53.29**	**51.01**

Table 6: Comparison of Baseline and *DiTTO* methods with the new Baseline method using Task-Adaptive Pre-training (TAPT) on MARC (S=1%, XLM-R).

7.4 Congruity in Feature Representation

As shown in Figure 1 earlier that the zero-shot performance and feature congruity between the source and target languages are highly correlated. To validate our hypothesis that increasing the congruity between the features (via language adaptation) will improve the performance, we have plotted the increment in CKA similarity from *DiTTO* over the Baseline method in Figure 6. We observe increment in CKA similarity across all the languages using *DiTTO* , which is comparatively higher for distant or low-resource target languages. We also visualize the t-SNE projection of the feature representations of the source and the target languages in Figure 2. It is difficult to distinguish between both languages in this figure, showcasing the quality of language adaptation.

8 Discussion and Conclusion

In this work, we propose a novel method to improve the cross-lingual transfer capability of pre-trained MMTs. We find that zero-shot performance is correlated with incongruency between the features of source and target languages. Experiments show that our proposed method *DiTTO* outperforms the standard fine-tuning approach across multiple setups. In general, the gains from our method are higher on the models (as in mBERT) with less cross-lingual transfer. AmNLI consists only of languages that were not present in the pre-trained MMTs leading to similar transfer performance to the Baseline method. *DiTTO* improves cross-lingual transfer using the pre-training features, hence the gains from *DiTTO* are similar on both mBERT and XLM-R. Due to a similar reason, the relative gains for unseen languages do not follow the trend observed on seen languages, where the gains are higher for languages with the lower cross-lingual transfer. We find higher relative gains on unseen and low-resource languages, followed by distant languages. We also notice that the cross-lingual transfer improves with the amount of source language data S for seen languages. In contrast, for unseen languages, improvements are limited. Due to this, the gains from *DiTTO* start to decrease for high values of S for seen languages but remain significant for unseen languages.

Our method provides similar gains using only unlabeled data compared to the fine-tuned Baseline model (using 500 instances for each target language). Our ablation study shows that both LA and SAM are essential components of *DiTTO*, with LA being the primary contributor to the gains. Experiments show that single language adaptation

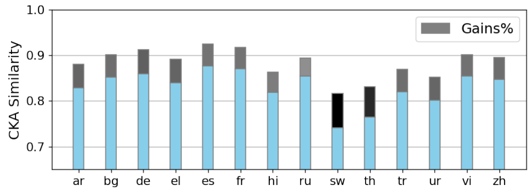


Figure 6: Gains in CKA similarity (between features of source and target language) from *DiTTO* over the Baseline method using mBERT on XNLI (S=10%).

improves on that corresponding target language but may regress on other languages as the feature may remain incongruent to the source. However, *DiTTO* that adapts to multiple target languages performs best. *DiTTO* tries to exploit the pre-training knowledge for improving the cross-lingual transfer, however few promising works such as adaptors (Pfeiffer et al., 2020; Ansell et al., 2021) have been proposed to improve the pre-training features for low-resource and unseen languages. However, task specific adaptors trained on the source language will also face the issue of incongruity in the feature representations. Hence, adaptors will not improve the cross-lingual transfer, but only improves the pre-trained features. We plan to extend our method towards integrating with adaptors to take advantage of pre-training features and improve performance.

9 Limitations

Unlabeled data in the target language is essential for the proposed method *DiTTO* for improving cross-lingual transfer. Obtaining unlabeled data can be challenging for specific tasks where the proposed approach may not be applicable. However, we recommend using the *DiTTO*-LA variant for these scenarios. Another limitation of *DiTTO* is that it requires all the target languages to be present during the fine-tuning stage to obtain the performances mentioned in our work, which might not be viable for all the tasks. Nevertheless, the gains from *DiTTO* may transfer to the new target languages if these languages are typologically similar to the target languages used during the fine-tuning of *DiTTO* . In the vanilla setup of *DiTTO* , the prior language probability depends upon the zero-shot accuracy using the Baseline method, which requires a validation or test dataset in each target language. This dependency may limit its application. However, *DiTTO* (UNF) can be used for obtaining similar gains if the validation sets are not available.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*.

Kabir Ahuja, Monojit Choudhury, and Sandipan Dandapat. 2022. On the economics of multilingual few-shot learning: Modeling the cost-performance trade-offs of machine translated and manual data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1369–1384, Seattle, United States. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. Mad-g: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781.

Dara Bahri, Hossein Mobahi, and Yi Tay. 2022. Sharpness-aware minimization improves language model generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7360–7371.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. 2021. When vision transformers outperform resnets without pre-training or strong data augmentations. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. 2009. Complex linguistic annotation – no easy way out! a case from Bangla and Hindi POS labeling tasks. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 4019–4028.

Gintare Karolina Dziugaite and Daniel M Roy. 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

Karën Fort. 2016. Collaborative annotation for reliable natural language processing: Technical and sociological aspects.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Haowei He, Gao Huang, and Yang Yuan. 2019. Asymmetric valleys: Beyond sharp and flat local minima. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2020. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR.

Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. "diversity and uncertainty in moderation" are the key to data selection for multilingual few-shot transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1042–1055, Seattle, United States. Association for Computational Linguistics.

Vinod Kumar Kurmi, Shanu Kumar, and Vinay P Namboodiri. 2019. Attending to discriminative certainty for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 491–500.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71, Online. Association for Computational Linguistics.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. 2012. Crowdsourcing research opportunities: Lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '12, New York, NY, USA. Association for Computing Machinery.

Samuel L. Smith and Quoc V. Le. 2018. A bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Shijie Wu, Benjamin Van Durme, and Mark Dredze. 2022. Zero-shot cross-lingual transfer is underspecified optimization. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 236–248.

Yue Xing, Qifan Song, and Guang Cheng. 2021. On the algorithmic stability of adversarial training. *Advances in Neural Information Processing Systems*, 34:26523–26535.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Tao Yu and Shafiq Joty. 2021. Effective fine-tuning methods for cross-lingual adaptation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8492–8501.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

A Data Statistics

We have provided the statistics of training and test data after removing any duplicates in each of the target languages for all the datasets in Tables 7, 8, and 9.

B Performance and Cost Trade-off

From the above results, it seems that *DiTTO* is more cost-effective in terms of both source and target language data We validate this by plotting the accuracy from both methods against the cost incurred while collecting the labeled data for fine-tuning. Assuming there is no cost associated with collecting unlabeled data, we define the cost $\mathbb{C}$ for building a fine-tuning dataset as follows:

$$\mathbb{C} = c_s * n_s^l + c_s * c_{t/s} * k * |\mathbb{T}| \qquad (12)$$

where c_s is the cost of obtaining one instance labeled in the source language and we assume it to be 3 cents considering EN as the source language. $c_{t/s}$ is the relative cost of obtaining labeled data in target language compared to the source language. We use Gaussian Process Regression with a dot product kernel for modeling performance with cost. In Figure 12 and 13, we plot the accuracy for various values of $c_{t/s}$ against the total cost incurred using mBERT on the MARC dataset, we observe a convex curve with increasing curvature as the value of $c_{t/s}$ increases. From the plot, we can see that higher accuracy can be achieved using *DiTTO* than Baseline at the same cost for all the values of $c_{t/s}$, showing the cost-saving nature of *DiTTO* with average savings of 7 times.

ISO	Language	Train	Test	XLM-R Group	mBERT Group
AR	Arabic	392403	5010	Distant	Distant
BG	Bulgarian	392335	5010	Distant	Distant
DE	German	392440	5010	Similar	Similar
EL	Greek	392331	5010	Distant	Distant
EN	English	392568	5010	Source	Source
ES	Spanish	392405	5010	Similar	Similar
FR	French	392405	5010	Similar	Similar
HI	Hindi	392356	5010	Distant	Low-Resource
RU	Russian	392318	5010	Similar	Similar
SW	Swahili	391819	5010	Low-Resource	Low-Resource
TH	Thai	392480	5010	Distant	Low-Resource
TR	Turkish	392177	5010	Distant	Distant
UR	Urdu	388826	5010	Low-Resource	Low-Resource
VI	Vietnamese	392416	5010	Distant	Distant
ZH	Chinese	392251	5010	Distant	Distant

Table 7: In this table, we have reported the target language categories and statistics of training and test data available in each language for XNLI dataset.

ISO	Language	Train	Test	XLM-R Group	mBERT Group
AYM	Aymara	743	750	Unseen	Unseen
CNI	Asháninka	658	750	Unseen	Unseen
BZD	Bribri	743	750	Unseen	Unseen
GN	Guaraní	743	750	Unseen	Unseen
NAH	Nahuatl	376	738	Unseen	Unseen
OTO	Otomí	222	748	Unseen	Unseen
QUY	Quechua	743	750	Unseen	Unseen
TAR	Rarámuri	743	750	Unseen	Unseen
SHP	Shipibo-Konibo	743	750	Unseen	Unseen
HCH	Wixarika	743	750	Unseen	Unseen

Table 8: In this table, we have reported the target language categories and statistics of training and test data available in each language for AmNLI dataset.

ISO	Language	Train	Test	XLM-R Group	mBERT Group
DE	German	199877	4993	Similar	Similar
EN	English	199891	4998	Source	Source
ES	Spanish	199726	4986	Similar	Similar
FR	French	199612	4986	Similar	Similar
JA	Japanese	199845	4995	Distant	Distant
ZH	Chinese	197418	4903	Distant	Distant

Table 9: In this table, we have reported the target language categories and statistics of training and test data available in each language for MARC dataset.

	$S = 1\%$			$S = 10\%$			$S = 100\%$		
Language	Baseline	*DiTTO*	RG	Baseline	*DiTTO*	RG	Baseline	*DiTTO*	RG
				XLM-R					
EN	67.84	69.00	1.68	78.08	79.24	1.48	83.83	82.59	-1.48
AYM	37.47	43.60	16.37	36.00	46.93	30.37	36.27	45.73	26.10
BZD	35.73	50.40	41.04	38.13	54.40	42.66	38.40	55.07	43.40
CNI	37.60	44.27	17.73	38.13	42.80	12.24	39.07	42.27	8.19
GN	42.46	49.53	16.67	42.86	56.21	31.15	39.92	48.60	21.74
HCH	33.51	41.92	25.10	38.72	41.39	6.90	37.92	41.39	9.15
NAH	38.89	44.99	15.68	42.01	45.26	7.74	42.14	46.21	9.65
OTO	37.43	42.91	14.64	38.24	43.72	14.34	38.90	45.45	16.84
QUY	37.47	42.27	12.81	37.60	43.87	16.67	37.20	46.53	25.09
SHP	38.93	46.13	18.49	42.27	46.67	10.41	41.07	45.73	11.36
TAR	40.05	40.45	1.00	35.11	44.33	26.24	36.45	43.52	19.41
AVG	37.95	44.65	17.95	38.91	46.56	19.87	38.73	46.05	19.09
				mBERT					
EN	62.53	64.83	3.67	71.20	73.05	2.61	81.18	79.64	-1.89
AYM	38.27	44.93	17.42	38.93	47.07	20.89	39.33	47.07	19.66
BZD	34.80	44.00	26.44	37.47	45.60	21.71	42.13	45.60	8.23
CNI	37.60	39.87	6.03	37.47	47.47	26.69	40.00	44.93	12.33
GN	40.19	46.86	16.61	38.85	49.80	28.18	41.79	51.67	23.64
HCH	34.98	40.85	16.79	36.98	45.79	23.83	39.92	44.59	11.71
NAH	40.79	44.72	9.63	42.28	46.07	8.97	43.90	48.92	11.42
OTO	38.10	38.64	1.40	37.43	41.58	11.07	37.97	44.39	16.90
QUY	38.67	39.47	2.07	36.53	42.80	17.15	38.00	43.87	15.44
SHP	38.40	40.53	5.56	40.13	46.93	16.94	41.73	46.67	11.82
TAR	35.91	40.99	14.13	36.85	44.86	21.74	35.65	42.72	19.85
AVG	37.77	42.09	11.61	38.29	45.80	19.72	40.04	46.04	15.10

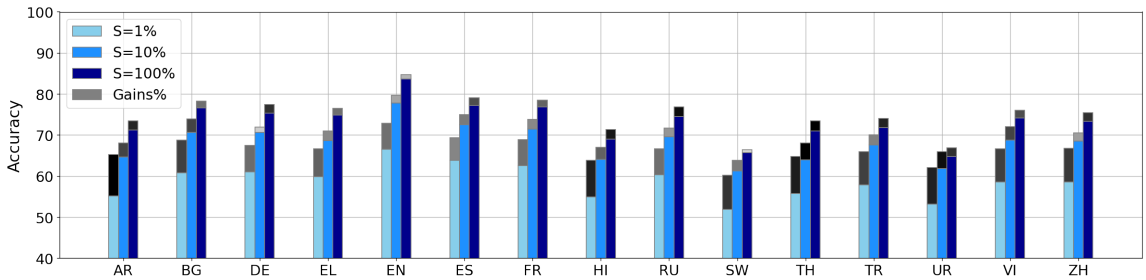
Table 10: We have reported the accuracy and relative gains using XLM-R and mBERT on AmNLI dataset. The average relative gain is denotes the average gains across all the languages except the source EN.

	$S = 1\%$			$S = 10\%$			$S = 100\%$		
Language	Baseline	*DiTTO*	RG	Baseline	*DiTTO*	RG	Baseline	*DiTTO*	RG
				XLM-R					
AR	55.25	65.19	17.99	64.69	67.96	5.06	71.16	73.39	3.14
BG	60.78	68.84	13.27	70.56	73.87	4.70	76.59	78.16	2.06
DE	60.98	67.50	10.70	70.64	71.92	1.81	75.33	77.37	2.70
EL	59.78	66.75	11.65	68.72	71.02	3.34	74.91	76.47	2.08
EN	66.49	72.91	9.67	77.80	79.72	2.46	83.71	84.65	1.12
ES	63.77	69.44	8.89	72.46	74.99	3.50	77.17	79.12	2.53
FR	62.51	68.86	10.15	71.52	73.77	3.15	76.85	78.50	2.16
HI	54.85	63.81	16.34	64.07	67.05	4.64	68.98	71.32	3.39
RU	60.32	66.67	10.52	69.62	71.66	2.92	74.49	76.93	3.27
SW	51.86	60.16	16.01	61.34	63.75	3.94	65.67	66.39	1.09
TH	55.81	64.75	16.02	63.89	68.14	6.65	70.96	73.45	3.52
TR	57.88	65.89	13.83	67.62	69.98	3.48	71.82	74.03	3.09
UR	53.17	62.12	16.82	61.94	65.91	6.41	64.83	66.95	3.26
VI	58.56	66.65	13.80	68.82	72.12	4.79	74.05	75.99	2.61
ZH	58.58	66.79	14.00	68.46	70.52	3.00	73.25	75.43	2.97
Average	58.15	65.96	13.57	67.45	70.19	4.10	72.57	74.54	2.71
				mBERT					
AR	47.09	56.75	20.52	57.78	62.87	8.81	63.07	65.21	3.39
BG	50.00	60.26	20.52	63.31	66.47	4.98	68.78	68.50	-0.41
DE	49.44	60.10	21.56	65.35	67.92	3.94	70.00	72.02	2.88
EL	48.70	59.08	21.31	60.12	64.63	7.50	65.91	66.99	1.64
EN	57.17	64.87	13.48	72.00	74.97	4.13	81.34	82.67	1.64
ES	50.12	62.38	24.45	66.11	70.88	7.22	73.11	75.43	3.17
FR	51.96	61.40	18.17	67.52	69.06	2.28	72.63	74.91	3.13
HI	46.57	54.93	17.96	57.25	60.58	5.82	60.26	62.02	2.91
RU	49.64	58.82	18.50	63.39	66.35	4.66	67.70	68.98	1.89
SW	37.82	46.51	22.96	45.91	49.20	7.17	50.68	49.42	-2.48
TH	36.61	53.31	45.64	48.20	56.21	16.60	53.85	57.03	5.89
TR	45.35	57.25	26.23	58.22	61.26	5.21	62.20	61.42	-1.25
UR	45.19	53.91	19.30	54.83	59.10	7.79	58.74	59.64	1.53
VI	49.20	59.98	21.91	63.45	67.25	5.98	69.46	71.44	2.84
ZH	48.74	60.26	23.63	63.97	66.63	4.15	68.64	71.60	4.30
Average	46.89	57.50	23.05	59.67	63.46	6.58	64.65	66.04	2.10

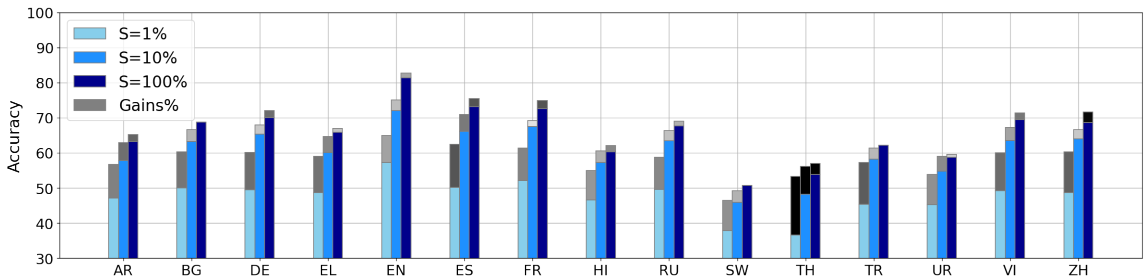
Table 11: We have reported the accuracy and relative gains using XLM-R and mBERT on XNLI dataset. The average relative gain is denotes the average gains across all the languages except the source EN.

Language	$S = 1\%$			$S = 10\%$			$S = 100\%$		
	Baseline	*DiTTO*	RG	Baseline	*DiTTO*	RG	Baseline	*DiTTO*	RG
XLM-R									
EN	56.40	61.28	8.66	64.17	64.37	0.31	66.81	66.91	0.15
DE	56.02	59.06	5.43	60.54	62.11	2.58	63.31	64.11	1.27
ES	53.29	54.87	2.97	56.34	56.96	1.10	57.68	58.80	1.95
FR	52.15	55.50	6.42	56.62	57.72	1.95	58.44	58.76	0.55
JA	49.77	53.29	7.08	52.93	55.46	4.77	53.77	55.98	4.10
ZH	48.05	51.01	6.15	50.34	52.74	4.78	51.50	53.66	4.20
Average	51.86	54.75	5.61	55.35	57.00	3.04	56.94	58.26	2.41
mBERT									
EN	54.32	55.82	2.76	60.80	62.77	3.22	65.53	65.71	0.27
DE	42.32	46.75	10.46	44.30	52.55	18.63	48.61	58.90	21.18
ES	42.30	45.81	8.30	45.77	51.18	11.83	49.56	54.75	10.48
FR	43.76	47.31	8.11	48.28	51.42	6.52	49.74	55.31	11.21
JA	36.82	38.66	5.00	39.00	43.78	12.27	39.32	48.77	24.03
ZH	32.31	41.85	29.55	36.32	46.40	27.74	38.67	49.60	28.27
Average	39.50	44.08	12.28	42.73	49.07	15.40	45.18	53.47	19.03

Table 12: We have reported the accuracy and relative gains using XLM-R and mBERT on MARC dataset. The average relative gain is denotes the average gains across all the languages except the source EN.

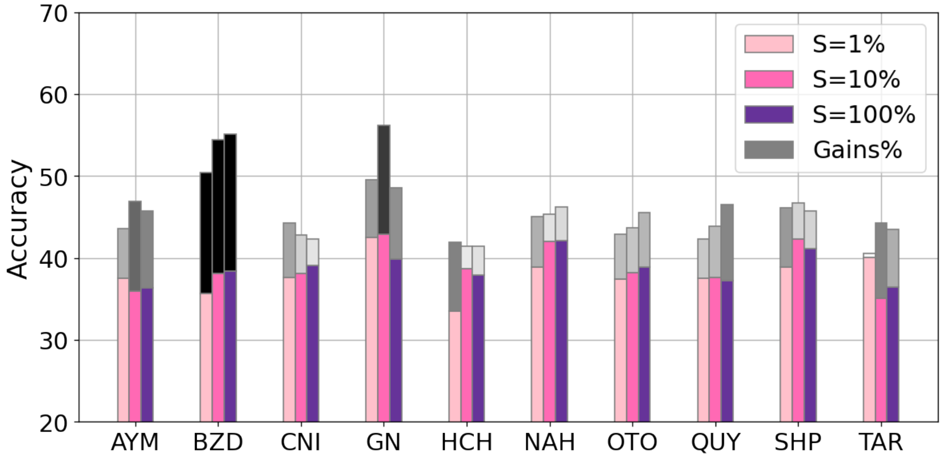


(a) XLM-R

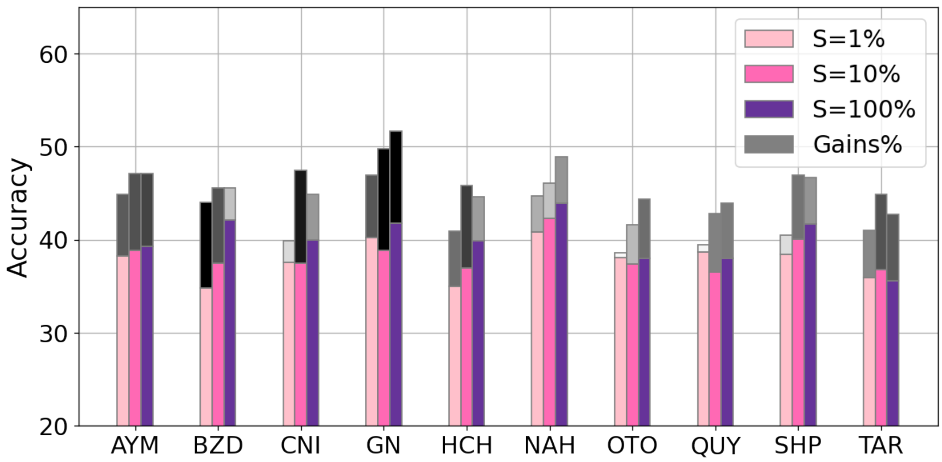


(b) mBERT

Figure 7: Absolute gains (darker shades of grey denotes higher gains) from *DiTTO* across all target languages on XNLI dataset.

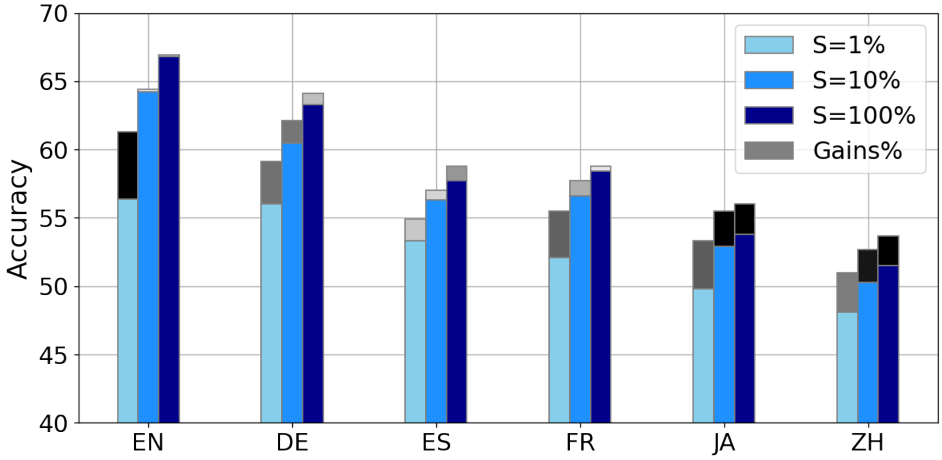


(a) XLM-R

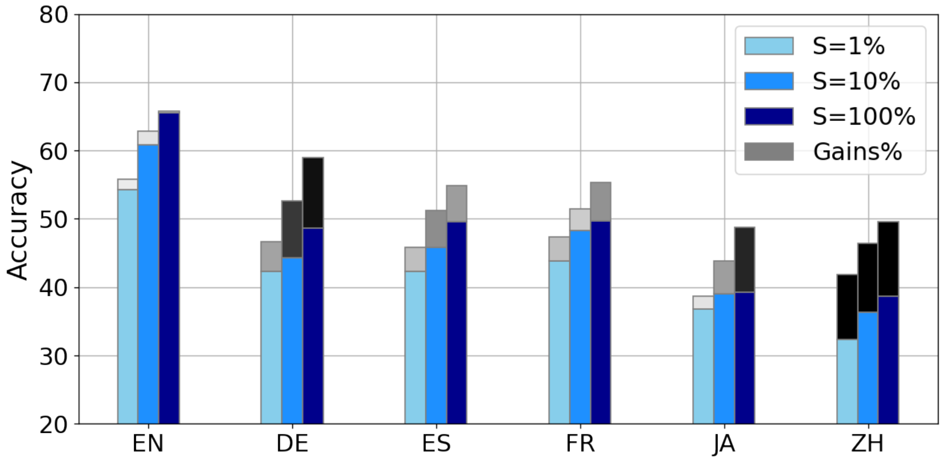


(b) mBERT

Figure 8: Absolute gains (darker shades of denotes higher gains) from *DiTTO* across all target languages on AmNLI dataset.



(a) XLM-R



(b) mBERT

Figure 9: Absolute gains (darker shades of denotes higher gains) from *DiTTO* across all target languages on MARC dataset.

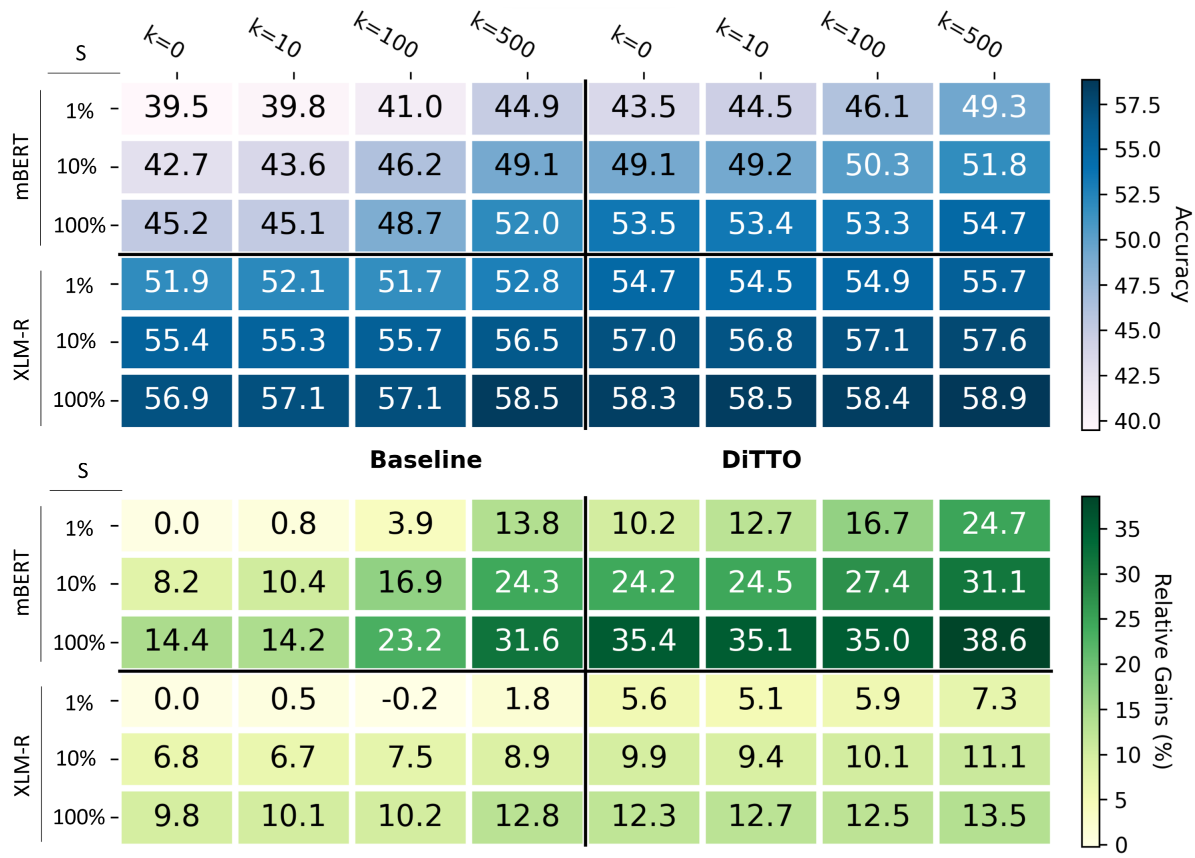


Figure 10: Accuracy/relative gains[2] on MARC dataset. Rows and columns denoting the amount of source and target language labeled instances, respectively.

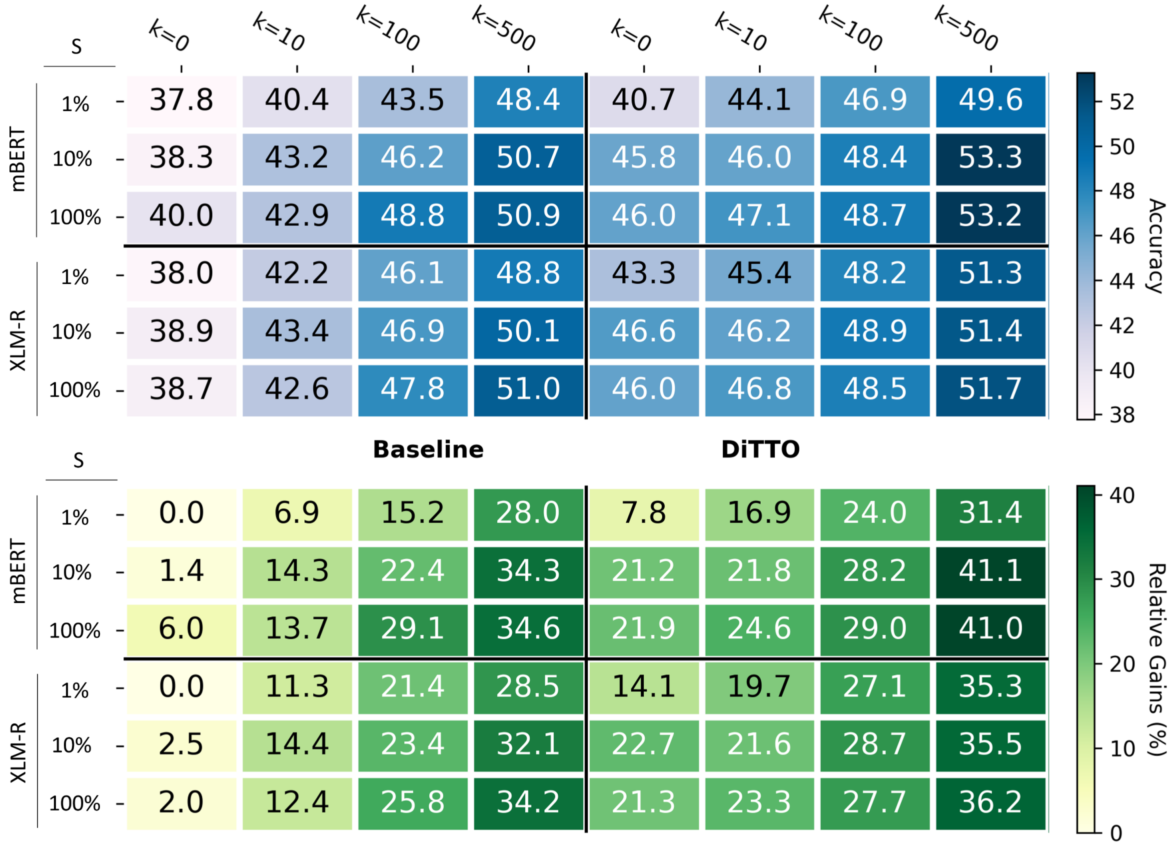


Figure 11: Accuracy/relative gains[3] on AmNLI dataset. Rows and columns denoting the amount of source and target language labeled instances, respectively.

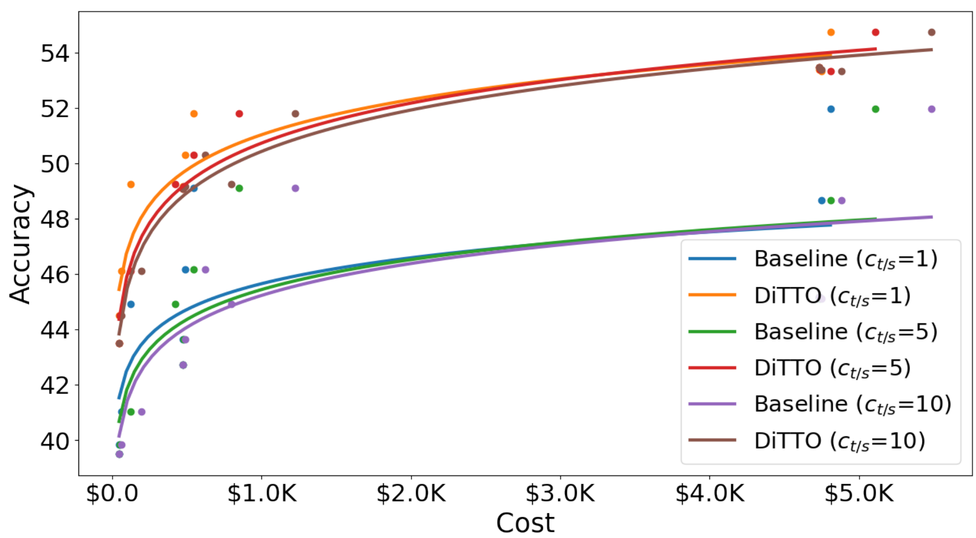


Figure 12: The plot shows Accuracy (vs) Cost graph with various values of $c_{t/s}$ for *DiTTO* and Baseline method trained using mBERT on XNLI (S=10%) dataset.

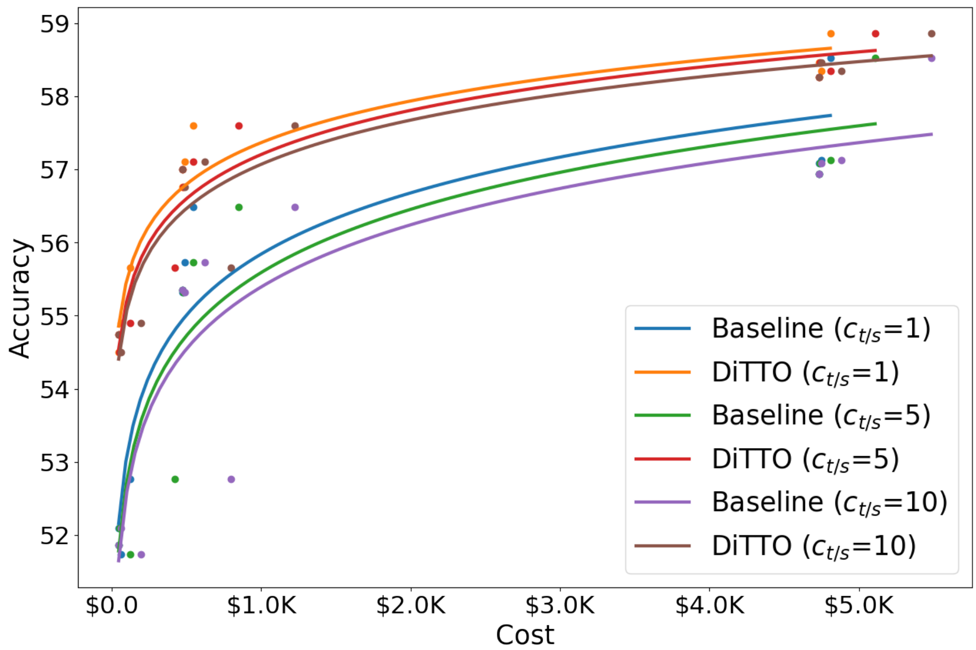


Figure 13: The plot shows Accuracy (vs) Cost graph with various values of $c_{t/s}$ for *DiTTO* and Baseline method trained using XLM-R on XNLI (S=10%) dataset.