Improving the Cross-Lingual Generalisation in Visual Question Answering

Farhad Nooralahzadeh, Rico Sennrich

Department of Computational Linguistics, University of Zurich
fahrad.nooralahzadeh@uzh.ch, sennrich@cl.uzh.ch

Abstract

While several benefits were realized for multilingual vision-language pretrained models, recent benchmarks across various tasks and languages showed poor cross-lingual generalisation when multilingually pre-trained vision-language models are applied to non-English data, with a large gap between (supervised) English performance and (zero-shot) cross-lingual transfer. In this work, we explore the poor performance of these models on a zero-shot cross-lingual visual question answering (VQA) task, where models are fine-tuned on English visual-question data and evaluated on 7 typologically diverse languages. We improve cross-lingual transfer with three strategies: (1) we introduce a linguistic prior objective to augment the cross-entropy loss with a similarity-based loss to guide the model during training, (2) we learn a task-specific subnetwork that improves cross-lingual generalisation and reduces variance without model modification, and (3) we augment training examples using synthetic code-mixing to promote alignment of embeddings between source and target languages. Our experiments on xGQA using the pretrained multilingual multimodal transformers UC2 and M3P demonstrate the consistent effectiveness of the proposed fine-tuning strategy for 7 languages, outperforming existing transfer methods with sparse models.

Introduction

Multimodal pretraining has established state-of-the-art performance for many multimedia tasks such as image-text retrieval, visual question, and answering, video localization, speech recognition, etc. Pretraining models outperforms traditional methods by providing stronger representation of different modalities learned in an unsupervised training fashion (e.g. Radford et al. 2021; Schneider et al. 2019; Sun et al. 2019). However, progress in this area has been limited mostly to the English language, whereas the main multimodal datasets consist only of English data. In order to generalize this achievement to non-English languages, recent works (e.g. Zhou et al. 2021; Ni et al. 2021; Liu et al. 2021; Bapna et al. 2022) attempt to learn universal representations to map objects that occurred in different modalities or texts expressed in various languages into shared semantic space.

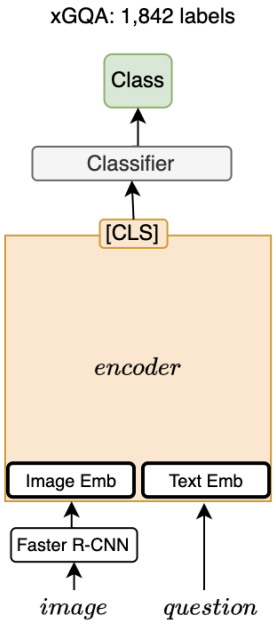

Figure 1: A standard setup (Bugliarello et al. 2022; Pfeiffer et al. 2022) to perform VQA task using UC2 or M3P.

IGLUE (Bugliarello et al. 2022), a recent benchmark spanning various tasks and languages has shown that performance degrades significantly when existing multilingual vision-language models are applied to non-English data, and there is a large gap between supervised performance and (zero-shot) cross-lingual transfer. This gap is most noticeable for resource-poor languages and languages that are distinct from English, attributed mostly to misalignment of text embeddings between the source and target languages (Liu et al. 2021; Pfeiffer et al. 2022).

In this work, we address a number of deficiencies in how these multilingual vision-language models are trained and evaluated on xGQA (Pfeiffer et al. 2022), a multilingual evaluation benchmark for the visual question answering task, where the source English dataset is extended to 7 typologically diverse languages. Specifically, we address the fol-

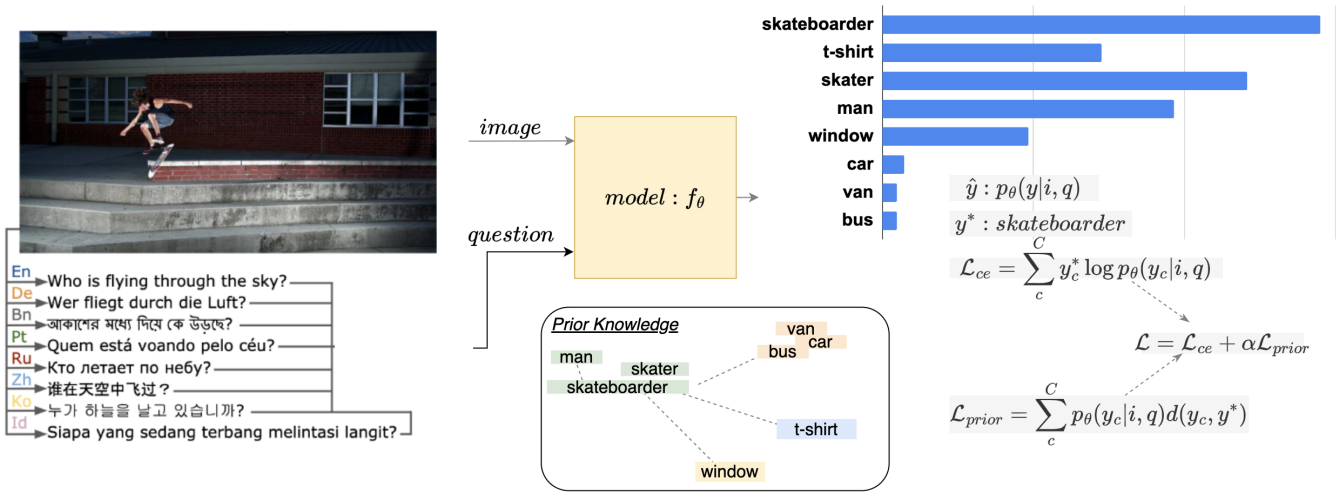

Figure 2: Illustration of how we augment the cross-entropy loss with a similarity-based loss using a linguistic prior knowledge in VQA task.

lowing issues: (i) The standard cross-entropy loss function fails to assess properly the different incorrect model outputs and results in treating equally all incorrect predictions during training, (ii) The label space is highly derived from the source language (i.e. English), resulting in language bias in the training material and hurting generalization to other languages, and (iii) The non-restricted fine-tuning of multilingual vision-language models likely neglect the task-specific and language-neutral components, resulting in over-fitting on the source language and poor cross-lingual generalisation. Our contributions are as follows:

1. We design an effective fine-tuning strategy by incorporating the linguistic prior, task-specific sparse subnetwork, and synthetic code-mixing augmentation to address the low performance of pretrained multilingual vision-language models on cross-lingual VQA task. Our strategy does not introduce extra trainable parameters or layers, and even reduces the number of model parameters. Code and data to reproduce our findings are publicly available.[1]

2. We evaluate the proposed strategy on cross-lingual zero-shot learning, across a total of 7 languages and observe consistent improvements over strong multilingual multimodal transformers including UC2 (Zhou et al. 2021) and M3P (Ni et al. 2021), achieving a substantial $+13.12\%$ and $+12.63\%$ gain in average accuracy over all languages in xGQA against UC2 and M3P baselines, respectively.

3. We perform an error analysis highlighting a substantial number of confusions between semantically related labels in xGQA, including synonyms, hypernyms, and hyponyms. We propose a metric that treats all synonyms of the ground truth label as correct.

[1]https://github.com/nooralahzadeh/CLG-VQA

Background

Having a pair of an image and a question, the task in visual question answering (VQA) is to provide an answer considering both modalities. This process has been treated as a classification task in most VQA benchmark datasets, where the underlying model should select one or multiple answers from a set of predefined labels. Recently, Pfeiffer et al. (2022) introduce a typologically diverse multilingual and multimodal benchmark for VQA task by extending the monolingual English-only GQA (Hudson and Manning 2019) dataset. They utilize 12,578 questions and 398 images from the test and development set of GQA, where the questions are manually translated into 7 different languages, covering 5 different scripts: Bengali (Bn), German (De), Indonesian (Id), Korean (Ko), Portuguese (Pt), Russian (Ru) and simplified Chinese (Zh). The xGQA benchmark also consists of new fixed data splits to guide cross-lingual few-shot learning experiments, where only a small number of examples in the target language are available. This dataset has been used in recent studies on cross-lingual transfer learning of vision-language models (e.g. Liu et al. 2022; Zeng et al. 2022) and includes several types of structured questions about an image. In this work, we base our approach on two state-of-the-art pretrained multilingual vision-language architectures, namely UC2 (Zhou et al. 2021) and M3P (Ni et al. 2021). These two transformer-based multimodal models accept the concatenation of image region features extracted with an object detector (i.e. Faster R-CNN (Ren et al. 2015)) and a sequence of BPE tokens (Sennrich, Haddow, and Birch 2016) representing the question using Sentence Piece model (Kudo and Richardson 2018) as an input. This input is then processed by a BERT-like encoder (Devlin et al. 2019) to obtain multimodal, contextualised representations. They are initialized from XLM-R (Conneau et al. 2020) and mainly differ in their pretraining strategy.

As Figure 1 depicts, the standard setup (Pfeiffer et al. 2022; Bugliarello et al. 2022) to perform cross-lingual VQA

task is to fine-tune the pretrained multilingual image-text model in the source language (i.e, English). Then, the representation of the [CLS] token as a multimodal and contextualized representation is fed into a non-linear two-layer feed-forward classifier head to predict an answer for a given image-question pair. For *zero-shot cross-lingual* evaluation, the fine-tuned model is evaluated on the multilingual test data, whereas in *few-shot cross-lingual* scenario, the fine-tuned model is additionally trained on image-question examples available in the target language.

Fine-Tuning Strategies

Multilingual vision-language pretrained models often suffer from poor cross-lingual generalisation compared with their corresponding monolingual baseline, achieving much better performance in the source language than in target languages unseen during fine-tuning. In this work, we aim to address the poor performance of these models on the xGQA benchmark, where models are fine-tuned on English data and evaluated on 7 typologically diverse languages. In particular, we investigate the impact of three fine-tuning strategies: (i) Incorporating linguistic prior, (ii) Task-specific sparse fine-turning, and (iii) Multilingual Code-Switching (i.e. Code-Mixing) data augmentation. In this section, we describe these three strategies in detail.

Incorporating Linguistic Prior We realize a number of deficiencies in how multilingual vision-language models are trained and evaluated cross-lingually in the VQA task. (i) The loss function fails to assess properly the different incorrect model outputs and results in treating equally all incorrect predictions during training, (ii) Training examples are only annotated with one label, where intuitively multiple labels are often plausible (e.g. lady vs. woman; couch vs. sofa), and (iii) The label space highly depends on the source language (i.e. English) and hurts generalization to other languages. For instance, there are singular and plural labels such as *car/cars*, *woman/women* and *laptop/laptops* while in some target languages such as Chinese, most nouns are not marked for grammatical numbers. In this section, we aim to address these issues.

Given an image i and a question q, the respective model f_θ for VQA task provides a probability distribution $\hat{y} = p_\theta(y|i,q)$ over a set of predefined answers. Commonly, VQA models are trained using the cross-entropy loss, in which parameters of the underlying model θ are optimized using the following objective function:

$$\mathcal{L}_{ce} = \sum_{c=1}^{C} y_c^* \log p_\theta(y_c|i,q)$$

where C is the number of classes in the answer set, and y^* is the one-hot vector that represents the ground truth answer. The objective loss function encourages the model to give a large probability mass to a correct class. It compares the predicted and ground-truth label and takes a once-for-all matching strategy, consequently evaluating all predictions as either correct or incorrect and ignoring the similarity between the correct and less incorrect predictions. As an example in the question-image pair shown in Figure 2, if the model receives a question as *Who is flying through the sky?* and the ground truth label is *skateboarder*, the underlying loss function will penalize the wrong predicted labels such as *skater*, *man*, *t-shirt* or *car* equally. We argue that the incorrect training predictions may be quite diverse and letting the model be aware of which incorrect predictions are more incorrect or less incorrect than others may more effectively guide the model during the training. Therefore, in our example, similar labels such as *skater* and *man* should be penalized much less than dissimilar words like *t-shirt* or *car*.

In order to alleviate the issue, we add a linguistic prior objective to augment the cross-entropy loss with a similarity-based loss. The loss can be conceived as a form of risk minimization, where the risk function is the distance d between a ground truth label y^* and the predicted label y_c. In other words, the objective function should give a small loss if the predicted and ground truth label are similar, and penalize dissimilar answers:

$$\mathcal{L}_{prior} = \sum_{c} p_\theta(y_c|i,q) d(y_c, y^*)$$
$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{prior}$$

The risk d is weighted by the probability distribution over all target labels $p_\theta(y|i,q)$, provided by the classification layer. We formalize the distance score $d(y_c, y^*)$ between the ground truth label and others in the label space by using two sources of linguistic knowledge:

WordNet (prior_{wn}): We extract the explicit relations among the labels using the synset structure of the English lexical database (i.e. WordNet (Fellbaum 1998)). To be more precise, we derive the synonymy, hyponymy, and hypernymy relations and formulate the distance as:

$$d(y_c, y^*) = \begin{cases} 0 & \text{if } y_c \text{ and } y^* \text{ are } synonyms \\ d_1 & \text{if } y_c \text{ is } hyponym \text{ of } y^* \\ d_2 & \text{if } y_c \text{ is } hypernym \text{ of } y^* \\ 1 & \text{otherwise} \end{cases}$$

, where $0 < d_1, d_2 < 1$.

Word Embeddings (prior_{em}): A distance d is extracted from implicit semantic proximity within pretrained word embeddings. We calculate the embeddings cosine distance as the distance of y^* and all other labels as:

$$d(y_c, y^*) = CosineDistance(emb_{y^*}, emb_{y_c})$$

Task-specific Sparse Fine-tuning (SFT) The success of multilingual pretrained models in cross-lingual genralisation is often attributed to task-specific and language-neutral components, which capture commonalities among languages (Libovický, Rosa, and Fraser 2020; Foroutan et al. 2022).

To this end, we are inspired by previous works (Frankle and Carbin 2019; Chen et al. 2020; Ansell et al. 2022) that claim there exists a sparse, separated trainable subnetwork (i.e. a winning ticket) capable to match or even outperform the original neural network. Similarly, we design a task-specific sparse fine-tuning strategy, here dubbed SFT, consisting of two steps:

Algorithm 1: Iterative Magnitude Pruning (IMP) with rewinding step (Han, Mao, and Dally 2016).

Input: Model $f(.;\theta)$ initialized with pretrained parameters θ^0.
Parameter: $p\%$: a pruning rate
Output: M

1: Set the initial pruning mask to $M = 1^{|\theta|}$.
2: **while** not done **do**
3: Train $f(.; M \odot \theta^0)$ to step t: $f(.; M \odot \theta^t)$.
4: Prune $p\%$ of remaining weights of $M \odot \theta^t$ and update M accordingly.
5: **end while**
6: Return $f(.; M \odot \theta^0)$.

$Step_0$**:** Considering the VQA model $f(.;\theta)$ initialized with pretrained weights θ^0, we obtain a subnetwork $f(.; M \odot \theta)$ where $M \in \{0,1\}^{|\theta|}$ represents a binary mask and $\odot$ is element-wise multiplication. More specifically as it is shown in Algorithm 1, we utilize *Iterative Magnitude Pruning* (IMP) (Han, Mao, and Dally 2016) to discover the pruning mask M, during the fine-tuning of the VQA model in English-only data. After each epoch, we prune a certain amount (e.g. $p\%$) of the original parameters. Then, we continue the fine-tuning by resetting the remaining parameters to their original value on the pretrained initialization θ^0.

$Step_1$**:** Having the pruning mask M, the model parameters are initialized with their original values θ^0 and are fine-tuned again. However, in this step, only the unmasked parameters are trained while the masked ones are kept frozen. It should be mentioned that following previous works (Zhou et al. 2019; Chen et al. 2020) the masked parameters are set to zero.

Code-Mixing (CDM) While each fine-tuning step only involves questions from the English language, the VQA task is unable to benefit properly from cross-lingual alignment information that exists in multilingual vision-language models. To make full use of this cross-lingual alignment information and better fine-tuning, we construct code-mixed data in target languages. To generate the code-mixed questions, we follow the mechanism of multilingual code-switching data augmentation (CoSDA) proposed by Qin et al. (2020). First, a set of words is randomly chosen in each question. Second, for each selected word, we randomly specify a target language to translate. Third, we replace the word with its translation in the selected language. If the word has multiple translations in the target language, then one of them is randomly selected for replacement. To increase the data diversity during the training, Qin et al. (2020) proposes to reset the replacement after each epoch and to replace different words at different epochs.[2] Figure 3 shows the result of applying the code-mixing procedure to our example.

[2]For further details regarding CoSDA we refer the reader to the original work.

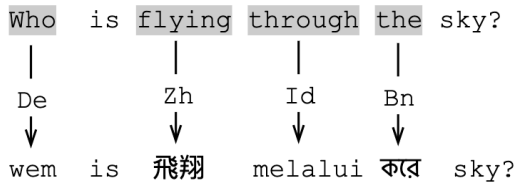

Figure 3: The code-mixed question, where a set of 4 words is randomly selected in order to be replaced by their translation using bilingual dictionaries of MUSE (Lample et al. 2018) into 4 randomly selected target languages in xGQA.

Experiments

To evaluate our proposed strategies, as explained in Section , we benchmark two state-of-the-art multilingual vision-language transformers, namely UC2 and M3P, as the base models. We study the impact of each strategy by fine-tuning the model on the monolingual English GQA dataset[3], then evaluating the cross-lingual transfer on the multilingual extension of GQA, known as xGQA. We adopt the codebase of IGLUE benchmark[4] to implement our proposed approach and we keep the value of the models and training hyper-parameters equal to the ones that are reported by Bugliarello et al. (2022). The results are reported for each experiment by averaging the performance over five different runs.

Model Configurations and Notation

On both UC2 and M3P models, we experiment with three different setups:

With prior$_{xx}$**:** The fine-tuning process is performed using a similarity-based loss together with cross-entropy loss. The prior knowledge based distances d are computed as follows: (i) $prior_{wn}$: The WordNet base distance is computed by using the NLTK library (Loper and Bird 2002), and (ii) $prior_{em}$: To compute the cosine distance among 1842 labels in xGQA, we use the spaCy[5] toolkit, where an embedding $emb_y \in \mathbb{R}^{300}$ of each label is derived from GloVe (Pennington, Socher, and Manning 2014) pretrained word embeddings.

To mitigate the negative influence of non-probable classes on the similarity-based loss, we consider the k most probable answers according to their probability $p_\theta(y_c|i,q)$ in both setups. We set hyper-parameters $d_1 = 0.8, d_2 = 0.8, k = 10$ and $\alpha = 10$ based on validation set performance.[6]

+SFT: Using PyTorch pruning module[7], we extract the subnetwork from the pretained weights θ^0 following Algorithm 1 and $Step_0$ of SFT strategy. More specifically, we

[3]We consider *balanced* subset of GQA as recommended in IGLUE benchmark (Bugliarello et al. 2022)

[4]https://iglue-benchmark.github.io/

[5]https://spacy.io/

[6]We performed a grid search using different values for these hyper-parameters. Note that $\mathcal{L}_{prior}$ is typically much smaller than $\mathcal{L}_{ce}$, hence the large α.

[7]https://pytorch.org/tutorials/intermediate/pruning_tutorial.html

	Model	En	Bn	De	Id	Ko	Pt	Ru	Zh	Avg
	Fine-tune model on English training set (Zero-Shot)									
UC2	Our Baseline	54.92	19.99	42.00	28.44	22.40	30.92	28.55	31.19	29.07
	Bugliarello et al. (2022)	55.19	19.98	42.85	28.67	21.36	30.41	30.99	31.15	29.35
	Liu et al. (2022)	58.57 ±0.2	26.23 ±1.5	49.51 ±1.1	38.92 ±1.3	36.48 ±1.3	39.76 ±0.6	41.72 ±0.3	**46.52** ±0.9	39.87
	With prior$_{wn}$	55.77 ±0.0	23.66 ±0.8	47.93 ±0.2	35.67 ±1.4	34.57 ±1.8	37.46 ±1.3	40.08 ±0.5	40.08 ±4.3	37.06
	With prior$_{em}$	56.09 ±0.1	23.97 ±2.6	48.13 ±0.8	36.87 ±1.9	34.14 ±3.6	38.18 ±2.5	41.07 ±0.9	41.76 ±1.9	37.73
	With prior$_{em}$+ SFT	56.56 ±0.1	23.53 ±2.0	49.54 ±0.3	36.79 ±0.5	34.56 ±0.5	38.95 ±0.2	41.18 ±0.2	43.40 ±0.2	38.28
	With prior$_{em}$ + CDM	54.37 ±0.0	27.38 ±0.0	46.66 ±1.7	20.88 ±2.3	36.32 ±1.1	40.81 ±2.1	43.48 ±0.2	30.62 ±1.5	35.16
	With prior$_{em}$ + SFT+CDM	55.21 ±0.1	**30.96** ±1.3	**50.30** ±0.2	**41.68** ±0.7	**39.57** ±0.6	**43.43** ±0.6	**44.58** ±0.9	44.80 ±0.8	**42.19**
M3P	Our Baseline	54.02	17.24	32.40	23.77	25.57	32.91	32.32	27.39	27.37
	Bugliarello et al. (2022)	53.75	18.64	33.42	32.48	25.11	31.40	27.50	28.65	28.17
	Liu et al. (2022)	46.70 ±0.7	29.75 ±1.4	39.52 ±1.3	36.73 ±1.6	35.67 ±1.1	37.59 ±0.8	37.93 ±0.9	36.15 ±0.9	36.19
	With prior$_{wn}$	55.91 ±0.2	22.38 ±0.4	39.48 ±1.7	29.31 ±2.3	35.15 ±0.9	39.00 ±0.2	38.92 ±0.3	35.74 ±0.8	34.28
	With prior$_{em}$	56.33 ±0.1	22.93 ±3.2	40.10 ±0.5	30.63 ±0.0	35.35 ±2.1	38.85 ±1.1	39.95 ±0.0	36.97 ±0.1	34.97
	With prior$_{em}$ + SFT	56.18 ±0.0	22.07 ±0.5	40.29 ±0.2	27.04 ±0.1	34.62 ±0.1	38.39 ±0.2	39.44 ±0.0	36.32 ±0.5	34.02
	With prior$_{em}$ + CDM	54.35 ±0.6	28.71 ±1.7	43.57 ±0.3	**38.89** ±2.1	38.06 ±0.4	41.93 ±0.6	41.64 ±1.1	38.80 ±1.2	38.80
	With prior$_{em}$ + SFT+CDM	55.58 ±0.1	**31.53** ±1.5	**46.19** ±0.5	34.60 ±0.5	**40.21** ±0.9	**42.87** ±0.6	**42.32** ±1.2	**42.25** ±0.6	**40.00**
	Translate everything to English and use the English-only model (Translate-Test)									
UC2	(Bugliarello et al. 2022)	55.19	49.31	52.61	50.34	48.62	52.17	49.95	48.32	50.19
M3P	(Bugliarello et al. 2022)	53.75	47.79	51.01	49.35	47.64	51.21	47.76	47.04	48.83

Table 1: Accuracy results on the xGQA test set for Zero-Shot transfer. Columns indicate the target languages. We also report the average (Avg.) accuracy across languages excluding English. For our baseline, we fine-tuned the model on the English balanced subset of GQA and evaluated it on the test set of xGQA. In *With prior$_{xx}$*, the original cross-entropy loss is augmented with a similarity-based loss, either using WordNet (i.e. *With prior$_{wn}$*) or Word Embeddings (i.e. *With prior$_{em}$)*. In *With prior$_{em}$*+SFT, we apply task-specific sparse fine-tuning strategy along with Word Embeddings based loss. In *With prior$_{em}$*+SFT+CDM is our final design where we employ the code-mixing augmentation on top of the previous strategy. For comparison, we report results from Bugliarello et al. (2022) and Liu et al. (2022). The performance of all proposed strategies is averaged from five runs with different random seeds.

consider metrics of the encoder part of the model (see Figure 1), excluding the image and text embeddings as well as the classifier layer in both UC2 and M3P.

Since the network architecture varies between UC2 and M3P, the pruning is applied to a different set of parameters. We perform IMP and prune a set of weights with the lowest-magnitude globally throughout the network after each fine-tuning epoch (number of epochs=5). Based on preliminary experiments, we iteratively prune a certain fraction of the lowest-magnitude weights (i.e. $p = 10\%$) at each epoch which results in the final sparsity level of around 40% in both models. Considering the exclusion of some parameters, the level of sparsity is 12.28% for UC2 and 13.44% for M3P.[8] As our focus is not on conducting a large-scale analysis over different sparsity levels, we leave this topic for future work.

Furthermore, following *Step*$_1$ of the SFT strategy, we fine-tune the model using the pruning mask M. In both steps, we incorporate similarity-based loss using word embeddings prior (i.e. *prior$_{em}$*) in our experiments. This effectively leads to a two-stage pruning and sparse fine-tuning process, termed as *With prior$_{em}$*+SFT.

[8]UC2 has around 281.66 M parameters where 85.52 M of them are involved during the pruning process. M3P has 376.90 M parameters and 123.67 M of them are considered for pruning.

+CDM: To perform code-mixing, the English questions and the bilingual dictionaries of MUSE (Lample et al. 2018) are used as the basis. We use all 7 target languages in xGQA during the code-mixing augmentation. To perform *With prior$_{em}$*+SFT+CDM experiments, we continue with fine-tuning by applying the code-mixing during the *Step*$_1$ of SFT after pruning the model according to the *Step*$_0$ of SFT. We find that including code-mixing during the pruning step (i.e. *Step*$_0$) negatively impacts the model performance in the experiments that follow *With prior$_{em}$*+SFT+CDM strategy.

Baselines and Previous Results:

We create *Our baselines* by directly evaluating the monolingual fine-tuned models on the test set of the target languages. For each model, we report another baseline using the results in Bugliarello et al. (2022). Moreover, we compare our model with a previous study (Liu et al. 2022), where the low performance of multilingual vision-language models (i.e. UC2 and M3P) in the xGQA dataset has been addressed through sophisticated classification architectures, fine-tuning strategies, and modifications of the model input via question-type conditioning. In addition, we report results of *Translate-Test* setup from Bugliarello et al. (2022) where target language test data is translated to English and

an English-only fine-tuned model is evaluated on the translated test set.

Results and Discussion

In this section, we report the results of the different fine-tuning strategies. The proposed strategies result in the best-performing models across all 7 target languages in the cross-lingual visual question answering task. A summary of the results with various strategies is provided in Table 1.

With prior$_{xx}$: Our first set of experimental results shows the advantage of using the proposed loss along with the standard cross-entropy loss for the VQA task. The proposed strategy (i.e. *With prior$_{xx}$*) improves the average cross-lingual zero-shot transfer accuracy by $+7.99$ and $+8.66$ points over the UC2 baseline using WordNet and GloVe embeddings, respectively. At the same time, it shows gains of $+6.9$ and $+7.6$ absolute accuracy points using different modeling choice (i.e. M3P) with *prior$_{wn}$* and *prior$_{em}$*, respectively. The results indicate that the similarity-based loss obtained from linguistic priors can effectively guide the models during training. They also support our hypothesis that incorporating additional semantic prior knowledge about the label space improves the cross-lingual generalisation. Among the proposed semantic distances, the GloVe embeddings-based distance delivers the greatest improvements in almost all languages. One major conceptual difference between our WordNet and GloVe-based distance that could explain this difference in performance is that the former is sparse and heuristic, whereas the latter is dense and continuous. GloVe will also capture relations such as antonym labels (e.g. male/female, boy/girl, or yes/no).

With prior$_{em}$+SFT: The results demonstrate the importance of a task-specific sparse fine-tuning strategy (i.e. SFT) for adapting the multilingual vision-language models in the downstream VQA task without modifications to the model. The SFT strategy brings further improvements (i.e $+0.55$) over the *With prior$_{em}$* strategy for UC2. Even though it does not surpass the previous strategy in M3P and provides slightly lower performance for some of the target languages in UC2, such as Bangali (Bn) and Indonesian (Id), it yields considerably more stable (lower variance) performance across random seeds in all 7 target languages. It is also worth noting that the SFT strategy offers a task-specific and parameter-efficient structure for both models, where a fraction of the encoder's parameters (12.28% of parameters in UC2 and 13.44% of parameters in M3P) are masked and ignored during the fine-tuning. These results suggest that SFT is successful in discovering language-neutral and task-specific parameters that generalise well cross-lingually for xGQA, similar to the finding by Foroutan et al. (2022) for text-only tasks.

With prior$_{em}$+SFT+CDM: The highest zero-shot transfer performance observed in our experiments is obtained by leveraging the code-mixing strategy on top of the previous best strategy (i.e. With prior$_{em}$+SFT). This strategy achieves much better performance than the previous strategies by a large margin on both transformer models compared to the baselines. The improvement is $+13.12$ and $+12.63$ in average accuracy compared to UC2 and M3P baselines, respectively. It can be observed that this approach outperforms the previous work by Liu et al. (2022), across most of the target languages with better performance and lower variance. Notably, our final strategy provides 42.19 versus 39.87 for UC2 model and 40.00 versus 36.19 for M3P model in terms of averaged accuracy across 7 languages. This confirms that our approach can better adapt the multilingual vision-language models for the cross-lingual VQA task.

We further aim to understand the impact of CDM in isolation where we do not perform SFT. It can be seen in Table 1, applying CDM as the only strategy results in a large performance drop for UC2 model in some of the target languages, especially in Indonesian (Id) and Chinese (Zh). It also leads to higher variance compared to its counterpart which only benefits from the SFT strategy in both models. This result demonstrates synergies between the proposed strategies, with the combination of CDM, which promotes alignment of word representations between source and target languages, and SFT, which discovers subnetworks that may be more language-neutral, achieving a large improvement in combination whereas effects are more moderate (or negative) when applied in isolation. It is worth to note that we also conduct experiments with only CDM strategy (i.e. excluding the With prior$_{em}$ strategy). However, the results were lower than applying the With prior$_{em}$+CDM (e.g Avg=32.76 compare to Avg=35.16 using UC2).

	Model	Avg.		
		w/o Syn.	with Syn.	Diff.
UC2	Our Baseline	29.07	29.96	+0.89
	With prior$_{wn}$	37.06	38.91	+1.85
	With prior$_{em}$	37.73	39.06	+1.33
	With prior$_{em}$+ SFT	38.28	39.67	+1.39
	With prior$_{em}$ + SFT + CDM	**42.19**	**43.90**	+1.71
M3P	Our Baseline	27.37	31.83	+4.56
	With prior$_{wn}$	34.28	37.70	+3.42
	With prior$_{em}$	34.97	38.85	+3.88
	With prior$_{em}$ + SFT	34.02	38.25	+4.23
	With prior$_{em}$ + SFT + CDM	**40.00**	**43.52**	+3.52

Table 2: Results of adjusting the evaluation metric to consider the synonym of the target label as a correct prediction (*with Syn.*). The *w/o Syn.* column indicates the results before the adjustment.

Further Analysis

To further investigate the effect of synonymy relations among the target labels on xGQA evaluation results, we

Model	L	4 most-confused labels label:prediction (rel.)							
BL	En	girl:woman	27	material:color	23	lady:woman	18	coffee table:table	17
	Bn	sailboats:sailboat	3	skater:skateboarder	3	plain:field	2	trees:tree	2
	De	girl:woman	33	material:color	21	lady:woman	16	woman:girl	13
	Id	girl:woman	28	lady:woman	18	skater:skateboarder	15	woman:girl	14
	Ko	girl:woman	7	skater:skateboarder	7	boy:man	2	fire truck:truck	2
	Pt	girl:woman	22	skater:skateboarder	17	lady:woman	13	zebras:zebra	12
	Ru	girl:woman	32	skater:skateboarder	17	lady:woman	17	woman:girl	14
	Zh	girl:woman	26	chairs:chair	15	cabinets:cabinet	15	skater:skateboarder	15
Best	En	girl:woman	28	material:color	24	cabinets:cabinet	20	woman:girl	18
	Bn	cabinets:cabinet	29	girl:woman	19	skater:skateboarder	15	woman:girl	12
	De	girl:woman	32	material:color	23	lady:woman	18	cabinets:cabinet	17
	Id	girl:woman	27	cabinets:cabinet	24	woman:girl	17	chairs:chair	17
	Ko	cabinets:cabinet	39	girl:woman	34	elephants:elephant	20	woman:girl	17
	Pt	material:color	25	girl:woman	24	woman:girl	20	zebras:zebra	15
	Ru	girl:woman	33	cabinets:cabinet	25	material:color	19	woman:girl	18
	Zh	cabinets:cabinet	32	girl:woman	27	chairs:chair	26	zebras:zebra	25

Table 3: The 4 most-confused labels for each language, specifically where the UC2 model predicts a *synonym*, *hypernym*, or *hyponym* of the target label. The number of "wrong" predictions that are in a synonym/hypernymy/hyponymy relationship with the ground truth label is reported in separate columns. Results shown for baseline (BL) and our best strategy.

modify the evaluation metric to consider synonyms of the ground truth label as a correct prediction. We use the WordNet synonym synset for this purpose. For instance we consider *couch* correct if the ground truth label is *sofa*, or *girls* if the ground truth label is *girl*. We note that confusion between synonymous labels is relatively common in xGQA; if we consider synonyms to be correct answers, model performance is actually higher than reported by the original accuracy by 0.9-1.85 and 3.42-4.56 percentage points with UC2 and M3P, respectively (see Table 2).

Table 3 shows the 4 most-confused labels for each language, specifically where the UC2 model predicts a synonym, hypernym, or hyponym of the target label. While synonyms are predominantly due to inflectional differences (singular/plural), we also find a large number of "wrong" predictions that are in a hypernymy/hyponymy relationship with the ground truth, and semantically plausible (*girl* or *lady* vs. *woman*). Although the confusion between similar labels motivated our use of linguistic priors, the performance improvement that we observe is not predominantly due to a reduction in this confusion. In fact, with our best strategy, the number of "wrong" predictions that are semantically plausible even increases for UC2, especially for some low-resource languages such as Bengali (Bn) and Korean (Ko), which we take as a positive result: our strict accuracy results in Table 1 already show a substantial improvement for these languages, and with a more permissive evaluation metric, gains over the baseline would be even greater. Similar results are observed when we only take into account synonymy relationships.

Related Works

The primary motivation for this work is the low cross-lingual generalization of multilingual vision-language pre-trained models. There are a number of works addressing this problem. Zeng et al. (2022) introduce cross-view language modeling by considering both image-caption pairs and parallel sentence pairs as two different views of the same object and train the model to align the two views by maximizing the mutual information between them with conditional masked language modeling and contrastive loss. Whereas they report a state-of-the-art zero-shot cross-lingual performance for xGQA, their method demands a pretraining step as well as high computing resources and multilingual language-vision datasets. In contrast, our proposed strategy can be applied on top of any multilingual vision-language pretrained model as an adaptation step. Our approach is similar to the work by Liu et al. (2022), where they propose a set of methods that improves previously low transfer performance and thus substantially reduce the gap to monolingual English performance. However, their approach is more complex and our final strategy provides better performance with a sparse encoder.

Similarity-based loss: There is an increasing interest in in-

corporating prior domain knowledge in neural NLP downstream tasks. Prior knowledge of the language has been applied recently to language generation learning. Li et al. (2020) introduces a technique that imposes the prior from (linguistic) data over the sequence prediction models and improves performance in typical language generation tasks, including machine translation, text summarization, and image captioning. Chousa, Sudoh, and Nakamura (2018) propose a novel NMT loss function that includes word similarity in forms of distances in a word embedding space and it leads to a substantial gain in the machine translation task.

Sparse fine-tuning: Our approach is inspired by studies of sparse fine-tuning methods (Ansell et al. 2022; Liang et al. 2021; Foroutan et al. 2022) . Ansell et al. (2022) and Liang et al. (2021) claim that non-restricted fine-tuning of multilingual models is prone to over-fitting on source language as well as catastrophic forgetting. They suppose that parameter interference is one of the causes of this degradation. Foroutan et al. (2022) suggest that language-specific and language-neutral subnetworks play a prominent role in the cross-lingual generalisation of the multilingual language model (i.e Multilingual BERT). In this work, we follow the above-mentioned ideas by looking at the structure and weights of multilingual vision-language models in the VQA task.

Code-Switching: Data augmentation training using Code-Switching offers a significant improvement to the low-resource languages. It helps the model explicitly learn the relationship among words in different languages. It has been applied to the training of various multimodal multilingual models such as M3P (Ni et al. 2021) and CCLM (Zeng et al. 2022). Raj Khan, Gupta, and Ekbal (2021) create a multilingual and code-mixed VQA dataset in eleven different language setups considering the multiple Indian and European languages as well as their code-mixed versions. They propose a knowledge distillation to extend an English language-vision model (teacher) into a multilingual and code-mixed model (student). However, this dataset is not diverse as xGQA in terms of covering the low-resource languages.

Conclusion

We present a series of strategies to fine-tune multilingual vision-language pretrained models for better cross-lingual generalisation in the visual question answering task. Our approach is based on various adaptation techniques aimed to mitigate the number of issues that we discovered regarding the training and evaluation of multilingual vision-language models on xGQA. Comparing our approach with the baseline and previous similar work in several pretrained models, the results indicate substantial improvements across target languages. The improvement is $+13.12$ and $+12.63$ in average accuracy over all 7 languages in xGQA compared to UC2 and M3P baselines, respectively.

We perform an analysis of closely related target labels in xGQA, proposing a new metric that rewards synonymous predictions and further demonstrates the success of the proposed strategies. This analysis also highlights the need for future research on the label space and evaluation metrics for cross-lingual VQA.

Acknowledgments

We would like to thank Xin Sennrich and Alham Fikri Aji for their helpful feedback on language resources. This work was funded by the Swiss National Science Foundation (project MUTAMUR; no. 176727) at the University of Zurich.

References

Ansell, A.; Ponti, E.; Korhonen, A.; and Vulić, I. 2022. Composable Sparse Fine-Tuning for Cross-Lingual Transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1778–1796. Dublin, Ireland: Association for Computational Linguistics.

Bapna, A.; Cherry, C.; Zhang, Y.; Jia, Y.; Johnson, M.; Cheng, Y.; Khanuja, S.; Riesa, J.; and Conneau, A. 2022. mSLAM: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*.

Bugliarello, E.; Liu, F.; Pfeiffer, J.; Reddy, S.; Elliott, D.; Ponti, E. M.; and Vulić, I. 2022. IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2370–2392. PMLR.

Chen, T.; Frankle, J.; Chang, S.; Liu, S.; Zhang, Y.; Wang, Z.; and Carbin, M. 2020. The lottery ticket hypothesis for pre-trained bert networks. *Advances in neural information processing systems*, 33: 15834–15846.

Chousa, K.; Sudoh, K.; and Nakamura, S. 2018. Training neural machine translation using word embedding-based loss. *arXiv preprint arXiv:1807.11219*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Foroutan, N.; Banaei, M.; Lebret, R.; Bosselut, A.; and Aberer, K. 2022. Discovering Language-neutral Subnetworks in Multilingual Language Models. *ArXiv*, abs/2205.12672.

Frankle, J.; and Carbin, M. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*.

Han, S.; Mao, H.; and Dally, W. J. 2016. Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. In Bengio, Y.; and LeCun, Y., eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6693–6702.

Kudo, T.; and Richardson, J. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 66–71. Brussels, Belgium: Association for Computational Linguistics.

Lample, G.; Conneau, A.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Li, Z.; Wang, R.; Chen, K.; Utiyama, M.; Sumita, E.; Zhang, Z.; and Zhao, H. 2020. Data-dependent Gaussian Prior Objective for Language Generation. In *ICLR*.

Liang, J.; Zhao, C.; Wang, M.; Qiu, X.; and Li, L. 2021. Finding Sparse Structures for Domain Specific Neural Machine Translation. In *AAAI*.

Libovický, J.; Rosa, R.; and Fraser, A. 2020. On the Language Neutrality of Pre-trained Multilingual Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1663–1674. Online: Association for Computational Linguistics.

Liu, C.; Pfeiffer, J.; Korhonen, A.; Vulic, I.; and Gurevych, I. 2022. Delving Deeper into Cross-lingual Visual Question Answering. *ArXiv*, abs/2202.07630.

Liu, F.; Bugliarello, E.; Ponti, E. M.; Reddy, S.; Collier, N.; and Elliott, D. 2021. Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10467–10485. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Loper, E.; and Bird, S. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 63–70. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Ni, M.; Huang, H.; Su, L.; Cui, E.; Bharti, T.; Wang, L.; Zhang, D.; and Duan, N. 2021. M¡sup¿3¡/sup¿P: Learning Universal Representations via Multitask Multilingual Multimodal Pre-training. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3976–3985.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Pfeiffer, J.; Geigle, G.; Kamath, A.; Steitz, J.-M.; Roth, S.; Vulić, I.; and Gurevych, I. 2022. xGQA: Cross-Lingual Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2497–2511. Dublin, Ireland: Association for Computational Linguistics.

Qin, L.; Ni, M.; Zhang, Y.; and Che, W. 2020. CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP. In *IJCAI*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Raj Khan, H.; Gupta, D.; and Ekbal, A. 2021. Towards Developing a Multilingual and Code-Mixed Visual Question Answering System by Knowledge Distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1753–1767. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 91–99.

Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, 3465–3469.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. Berlin, Germany: Association for Computational Linguistics.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K. P.; and Schmid, C. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7463–7472.

Zeng, Y.; Zhou, W.; Luo, A.; and Zhang, X. 2022. Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training. *ArXiv*, abs/2206.00621.

Zhou, H.; Lan, J.; Liu, R.; and Yosinski, J. 2019. Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask. In *Advances in Neural Information Processing Systems*.

Zhou, M.; Zhou, L.; Wang, S.; Cheng, Y.; Li, L.; Yu, Z.; and Liu, J. 2021. UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2021)*.