Stop Pre-Training: Adapt Vision-Language Models to Unseen Languages

Jeremy Di Dio
Distributed Information Systems Laboratory
jeremy.didio@epfl.ch

Abstract

Vision-Language Pretraining has significantly improved the performance of various vision-language tasks, including image-text retrieval, Visual Entailment, and Visual Question Answering. However, existing pretraining methods primarily rely on English-based datasets, resulting in a bias towards North American or Western European languages. This bias poses challenges for zero-shot or few-shot cross-lingual transfer learning, where the model's performance suffers.

In this project, we propose a novel approach to enhance a predefined vision-language model by directly incorporating adapter modules into the text encoder. This integration of adapter layers enables efficient parameter transfer to diverse tasks and languages, addressing the limitations of existing models. The key focus of our work is to leverage these adapter layers to facilitate transfer learning for low-resource languages, where data availability is limited. Through comprehensive experiments and evaluations, we demonstrate the effectiveness of our approach in improving cross-lingual transfer learning performance.

1 Introduction

Recently, the field of multimodal learning has been revolutionized by Vision-Language (VL) models, which effectively capture the intricate relationship between textual and visual informations. VL models, such as ALBEF (Li et al., 2021) or BLIP (Li et al., 2022), have demonstrated remarkable performance across various tasks by leveraging large-scale dataset during pretraining. However, the reliance on extensive pretraining poses challenges when applying these models to languages with limited pretraining data. Furthermore, the majority of VL datasets predominantly consist of English samples, thereby introducing an inherent bias towards the Anglo-Saxon world when employing these pretrained models on different languages. This issue has created significant attention within the research community, leading to an emergence of dedicated multilingual VL datasets and tasks (Liu et al., 2021).

This project arises from this need to extend the capabilities of VL models to low resources languages (Magueresse et al., 2020). Existing approaches heavily rely on extensive pretraining using massive multilingual corpora (Raffel et al., 2019), making it challenging to adapt these models to languages with scarce data. To overcome this issue, we propose an approach that enables the adaptation of such models to unseen languages without the need for a large dataset. By leveraging the capabilities of adapter layers in the transformer-based architecture of these models, our objective is to fine-tune a pretrained model for a low resource target language. This approach not only enables us to achieve desired results but also reduces computational complexity. Previous research in transfer learning (Jafari et al., 2021) and adapter-based architectures (Pfeiffer et al., 2020a) has laid the groundwork for our study. Researchers have successfully employed adapter layers in various NLP tasks, showcasing their ability to improve performance with reduced computational requirements. Our work extends this research to the domain of VL models and addresses the specific challenge of adapting pretrained VL models to unseen languages.

In this article, we present our methodology for adapting a particular VL models to new languages using adapter layers. We outline the experimental setup employed to assess the efficacy of our approach in multimodal cross-lingual learning. We provide a comprehensive overview of the chosen VL model, the available training data used and the results obtained in our study.

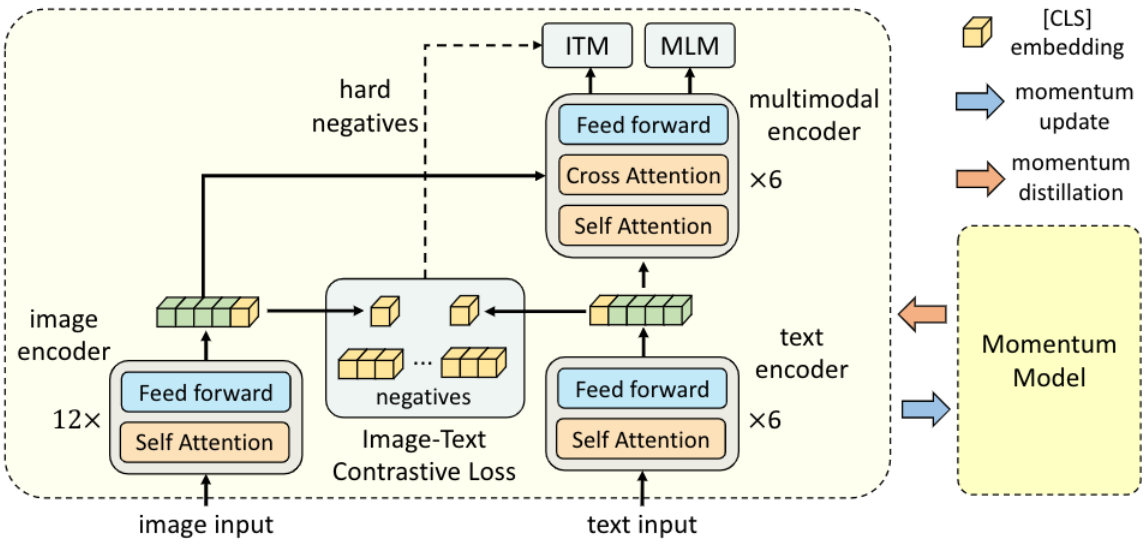


Figure 1: Illustration of ALBEF architecture from (Li et al., 2021)

2 Model and Adapters

In order to carry out this project, we made use of a novel framework for vision-language representation learning known as ALBEF (Li et al., 2021). Our primary objective was to improve this existing model by integrating adapter modules, allowing the model to acquire task or language-specific parameters. In pursuit of this goal, we adopted different adapter architecture, which are elaborated upon in Section 2.2.

2.1 ALBEF

ALBEF (ALign BEfore Fuse) is a VL model that effectively combines image and text information. The model independently encodes image and text informations, and subsequently fuses them through cross-modal attention. Prior to the fusion process, an intermediate contrastive loss is employed to align both modalities, ensuring optimal integration. To address the challenge of training with noisy supervision, the model incorporates Momentum Distillation (MoD) methodology.

This particular model consists of three distinct components, namely an image encoder, a text encoder, and a multimodal encoder, as illustrated in Figure 1. The text encoder, known as XBert in the following results, is a modified version of BERT^{base} (Devlin et al., 2019) with 123.7M parameters. Similarly, the image encoder is a modified variation of ViT-B/16 (Dosovitskiy et al., 2021) with 85.8M parameters.

2.2 Adapters

In the context of Large Language Model (LLM), adapters (Houlsby et al., 2019) are additional bottleneck modules inserted in each layers of the model without modifying the original model architecture. Adapters are designed to capture task or language-specific informations and adapt pre-trained models to new tasks or languages. Adapters are typically composed of a down-projection, followed by a non-linear function and then an up-projection. By adding adapters, a pre-trained model can learn task-specific parameters while preserving the knowledge gained from the previous pre-training. Resulting from the fact that during the fine-tuning process, the adapter parameters are optimized for the specific downstream task while keeping the original model parameters frozen.

The use of adapters in LLMs, allows for efficient transfer learning and customization as it reduces the computational and memory requirements associated with fine-tuning the entire model.

In the article (Pfeiffer et al., 2020b), various types of adapters are outlined, including task adapters, language adapters, and invertible adapters. For our study, we employed these adapter architectures, each possessing distinct characteristics as described in the following sections. Figure 2 visually depicts the integration of these specific adapters into a transformer model's layer.

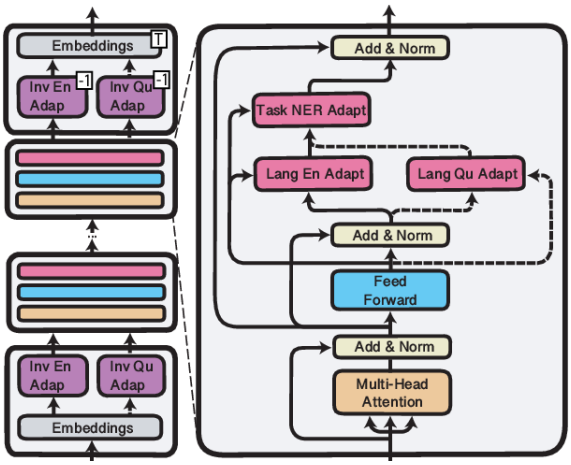


Figure 2: Illustration of task, language and invertible adapters implementation in a transformer model from (Pfeiffer et al., 2020b)

2.2.1 Language adapters

Language adapters are used to learn the language-specific transformation. The language adapter LA_l at layer l consists of a down projection $\mathbf{D}_l : D \in \mathbb{R}^{h \times d}$ where h is the hidden size of the transformer and d is the latent size of the adapter. This down projection is followed by a GeLU activation function and an up projection $\mathbf{U}_l : U \in \mathbb{R}^{d \times h}$.

Equation 1 showcases the output of such language adapter:

$$\mathsf{LA}_l(\mathbf{h}_l, \mathbf{r}_l) = \mathbf{U}_l(\mathsf{GeLU}(\mathbf{D}_l(\mathbf{h}_l))) + \mathbf{r}_l \quad (1)$$

where $\mathbf{h}_l$ and $\mathbf{r}_l$ denotes the transformer hidden state and residual at layer l, respectively.

2.2.2 Task adapters

Task adapters share the identical architecture as language adapters and are stacked on top of the language adapter as depicted in Figure 2.

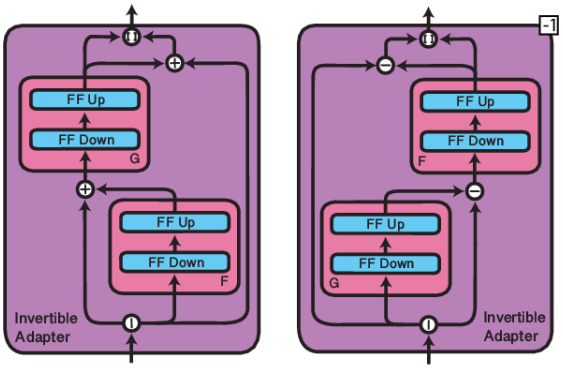
Consequently, task adapters receive two inputs: the output LA_l of the language adapter at layer l, and the residual $\mathbf{r}_l$ from the transformer's feed-forward layer:

$$\mathsf{TA}_l(\mathsf{LA}_l, \mathbf{r}_l) = \mathbf{U}_l(\mathsf{GeLU}(\mathbf{D}_l(\mathsf{LA}_l))) + \mathbf{r}_l \quad (2)$$

2.2.3 Invertible adapters

Invertible adapters are used in combination with language adapters and are designed to address vocabulary mismatch between source and target languages. They aim to facilitate effective token-level language-specific transformations. Positioned above the embedding layer, they are paired with their respective inverses, which are located before the output embedding layer (see Figure 2). This configuration ensures that the invertible adapters play a crucial role in aligning vocabularies and enabling accurate language-specific transformations.

The architecture of these adapters is illustrated in Figure 3, with F and G representing arbitrary non-linear functions that are applied within the adapter.



(a) The invertible adapter (b) The inversed adapter

Figure 3: Illustration of an invertible adapter (a) and its inverse (b) from (Pfeiffer et al., 2020b). The input is initially split in half and then subjected to a transformation using alternating projections F and G. The symbols +/− denote element-wise addition/subtraction, | denote the splitting of the input vector and [ ] represent the concatenation of two vectors.

3 Methods

3.1 Implementation details

In order to allow the ALBEF model to work on unseen languages without requiring a full pretraining on the target language, we took advantage of the architecture of the text encoder to implement adapter modules between each layer of it. Specifically, we employed task adapters, language adapters, and invertible adapters at each layer of the text encoder, thereby empowering the model with language-specific capabilities while preserving its core architecture and reducing the computing time.

The training process first consisted of the training of language-specific adapter modules via Masked Language Modeling (MLM) on unlabeled target language data, resulting in the acquisition of language-specific parameters. Following the language adapter training phase, we further fine-tuned the model by stacking the task adapter on top of the source language adapter and training it on the downstream task in the source language.

During the training process, it is important to highlight that all parameters of the model were frozen, except for the adapter parameters that were specifically trained. This approach allowed us to concentrate uniquely on fine-tuning the adapter parameters, resulting in reduced computational time while preserving theexisting knowledge of the model.

To summarize, the training process for both language and task adapters consists of the following six steps:

1. Addition of language and invertible adapter in the text encoder layers
2. Freezing of all the model's parameters except for the language adapter module parameters
3. Training of the language adapter via MLM on each language separately
4. Stacking of a task adapter on top of the source language adapter
5. Freezing of all the model's parameters except for the task adapter module parameters
6. Training of the task adapter on the downstream task

Upon completion of the training phase, we obtained a task adapter module and multiple language adapter modules, with each language having its own dedicated language adapter. To evaluate our models, we adopted an approach where we alternatively combined the task adapter with each language adapter. Consequently, we tested multiple models, each tailored to a specific language, by selecting the corresponding language adapter.

3.2 Datasets

During the training and evaluation of our models, we leveraged multiple datasets that are specifically relevant to the given task. For clarity, we provide the details of each dataset, including its modality and the corresponding downstream task.

3.2.1 XNLI

The Cross-lingual Natural Language Inference (XNLI) (Conneau et al., 2018) dataset is an extension of the Multi-Genre NLI (MNLI) (Williams et al., 2018) dataset to 15 languages. The corpus was created by manually translating the validation and test sets of MNLI into each of those 15 languages. The English training set was machine translated for all languages. This dataset is used for text classification, each sample is composed of two different sentences, a hypothesis and a premise, in a particular language. The goal is therefore to classify these two sentences as either entailment, contradiction, or neutral.

For this project, we selected a subset of the 15 available languages to keep only the one using the Latin alphabet. Therefore, we only used data in the following languages: English, French, Spanish, Turkish, and Swahili.

3.2.2 WIT

The Wikipedia-based Image Text Dataset (WIT) (Srinivasan et al., 2021) dataset is a large multimodal, multilingual dataset composed of a set of 37.6 million entity rich image-text examples with 11.5 million unique images across 108 Wikipedia languages. Each sample consists of an image and its caption.

3.2.3 Flickr30k

The Flickr30k dataset (Young et al., 2014) consists of 31,000 images collected from the photo-sharing website Flickr. Each image in the dataset is paired with five English sentences that describe the content of the image. These sentences were created by human annotators and provide a rich source of textual descriptions for training and evaluating image captioning models.

3.2.4 XVNLI

The Cross-lingual Visual Natural Language Inference (XVNLI) dataset (Bugliarello et al., 2022) is the combination of the text-only dataset SNLI (Bowman et al., 2015), with its multimodal (Xie et al., 2019) and cross-lingual (Agić and Schluter, 2018) counterparts. The main task of this dataset is cross-lingual visual entailment. In other words, the goal is to predict if a text-hypothesis 'entails', 'contradicts', or is 'neutral' to an image-premise.

4 Experiments

Given that the text encoder was the only modified part in ALBEF, we opted to initially assess the performance of our implementation uniquely on cross-lingual textual tasks. This approach allowed us to verify the correctness of our implementation before evaluating our modified VL model on multimodal cross-lingual tasks.

4.1 XBert experiments

To evaluate our modified text encoder, we focused exclusively on the XNLI dataset. As outlined in Section 3.1, our initial approach involved training language adapters for multiple languages: English (as the source language), along with French, Spanish, Turkish, and Swahili (as target languages). Once the language adapters were trained, we proceeded to stack the task adapter on top of the source language adapter and train it on the downstream task, here text classification. To assess the performance of our models, we conducted evaluations in each language using the XNLI dataset's testing set.

Additionally, we investigated the influence of the loaded pretrained model by repeating the experiment using mBERT, a multilingual version of BERT (Devlin et al., 2019). Moreover, we explored the impact of the latent size of the hidden adapter, considering its potential implications on overall performance. Section 5.1 exposes the results obtained from these different experiments.

4.2 ALBEF experiments

In order to evaluate the performance of our method on Vision Language tasks, we conducted experiments using the complete ALBEF model with our modified text encoder. The process of adapter training followed the procedure outlined in Section 3.1. The difference lies in the choice of datasets and tasks employed in these experiments.

For training our language adapters via MLM, we used the WIT dataset. Our task adapters, combined with the trained source language adapter, were trained on the Flickr30k dataset using visual entailment as the downstream task. As the Flickr30k dataset is not multilingual, we employed the XVNLI dataset for evaluating our models in each target language. In consideration of the number of different used datasets, we selected French and Spanish as the target languages for evaluation purposes. Detailed results from these diverse experiments are presented in Section 5.3.

Furthermore, we note that during the training of our language adapters using the WIT dataset, we processed the dataset to retain only 100,000 samples at random. This was done to simulate low-resource languages and investigate the performance of our method under such conditions.

5 Main results and discussion

This section presents a comprehensive analysis of the results obtained from the experiments discussed in Section 4.

5.1 XBert results

The provided results in Table 1 showcase the performance of different models on the XNLI text classification task across multiple languages.

XBert^{base}: This model without any additional modifications or adapters serves as a reference point. It achieves the highest accuracy of 82.2% for the English language. The accuracies for the other target languages (fr, es, tr, sw) are comparatively lower, which is expected.

XBert^{base} + L/T$_{64}$: The inclusion of language adapters and task adapters on top of the baseline model shows slight improvements in some languages. Notably, it enhanced the performance for French and Spanish, as their accuracies increased by 4.4% and 1.2%, respectively. However, it does not provide any substantial improvements for other languages.

XBert^{mBert}: Using the multilingual BERT (mBERT) as the pretrained model yields lower accuracy for English compared to the baseline model. However, it shows significant improvements for French, Spanish, and Swahili, surpassing the baseline accuracy in those languages. The accuracy for Turkish is the lowest among all languages.

XBert^{mBert} + L/T$_{64}$: As expected, the addition of language adapters and task adapters on top of the XBert^{mBert} model shows slight improvements in some languages. It notably improves accuracy for French, Spanish, and Swahili, with accuracies increasing by 0.6%, 0.7% and 4.4%, respectively.

The overall results demonstrate the significant performance advantage of multilingual models in cross-lingual transfer learning compared to our adapter configurations using the baseline model. The multilingual models consistently outperform our adapter-based approaches.

However, we still observe that the inclusion of language adapters and task adapters still yields improvements for specific languages, it is important to

Model	en	fr	es	tr	sw
XBert^{base}	**82.2**	39.1	44.0	38.3	37.3
XBert^{base} + L/T$_{64}$	82.2	43.5	45.2	35.9	37.3
XBert^{mBert}	78.9	71.9	72.8	**61.0**	51.3
XBert^{mBert} + L/T$_{64}$	78.9	**72.5**	**73.5**	59.4	**55.7**

Table 1: Reported accuracy on XNLI text classification task using both bert-based-uncased and bert-based-multilingual-uncased as pretrained models. L/T indicates the addition of language and task adapters, if not present, we assume zero-shot cross-lingual setting.

note that the impact of these adapters is not consistent across all languages. The effectiveness of the adapters varies depending on the language, highlighting the complexities and nuances associated with language-specific adaptations.

5.2 Adapters latent size impact

The provided results in Table 2 showcase the accuracy of XBert^{base} + L/T with varying latent sizes of language and task adapters on the XNLI text classification task.

These findings clearly indicate that the latent size of the adapters plays a crucial role in the model's performance. A larger latent size can significantly help capture more complex patterns and improve overall accuracy.

While the results demonstrate the potential benefits of increasing the latent size of the adapters, it is crucial to acknowledge that this approach contradicts with the project's primary objective, which aims to find a balance between performance improvements and computational complexity.

Therefore, finding the optimal balance becomes essential, where the performance improvements achieved by increasing the latent size are weighed against the computational costs incurred.

Model	en	fr	es
XBert^{base} + L/T$_{64}$	82.2	43.5	45.2
XBert^{base} + L/T$_{256}$	82.1	51.0	49.2
XBert^{base} + L/T$_{512}$	**84.8**	55.7	52.8
XBert^{base} + L/T$_{1024}$	84.1	**60.6**	**59.3**

Table 2: Accuracy of XBert with varying latent size of adapters. The evaluation is made on the XNLI dataset with text classification as downstream task.

5.3 ALBEF results

The results showcased in Table 3 presents the accuracy for different versions of the ALBEF model on three different languages.

ALBEF^{base}: This version of ALBEF is the base model that has not been specifically trained on the downstream task. As expected, the accuracies obtained with this model are relatively low due to the lack of task-specific training.

ALBEF^{pretrained}: In this version, ALBEF underwent full pretraining on the downstream task, but exclusively for the English language. The model was then evaluated on the other languages in a zero-shot cross-lingual setting.

ALBEF^{base} + L/T$_{256}$: This version is the base version of ALBEF with the incorporation of language and task adapters.

Based on these results, we observe that the complete pretraining of ALBEF on the specific downstream task yields improved accuracy, particularly for the English language. Furthermore, incorporating language and task adapter modules enhances the model's ability to adapt to different languages, resulting in improved performance for French and Spanish.

Model	en	fr	es
ALBEF^{base}	33.1	34.3	34.2
ALBEF^{pretrained}	**78.7**	37.6	34.2
ALBEF^{base} + L/T$_{256}$	78.1	**45.7**	**38.3**

Table 3: Accuracy of different training configuration of ALBEF on three different languages: English (en), French (fr) and Spanish (es). The evaluation is made on the XVNLI dataset with cross-lingual visual entailment as downstream task.

6 Conclusion

In conclusion, this project has presented a comprehensive analysis and evaluation of the impact of

adapter modules in Large Language Models and Vision-Language Models on cross-lingual transfer learning.

The results obtained indicate that the choice of pretrained models plays a significant role in achieving high accuracy. The XBert^{mBert} model, leveraging the BERT-based-multilingual-uncased pretrained model, demonstrated superior performance across multiple languages. This highlights the advantages of using multilingual models for multilingual tasks.

Furthermore, the incorporation of language adapters and task adapters showcased the potential for performance improvements. The addition of adapters, especially with appropriate latent sizes, enhanced accuracy for certain languages.

Overall, this project contributes to the understanding of multimodal cross-lingual transfer learning and offers insights into the effectiveness of adapters module used in different models and configurations.

7 Limitations and Future work

In order to enhance the overall effectiveness of our approach, it would be advantageous to replace our current text encoder, XBert, with a more performing model like XLM-R (Conneau et al., 2020). XLM-R integrates a larger token vocabulary, making it a superior choice for our needs. Numerous studies have demonstrated that using XLM models yields improved outcomes in cross-lingual transfer learning (Lample and Conneau, 2019).

Moreover, our investigation into the effects of adapter latent size, as depicted in Figure 2, revealed a significant impact on performance. To further explore the potential benefits of increasing the latent size, it would be beneficial to conduct experiments using a multilingual model rather than restricting ourselves to only English pre-trained models, as we did in our previous analysis in Section 5.2. By doing so, we could gain deeper insights of the impact of this parameter and potentially uncover new avenues for enhancing our method's effectiveness.

It is important to acknowledge that this study focused exclusively on text classification and visual entailment tasks. Conducting evaluations on a broader range of downstream tasks would significantly enhance our understanding of the method's overall performance.

Finally, expanding the evaluation to include non-Latin alphabet languages, such as Chinese or Arabic, would provide valuable insights into the model's adaptability and efficacy across different writing systems. These languages present unique challenges due to their complex character structures and linguistic characteristics. By assessing the model's performance on these languages, we could ascertain its ability to handle diverse scripts and further enhance its cross-lingual capabilities.

References

Željko Agić and Natalie Schluter. 2018. Baselines and test data for cross-lingual inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. *ArXiv*, abs/2201.11732.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. 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. ArXiv:1810.04805 [cs].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image

is worth 16x16 words: Transformers for image recognition at scale.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Amir Reza Jafari, Behnam Heidary, Reza Farahbakhsh, Mostafa Salehi, and Mahdi Jalili. 2021. Transfer learning for multi-lingual tasks - a survey. *CoRR*, abs/2110.02052.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ArXiv:2201.12086 [cs].

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. ArXiv:2107.07651 [cs].

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually Grounded Reasoning across Languages and Cultures. ArXiv:2109.13238 [cs].

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. ArXiv:2005.00052 [cs].

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

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.