BAD-X: Bilingual Adapters Improve Zero-Shot Cross-Lingual Transfer

Abstract

Adapter modules enable modular and efficient zero-shot cross-lingual transfer, where current state-of-the-art adapter-based approaches learn specialized language adapters (LAs) for individual languages. In this work, we show that it is more effective to learn bilingual language pair adapters (BAs) when the goal is to optimize performance for a particular source-target transfer direction. Our novel BAD-X adapter framework trades off some modularity of dedicated LAs for improved transfer performance: we demonstrate consistent gains in three standard downstream tasks, and for the majority of evaluated low-resource languages.

1 Introduction

Massively multilingual Transformers (MMTs) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021) have dominated research in multilingual NLP and cross-lingual transfer recently. Pretrained on large amounts of unlabelled data in 100+ languages, they have been shown to achieve impressive performance for a wide range of languages and tasks, and in zero-shot cross-lingual transfer in particular (Wu and Dredze, 2019; K et al., 2019). However, their representational capacity is known to be limited by the curse of multilinguality: a trade-off between the language coverage and model capacity (Conneau et al., 2020), which typically favors high-resource languages. Their limitations are thus especially pronounced in low-resource scenarios, in transfer between distant languages and towards resource-poor target languages (Hu et al., 2020; Lauscher et al., 2020; Ansell et al., 2021b, inter alia).

A standard approach to zero-shot cross-lingual transfer with MMTs (i) fine-tunes the full MMT on task-specific data in the source language and then (ii) applies it directly to make predictions in the target language (Hu et al., 2020). On top of the expensive fine-tuning of the entire large model, this standard procedure also does not ‘prepare’ the MMT to excel at a particular target language or for a particular source-target transfer direction.

This has been alleviated through modular parameter-efficient adaptations of the MMTs (Bapna and Firat, 2019; Philip et al., 2020; He et al., 2021) which bypass full fine-tuning, most prominently through lightweight adapters (Rebuffi et al., 2017; Houlsby et al., 2019): additional trainable parameters inserted into the MMT’s layers. They have recently been used for language and task specialization of the MMTs (Pfeiffer et al., 2020b), offering improved and more efficient zero-shot cross-lingual transfer.

Previous work (Pfeiffer et al., 2020b; Üstün et al., 2020, 2021; Vidoni et al., 2020; Ansell et al., 2021b, inter alia) focused on creating: 1) dedicated language adapters (LAs) for each individual language, and 2) individual task adapters (TAs). Creating single-language LAs enables a very modular approach to cross-lingual transfer, where a source language LA (used in training) can be directly swapped with any target language LA at inference. Yet, this procedure still does not prepare nor adapt the MMT for a particular source-target transfer direction. Put simply, if one’s incentive is to optimize the performance of a particular target language \(L_t\) given annotated data in a particular source language \(L_s\), especially under low-data regimes, one might try to capture the interplay between the two languages instead of learning separate LAs.

To address this gap, in this work we introduce the BAD-X framework: bilingual adapters (BAs) for zero-shot cross-lingual transfer (see Figure 1), designed towards improving transfer performance for a particular transfer direction, with a focus on low-resource target languages. The goal of BAD-X is to specialize the MMT for a particular language pair, while preserving all its existing knowledge encoded into the MMT’s parameters.

We experiment with three standard tasks in cross-
lingual transfer (Lauscher et al., 2020; Ansell et al., 2021b): part-of-speech tagging (POS), dependency parsing (DP) and natural language inference (NLI), and with a total of 20 low-resource target languages. Our results demonstrate that trading off modularity of single-language LAs for less modular BAs (tailored for language pairs) indeed yields improved transfer performance over the current state-of-the-art (SotA) adapter-based transfer framework MAD-X (Pfeiffer et al., 2020b), in all three tasks and for the large majority of target languages. Moreover, we show that, under the fixed fine-tuning budget and resources, further task performance gains can be achieved by varying the ratio of $L_s$-vs-$L_t$ unannotated data when learning BAs. We will share our code and pretrained BAs online at: [URL].

2 BAD-X: Methodology

Motivation and Overview. The main idea can be summarized into the following: instead of adapting the MMT to languages $L_s$ and $L_t$ separately as done in the SotA adapter-based MAD-X framework (Pfeiffer et al., 2020b), cross-lingual transfer might be more effective by adapting the MMT directly to the language pair $(L_s, L_t)$. This means that we learn a bilingual language-pair adapter instead of two separate monolingual LAs. We then learn a task adapter directly on top of the BA: since we focus on the zero-shot setting, this means using task-annotated examples only from $L_s$ to fine-tune the TA. This procedure is summarized in Figure 1.1

BAD-X Adapters. BAD-X adapts the MAD-X adapter framework, where BAs are learnt instead of single-language LAs. The architecture of the adapter in each layer $l$ consists of a down- and up-projection with a residual connection. More specifically, let the down-projection be a matrix $D_l \in \mathbb{R}^{h \times d}$ and the up-projection be a matrix $U_l \in \mathbb{R}^{d \times h}$, where $h$ is a hidden size of the MMT and $d$ is the hidden size of the adapter. Let us denote MMT’s hidden state and the residual at layer $l$ as $h_l$ and $r_l$, respectively. The adapter computation of layer $l$ is then given by:

$$A_l(h_l, r_l) = U_l(ReLU(D_l(h_l))) + r_l,$$  \hspace{1cm} (1)

1Inspiration for BAD-X originates from neural machine translation (NMT), where bilingual adapters have been trained on parallel corpora of two languages to recover performance of a massively multilingual NMT model for high-resource languages (Bapna and Firat, 2019). BAD-X, however, proposes bilingual adapters (i) without the use of any parallel data, (ii) with the goal to support downstream cross-lingual transfer, and (iii) targets low-resource target languages.

with ReLU as the activation. This formulation subsumes LAs and TAs in MAD-X, as well as BAs and TAs in BAD-X, where LAs/BAs receive the input from the (frozen) Transformer layer, while TAs receive the input from the (frozen) LA/BA put on top of the frozen Transformer layer (Figure 1).2

MAD-X LAs are trained via masked-language modeling (MLM) objective on the Wikipedia of the corresponding language, while TAs are trained on annotated task data. Once LA for $L_s$ is available, TA is trained by stacking it on top of the fixed source LA. Transfer is done by replacing the $L_s$ LA with the $L_t$ LA. Unlike MAD-X, BAD-X trains a single bilingual adapter via MLM, alternating between the unlabelled (Wikipedia) data from both $L_s$ and $L_t$. The ‘data alternations’ are done according to a predefined ratio: e.g., the ratio of $N:1$ denotes that the model would see $N$ $L_s$ sentences followed by 1 $L_t$ sentence. The motivation for this is twofold: 1) seeing a data mixture from the two languages could produce a BA that is better for transfer than having two independent LAs; 2) LAs for low-resource $L_t$-s might otherwise overfit due to unlabelled data scarcity in $L_t$, and thus could benefit from additional $L_s$ data.

In BAD-X, TA is then again trained on top of the fixed BA, and the same BA-TA configuration is retained at inference, see Figure 1 again.

Advantages and Limitations. BAD-X allows parameter-efficient transfer to arbitrary tasks and languages by learning modular bilingual and task representations. It trades-off some modularity of MAD-X for increased performance and expressiveness when the goal is to perform a transfer for a fixed pair of languages. A disadvantage of BAD-X with respect to modularity is that it no longer offers a zero-cost transfer (once all LAs are learnt) between all language pairs under consideration: it

Figure 1: BAD-X adapter module at one MMT layer, showing the BAD-X BA for one language pair (English-Wolof: En-Wo) and the POS TA. The same module (but different parameters) is added at each MMT layer.
requires training of separate BAs for all pairs of interest. However, as we show further in §3, BAD-X might be preferable over MAD-X in the cases when the goal is to improve a particular source-target direction, which is our targeted use-case.

3 Experiments and Results

Tasks and Languages. We treat MAD-X as our principal baseline, and conduct all evaluations and analyses on three standard cross-lingual tasks which allow for experimentation with low-resource target languages: POS, DP, and NLI.

For POS and DP, we sample ten low-resource languages from the Universal Dependencies (UD) 2.7 dataset (Zeman et al., 2020), taking into account: 1) the availability and the size of the corresponding Wikipedia; and 2) typological diversity to ensure that different language families are covered.\(^3\) For NLI, we rely on the recent AmericasNLI dataset (Ebrahimi et al., 2021), spanning ten low-resource languages from the Americas. For AmericasNLI languages, we use Wikipedia if available; otherwise we use the unlabelled data previously used by Ansell et al. (2021a). English is the source language in all experiments for all tasks.\(^4\)

All languages along with their language codes are listed in Table 2 in the Appendix.

3.1 Experimental Setup

MMT. In all our experiments, we use mBERT, an MMT model pretrained on the Wikipedias of 104 languages (Devlin et al., 2019).\(^5\)

Training Setup: LAs and BAs. To enable a fair comparison between MAD-X and BAD-X under the same training and inference conditions, we train our own MAD-X LAs from scratch with the MLM objective on monolingual Wikipedias: training is run for 25,000 steps, with a batch size of 64 and a learning rate of \(1\times10^{-4}\). We evaluate the LAs every 500 training steps and finally choose the LA that yields the lowest perplexity, as evaluated on the 5% of the Wikipedia data that acts as a validation set.

Pfeiffer et al. (2020b) empirically established that strong task performance of MAD-X on low-resource languages can be achieved already after 20,000 LA training steps, and that longer training offers only modest to negligible performance gains. Driven by their findings, we train MAD-X LAs for 25,000 iterations due to computational constraints, a large number of experiments, and the low-resource nature of our target languages.

BAD-X BAs are trained on the Wikipedia data of both \(L_s\) and \(L_t\). The standard BAD-X variant termed Balanced BAD-X (also BAD-X 1:1) is trained by alternating one batch of the \(L_s\) data (i.e., English) followed by one batch of the \(L_t\) data, for 50,000 iterations (i.e., this way we match the total number of iterations performed by training MAD-X \(L_s\) and \(L_t\) LAs for 25,000 iterations each), and we adopt all the hyperparameters from MAD-X LA training. We select as the final BA the one with the lowest \(L_t\) perplexity. Bilinguality of the BAD-X BAs allows us to directly train TA on top of it and perform the inference with the same setup.

Training Setup: TAs. For POS and DP, TA is trained by stacking it on top of the source (i.e., English) LA (with MAD-X) or the English-\(L_t\) BA (with BAD-X) and performing 15,000 steps with a batch size of 8 and a learning rate of \(5\times10^{-5}\). We evaluate the TAs every 250 steps on English validation sets, and select as the final TAs the ones with the best accuracy (POS) and LAS scores (DP). The adapter reduction factor (Pfeiffer et al., 2020a) is 2 for LAs and 16 for TAs. For AmericasNLI, we train its TA on the English MultiNLI dataset (Williams et al., 2018) following the setup of Ebrahimi et al. (2021): 5 epochs with a batch size of 32, and a learning rate of \(2\times10^{-5}\). We evaluate the TA every 625 steps and choose the one with the best accuracy on the English validation set.

BAD-X: BA Variants. Besides Balanced BAD-X, we consider other variants of BAD-X BAs that differ in the data ratios between \(L_s\) and \(L_t\); we denote these variants as BAD-X 1:N, where 1 batch of \(L_s\) data is followed by \(N\) batches of \(L_t\) data, and vice versa: BAD-X N:1. With these variants, we aim to answer the following question: given a fixed number of MLM training steps (i.e., a fixed computational budget) for BAs, is it possible to further impact/improve transfer performance? Is the optimal data sampling ratio task-dependent?

3.2 Results and Discussion

The results for all languages and tasks with MAD-X and Balanced BAD-X are summarized in Table 1. As a general trend, we observe that the proposed
Balanced BAD-X variant outperforms MAD-X over a majority of languages and across all three tasks: besides offering higher average results, we also report gains on 8/10 (POS; accuracy), 10/10 (DP; UAS), and 9/10 (NLI; accuracy) target languages. This confirms the positive impact of BA training, which is able to capture additional interactions of each language pair, in lieu of LA training.

**Performance across Tasks.** In particular, BAD-X gains on average 1.06% in accuracy and 0.66% in $F_1$ compared to MAD-X on POS task. The gains are even more pronounced on the more complex DP task, which shares the target language set with POS: BAD-X outperforms MAD-X on average with an even larger gap of 2.62% in UAS and 2.38% in LAS scores, on average. The gain is particularly high for Wolof, a West-African language spoken by more than five million people, with ~9% improvement over MAD-X in both UAS and LAS scores. Wolof is also a language with one of the highest gains in POS. We also observe the superior performance of BAD-X over MAD-X on all tasks and languages. BAD-X targets improving transfer performance for all tasks, that is, the ‘winning’ variant seems to be task-dependent. In particular, DP benefits the most from 5:1 sampling, while for POS and NLI the 1:2 variant outscores the others although DP and POS share exactly the same BA training data.

Note that, due to computational constraints, we did not extensively search for the best sampling ratios of the source and target language during BA training, thus the optimal strategy might not be covered by our experiments. However, these findings warrant further investigation in future work.

**4 Conclusion**

We have presented BAD-X, a novel adapter-based framework for zero-shot cross-lingual transfer. BAD-X targets improving transfer performance for particular fixed source-target transfer directions through the introduction and use of dedicated bilingual language-pair adapters (BAs). The effectiveness of the BAs and the BAD-X framework has been demonstrated on three standard transfer tasks, across a plethora of low-resource languages.
References


Ahmet Üstün, Arianna Bisazza, Gossé Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation


A Details of the Experimental Setup

Computing Infrastructure. All experiments were run on a single NVIDIA GeForce RTX 3090 GPU; training one BAD-X BA for 50,000 iterations took around 24 hours (MAD-X LA for 25,000 steps took around 12 hours). Training of any TA took less than two hours. Evaluation is performed within the AdapterHub framework (Pfeiffer et al., 2020a).

Hyperparameters. All hyperparameters were taken from (Pfeiffer et al., 2020b), as discussed in the main paper, and no hyperparameter search was done. All reported results are from a single run.

B Languages

The list of languages in each task along with their language codes is provided in Table 2.

C BAD-X: Full results

Full results on all languages for MAD-X and all BAD-X variants are given in Tables 3, 4 and 5 for POS, DP and NLI, respectively.
Table 2: Lists of tasks with all the languages.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS, DP</td>
<td>Afrikaans, Bambara, Basque, Erzya, Komi-Zyryan, Maltese, Marathi, Telugu, Uyghur, Wolof</td>
</tr>
<tr>
<td>NLI</td>
<td>Asháninka, Aymara, Bribri, Guaraní, Náhuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, Wixarika</td>
</tr>
</tbody>
</table>

Table 3: Results of MAD-X and all BAD-X variants on POS. Scores are accuracy/F1. The last column is the average score over all languages.

<table>
<thead>
<tr>
<th>Method</th>
<th>AF</th>
<th>BM</th>
<th>EU</th>
<th>MYV</th>
<th>KPV</th>
<th>MT</th>
<th>MR</th>
<th>TE</th>
<th>UG</th>
<th>WO</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAD-X</td>
<td>88.97/85.43</td>
<td>45.92/41.61</td>
<td>70.64/58.90</td>
<td>72.92/66.84</td>
<td>57.18/52.63</td>
<td>74.12/69.94</td>
<td>57.58/52.65</td>
<td>79.81/75.27</td>
<td>60.26/47.07</td>
<td>68.00/51.78</td>
<td>67.34/50.71</td>
</tr>
<tr>
<td>BAD-X 1-2</td>
<td>87.09/85.53</td>
<td>48.40/43.91</td>
<td>72.03/60.88</td>
<td>75.55/69.49</td>
<td>57.88/48.43</td>
<td>72.70/68.40</td>
<td>59.45/54.31</td>
<td>81.33/76.63</td>
<td>63.86/56.53</td>
<td>71.79/65.74</td>
<td>69.02/61.98</td>
</tr>
<tr>
<td>BAD-X 1-1</td>
<td>86.68/85.94</td>
<td>47.05/42.40</td>
<td>71.16/59.48</td>
<td>75.55/69.49</td>
<td>57.88/48.43</td>
<td>72.70/68.40</td>
<td>59.45/54.31</td>
<td>81.33/76.63</td>
<td>63.86/56.53</td>
<td>71.79/65.74</td>
<td>69.02/61.98</td>
</tr>
<tr>
<td>BAD-X 2-1</td>
<td>87.01/85.26</td>
<td>45.59/40.96</td>
<td>73.17/67.28</td>
<td>75.55/69.49</td>
<td>57.88/48.43</td>
<td>72.70/68.40</td>
<td>59.45/54.31</td>
<td>81.33/76.63</td>
<td>63.86/56.53</td>
<td>71.79/65.74</td>
<td>69.02/61.98</td>
</tr>
<tr>
<td>BAD-X 5-1</td>
<td>86.98/85.44</td>
<td>48.67/44.35</td>
<td>70.75/59.76</td>
<td>75.90/69.59</td>
<td>57.68/49.52</td>
<td>71.62/67.66</td>
<td>58.81/54.21</td>
<td>80.40/75.63</td>
<td>58.39/43.45</td>
<td>70.30/64.55</td>
<td>67.85/61.21</td>
</tr>
</tbody>
</table>

Table 4: Results of MAD-X and all BAD-X variants on DP. Scores are UAS/LAS. The last column is the average score over all languages.

<table>
<thead>
<tr>
<th>Method</th>
<th>CNI</th>
<th>AYM</th>
<th>BZD</th>
<th>GN</th>
<th>NAH</th>
<th>OTO</th>
<th>QUY</th>
<th>TAR</th>
<th>SHP</th>
<th>HCH</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAD-X</td>
<td>66.64/54.50</td>
<td>35.19/12.17</td>
<td>54.71/32.06</td>
<td>55.18/33.64</td>
<td>43.74/22.03</td>
<td>60.42/44.16</td>
<td>46.09/27.49</td>
<td>64.77/48.54</td>
<td>33.74/15.13</td>
<td>46.04/24.84</td>
<td>50.58/21.55</td>
</tr>
<tr>
<td>BAD-X 1-2</td>
<td>67.83/55.42</td>
<td>37.70/15.10</td>
<td>53.88/31.84</td>
<td>54.26/38.07</td>
<td>44.20/22.95</td>
<td>61.79/43.29</td>
<td>48.71/30.51</td>
<td>68.93/52.58</td>
<td>33.03/15.94</td>
<td>51.72/30.77</td>
<td>52.62/33.55</td>
</tr>
<tr>
<td>BAD-X 1-1</td>
<td>68.02/55.75</td>
<td>37.20/14.47</td>
<td>55.42/33.30</td>
<td>58.61/37.74</td>
<td>44.34/25.81</td>
<td>61.37/42.45</td>
<td>48.01/29.19</td>
<td>68.69/51.91</td>
<td>35.07/15.11</td>
<td>54.82/33.93</td>
<td>53.20/33.93</td>
</tr>
<tr>
<td>BAD-X 2-1</td>
<td>67.81/55.70</td>
<td>36.35/14.11</td>
<td>54.70/33.40</td>
<td>58.78/37.58</td>
<td>43.04/22.81</td>
<td>63.18/44.68</td>
<td>49.88/30.40</td>
<td>66.90/49.98</td>
<td>34.31/14.40</td>
<td>55.66/33.69</td>
<td>53.07/33.58</td>
</tr>
<tr>
<td>BAD-X 5-1</td>
<td>68.03/56.03</td>
<td>36.50/14.40</td>
<td>53.65/31.84</td>
<td>62.03/42.22</td>
<td>45.26/24.67</td>
<td>62.68/42.28</td>
<td>49.52/30.40</td>
<td>66.65/48.54</td>
<td>35.47/14.31</td>
<td>57.08/36.78</td>
<td>53.78/34.15</td>
</tr>
</tbody>
</table>

Table 5: Results of MAD-X and all BAD-X variants on NLI. Scores are accuracy. The last column is the average score over all languages.

<table>
<thead>
<tr>
<th>Method</th>
<th>CNI</th>
<th>AYM</th>
<th>BZD</th>
<th>GN</th>
<th>NAH</th>
<th>OTO</th>
<th>QUY</th>
<th>TAR</th>
<th>SHP</th>
<th>HCH</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAD-X</td>
<td>42.53</td>
<td>46.67</td>
<td>44.53</td>
<td>54.53</td>
<td>47.56</td>
<td>41.18</td>
<td>49.47</td>
<td>37.87</td>
<td>41.73</td>
<td>38.40</td>
<td>44.45</td>
</tr>
<tr>
<td>BAD-X 1-2</td>
<td>45.60</td>
<td>52.13</td>
<td>45.47</td>
<td>56.93</td>
<td>45.53</td>
<td>45.05</td>
<td>54.13</td>
<td>39.07</td>
<td>47.20</td>
<td>45.47</td>
<td>47.66</td>
</tr>
<tr>
<td>BAD-X 1-1</td>
<td>48.13</td>
<td>47.33</td>
<td>44.93</td>
<td>58.00</td>
<td>48.24</td>
<td>41.44</td>
<td>49.33</td>
<td>38.93</td>
<td>47.07</td>
<td>45.07</td>
<td>46.85</td>
</tr>
<tr>
<td>BAD-X 2-1</td>
<td>46.27</td>
<td>50.27</td>
<td>46.13</td>
<td>51.47</td>
<td>48.10</td>
<td>40.51</td>
<td>53.20</td>
<td>37.60</td>
<td>48.18</td>
<td>43.60</td>
<td>46.53</td>
</tr>
<tr>
<td>BAD-X 5-1</td>
<td>43.20</td>
<td>52.13</td>
<td>45.73</td>
<td>56.27</td>
<td>46.75</td>
<td>43.18</td>
<td>55.73</td>
<td>37.47</td>
<td>50.40</td>
<td>42.53</td>
<td>47.34</td>
</tr>
</tbody>
</table>