Multilingual Multimodal Learning with Machine Translated Text

Chen Qiu[α] **Dan Oneață**[β] **Emanuele Bugliarello**[ç]
Stella Frank[ç,ð] **Desmond Elliott**[ç,ð]
[α]School of Computer Science and Technology,
Wuhan University of Science and Technology, China
[β] University Politehnica of Bucharest, Romania
[ç]Department of Computer Science, University of Copenhagen, Denmark
[ð]Pioneer Centre for AI, Denmark
chen@wust.edu.cn dan.oneata@speed.pub.ro {emanuele, stfr, de}@di.ku.dk

Abstract

Most vision-and-language pretraining research focuses on English tasks. However, the creation of multilingual multimodal evaluation datasets (e.g. Multi30K, xGQA, XVNLI, and MaRVL) poses a new challenge in finding high-quality training data that is both multilingual and multimodal. In this paper, we investigate whether machine translating English multimodal data can be an effective proxy for the lack of readily available multilingual data. We call this framework TD-MML: Translated Data for Multilingual Multimodal Learning, and it can be applied to any multimodal dataset and model. We apply it to both pretraining and fine-tuning data with a state-of-the-art model. In order to prevent models from learning from low-quality translated text, we propose two metrics for automatically removing such translations from the resulting datasets. In experiments on five tasks across 20 languages in the IGLUE benchmark, we show that translated data can provide a useful signal for multilingual multimodal learning, both at pretraining and fine-tuning.

1 Introduction

Vision-and-language (V&L) pretraining is the process of learning deep contextualised cross-modal representations from large collections of image–sentence pairs (Li et al., 2019; Tan and Bansal, 2019; Chen et al., 2020, *inter-alia*). These pretrained models are an excellent backbone for transfer learning to a wide range of downstream tasks, such as visual question answering (Antol et al., 2015; Gurari et al., 2018; Agrawal et al., 2022), referring expression alignment (Kazemzadeh et al., 2014; Mao et al., 2016), and image–sentence retrieval (Young et al., 2014; Lin et al., 2014). Thus far, downstream evaluations have mostly focused on English tasks due to the availability of datasets, but the recent IGLUE benchmark (Bugliarello et al., 2022) makes it now possible to evaluate models on several downstream tasks across 20 languages.

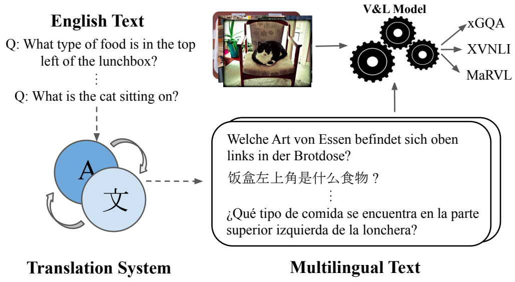

English Text
Q: What type of food is in the top left of the lunchbox?
⋮
Q: What is the cat sitting on?
V&L Model
xGQA
XVNLI
MaRVL
A
文
Welche Art von Essen befindet sich oben links in der Brotdose?
饭盒左上角是什么食物？
⋮
¿Qué tipo de comida se encuentra en la parte superior izquierda de la lonchera?
Translation System
Multilingual Text

Figure 1: Multilingual multimodal data is a scarce resource compared to English multimodal data. Given an English multimodal dataset, we generate a multilingual dataset using a black box translation system. We explore the utility of this approach to creating multilingual text for both downstream task fine-tuning and pretraining.

Success in multilingual multimodal tasks, such as those in IGLUE, is expected to depend on models with grounded representations that transfer across languages (Bugliarello et al., 2022). For example, in the MaRVL dataset (Liu et al., 2021), models need to deal with a linguistic and cultural domain shift compared to English data. Therefore, an open problem is to define pretraining strategies that induce high-quality multilingual multimodal representations. Existing work has tackled this problem by either jointly training on English multimodal data and multilingual text-only data (Liu et al., 2021; Ni et al., 2021), pretraining with a private dataset of multilingual captioned images (Jain et al., 2021), or machine translating multimodal pretraining data (Zhou et al., 2021).

In this paper, we further investigate the potential of machine translated text for both fine-tuning *and* pretraining across four diverse V&L tasks.[1] The overarching motivation is that machine translation is an inexpensive approach to producing large amounts of multilingual text compared to collecting data from humans, or scraping high-quality image–caption pairs from the web. Having access to thou-

[1] The models and the machine translated text are available at https://github.com/danoneata/td-mml

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4178–4193
December 7-11, 2022 ©2022 Association for Computational Linguistics

sands of data points in a target language might indeed be necessary to improve cross-lingual performance in downstream tasks (Bugliarello et al., 2022). As such, translating fine-tuning data into multiple languages may be a compelling approach towards downstream task success. Moreover, if this can be achieved through machine translated text, it raises the question of whether we can also pretrain on many millions of multilingual translated examples. Motivated by the initial experiments of Zhou et al. (2021), we test this hypothesis further, on more languages and more tasks, reporting more nuanced results from large-scale translated text.

Overall, we show that machine translation can provide inexpensive and impressive improvements when fine-tuning models for multilingual multimodal tasks. Moreover, translation-based pretraining leads to significant gains in zero-shot cross-lingual transfer over existing approaches. However, we find mixed results when combining this with multilingual fine-tuning. There are still opportunities to realise further benefits from machine translated text, which may be found through more compute-intensive pretraining.

Contributions. **1)** We present the TD-MML framework to narrow the gap between English and non-English languages in multimodal research. **2)** In the process of translation-based pretraining, we present a reliable strategy to filter out bad translations. **3)** We conduct systematic evaluations in zero-shot and machine translated scenarios, and show the benefits that can be gained from simply having more data in the target languages.

2 Related Work

Inspired by the success of self-supervised language model pretraining (Devlin et al., 2019, *inter-alia*), researchers have also explored this paradigm with multimodal models (Gella et al., 2017; Ákos Kádár et al., 2018). The first wave (Li et al., 2019; Tan and Bansal, 2019; Li et al., 2020; Chen et al., 2020) were initialised from BERT and pretrained on English image–text datasets like Conceptual Captions (Sharma et al., 2018) and COCO (Lin et al., 2014), where the visual modality was represented using feature vectors extracted from 10–100 automatically detected object proposals (Anderson et al., 2018). More recent models (Kim et al., 2021; Li et al., 2021; Singh et al., 2022) represent the visual modality using a Vision Transformer (Dosovitskiy et al., 2021), which can be end-to-end fine-tuned during pretraining, as opposed to working with pre-extracted object proposals.

More related to our work are the multilingual variants of these models (Liu et al., 2021; Zhou et al., 2021; Ni et al., 2021; Jain et al., 2021). The lack of large-scale multilingual multimodal datasets has resulted in different strategies to train such models. Liu et al. (2021) simply augment English caption data with text-only multilingual Wikipedia data. In addition to this, Ni et al. (2021) further create code-switched multimodal data[2] by randomly swapping English words in Conceptual Captions with the corresponding translation in one of 50 other languages, obtained through the *Panlex* dictionary. On the other hand, Zhou et al. (2021) machine translate the Conceptual Captions dataset into German, French, Czech, Japanese, and Chinese, for a total of 19.8M pretraining data points. Finally, Jain et al. (2021) pretrain on 3.6B multilingual captions by extending the Conceptual Captions collection pipeline to multiple languages.[3]

In this paper, we further explore the potential of machine translation for pretraining and fine-tuning. Zhou et al. (2021) first pretrained a model on machine translations of the Conceptual Captions pretraining data in five high-resource languages (Mandarin Chinese, Czech, French, German, and Japanese), which then resulted in overall better multilingual representations across a number of diverse languages (Bugliarello et al., 2022). Here, we explore the potential of training multimodal models on a much larger and diverse set of languages, including low-resource ones. Effectively doing so requires tackling issues and limitations with machine translation systems, which do not produce high quality translations across all languages. This is especially relevant when translating a large corpus, which might include a large number of data points with low-quality text.

3 The IGLUE Benchmark

The impetus of our work is the recent creation of the Image-Grounded Language Understanding Evaluation (IGLUE; Bugliarello et al. 2022) benchmark for evaluating multimodal models across twenty languages and four tasks, using five different datasets. Specifically, the benchmark focuses on zero- and few-shot transfer, where models are

[2] French code-switching might transform "The dog is chasing a ball" into "The *chien est* chasing a ball", for example.

[3] This large-scale dataset is not publicly available.

fine-tuned on English data and then tested to cross-lingually generalise with no or few samples in the target language for the target downstream task. The following datasets are included in IGLUE:

XVNLI is a cross-lingual Visual Natural Language Inference task (Bugliarello et al., 2022), which requires models to predict the relation (entailment, contradiction, or neutral) between a premise in the form of an image, and a hypothesis in the form of a sentence.

xGQA is a cross-lingual Grounded Question Answering task (Pfeiffer et al., 2022), using images from Visual Genome (Krishna et al., 2017) and translations of the English questions from GQA (Hudson and Manning, 2019). The questions in GQA are automatically generated from the image scene graphs.

MaRVL focuses on multicultural reasoning over images (Liu et al., 2021). The task is in the same format as the English NLVR2 (Suhr et al., 2019) data, namely to judge whether a sentence is true or false for a pair of images. However, the images and the descriptions are sourced directly in the target languages.

xFlickr&CO is an evaluation dataset for image–text retrieval in eight high-resource languages (Bugliarello et al., 2022). The images are collected from Flickr30K (Young et al., 2014) and COCO (Lin et al., 2014), while the captions are new descriptions sourced from native speakers in the target languages.

WIT is a second image–text retrieval dataset based on the Wikipedia-based Image Text dataset (Srinivasan et al., 2021). WIT is scraped directly from Wikipedia and contains a much more diverse set of image types than the other datasets, as well as more complex and entity-centric descriptions.

Each of the tasks has a natural English training counterpart: SNLI-VE (Xie et al., 2019) for XVNLI; GQA for xGQA, NLVR2 for MaRVL, and English training splits of Flickr30K and WIT.

Bugliarello et al. (2022) found that current multilingual V&L models show a large gap in performance, in each of these tasks, when evaluating on non-English data. Moreover, further training these models on a few examples in a target language only slightly improved their cross-lingual capabilities.

Approach	**ENG**	**IND**	**SWA**	**TAM**	**TUR**	**CMN**	**avg**
English	71.6	55.1	55.5	53.1	56.2	53.1	54.6
MT	67.9	59.6	61.4	60.4	64.3	59.4	61.0

Table 1: MaRVL accuracy results for zero-shot cross-lingual evaluation, i.e. English-only NLVR2 fine-tuning, and multilingual fine-tuning using machine translated NLVR2 data (MT). The average results exclude ENG accuracy.

Approach	**ENG**	**BEN**	**DEU**	**IND**	**KOR**	**POR**	**RUS**	**CMN**	**avg**
English	54.8	10.8	34.8	33.7	12.1	22.1	18.8	19.6	21.7
MT	48.1	41.8	46.5	45.7	44.8	46.8	46.2	45.7	45.3

Table 2: xGQA accuracy results for zero-shot cross-lingual evaluation, i.e. English-only GQA fine-tuning, and multilingual finetuning using machine translated GQA data (MT). Average results exclude ENG accuracy.

4 Fine-Tuning with Translated Data

As an initial experiment, we investigate the extent to which performance can be improved by fine-tuning on multilingual machine-translated data instead of only English data. We conduct this experiment on the MaRVL and xGQA datasets. The results can be seen in Tables 1 and 2, respectively.

We use the M2M-100-large model (Fan et al., 2021) to translate the NLVR2 training data into the 5 MaRVL languages, and the GQA training data into the 7 xGQA languages. For the model, we use the xUNITER (Liu et al., 2021) implementation from VOLTA (Bugliarello et al., 2021). xUNITER extends the UNITER architecture (Chen et al., 2020) multilingually, by initializing the model from XLM-RoBERTa (Conneau et al., 2020) and pre-training on English image captions and text-only multilingual Wikipedia paragraphs. Starting from the publicly-released xUNITER checkpoint, we fine-tune on the machine translated training sets for each task. For a fair comparison to English-only fine-tuning, we ensure that the multilingual fine-tuning is based on the same number of parameter updates. In effect, this reduces the number of training epochs from 20→3 for MaRVL, and 5→1 for xGQA. We round number of epochs so it is close to the English-only setup.[4] This means that, in our setup, all the images are seen for the same number of times, but each unique caption will be seen fewer times in each of the target languages.

Using machine translated data for fine-tuning

[4] Note: This is an approximation. For MaRVL, 3 epochs are equivalent to 18 (rather than 20) of English-only data (6 languages). For xGQA, 1 translated text epoch is equivalent to 8 English-only epochs (8 languages) rather than 5.

brings large improvements in performance for both MarVL and xGQA. Table 1 shows the results for MaRVL, where each non-English language increases by between 4.5–8.1 accuracy points. Table 2 shows the results for xGQA, where the performance for the non-English languages increases by 11.7–32.7 points. We also observe small decreases in performance on English for each task but this is expected. Recall that the models were fine-tuned for the same number of steps, which means the model fine-tuned on translations has been exposed to less English text in order to process multilingual text. We conclude that using machine translated fine-tuning data is an inexpensive and viable path to better task-specific performance.

5 On Pretraining with Translated Data

The previous section showed the benefits of using machine translated data for multilingual fine-tuning. We now turn our attention to whether further improvements can be realised by adding multilinguality via machine translated data is useful for pretraining. This requires two components: (i) a large-scale translation pipeline and the means to deal with potential data quality issues, and (ii) a model that can exploit the machine translated training data, which we dub TD-MML for Translated-Data Multimodal Multilingual Learning.

5.1 Translation and Data Preparation

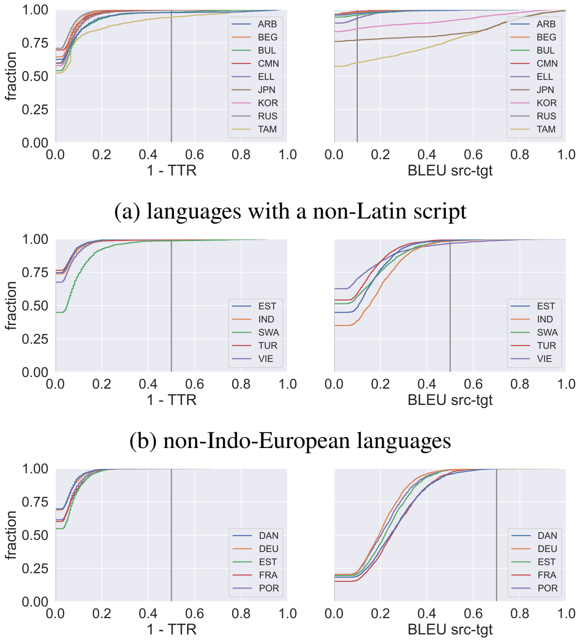
A commonly used dataset for multimodal pretraining is Conceptual Captions (Sharma et al., 2018), gathered from alt-text on the Internet and post-processed to remove proper names. We translate 2.77M English sentences from the Conceptual Captions training split into the twenty target languages in IGLUE. Once again, we use the M2M-100-large model (Fan et al., 2021), with 1.2B parameters.

We notice that the quality of the translations varies across languages, presumably due to the amount of data used to train M2M-100. Moreover, captions in this dataset often consist of sentence fragments, which may be harder to translate well.

In order to prevent bad data from corrupting the model, we apply a filtering step to the translated data. The two most frequent types of errors are single words being repeated multiple times and English words being copied into the translation. We discard sentences that exhibit these characteristics based on the following two "badness" scores:

- *Complement of the token-to-type ratio.* The token-to-type ratio (TTR) measures the fraction of unique tokens in a given text. We use its complement ($1 - \text{TTR}$), such that a large score (close to one) indicates repetition.

- *BLEU score between the source sentence and its translation.* We measure the similarilty between the English source and the (non-English) target by computing the BLEU score using the NLTK toolkit (Bird, 2006). A large score (close to one) indicates that English text has been copied into the translation.



(a) languages with a non-Latin script

(b) non-Indo-European languages

(c) Indo-European languages with Latin script

Figure 2: Cumulative distributions of the two badness scores (1 - TTR, the complement of the token-to-type ratio, and BLUE src-tgt, the BLEU score between the source and target sentence) for the nineteen non-English languages in IGLUE. The languages are grouped in three categories, and the vertical lines denote the filtering thresholds for each of the categories and two scores.

We estimate thresholds for the two scores by manually inspecting a subset of 2,000 sentences from each of the twenty target languages. We use the same TTR threshold (0.5) for all languages (since repetition is language-independent). We observe different patterns of English copying so we set different thresholds for different language groups (Figure 2): Indo-European languages with a Latin script, all languages with a non-Latin script, and non-Indo-European languages using a Latin script. We will discard all sentences with scores above *either* threshold from the multilingual pre-

✗	*funny animals of the week,* *funny animal photo, cute animal pictures*	→	*Animaux drôles, Animaux drôles,* *Animaux drôles, Animaux drôles* (FRA)
✗	*damask seamless floral pattern, ornament*	→	*Mifano ya Mifano ya Mifano ya Mifano ya* *Mifano ya Mifano ya Mifano ya Mifano ya Mifano ya Mifano* (SWA)
✗	*plaid, over garment , outfit idea cute fall outfit idea*	→	*方格, over garment, cute fall* (CMN)

Table 3: Examples of translations that are filtered out by the proposed procedure.

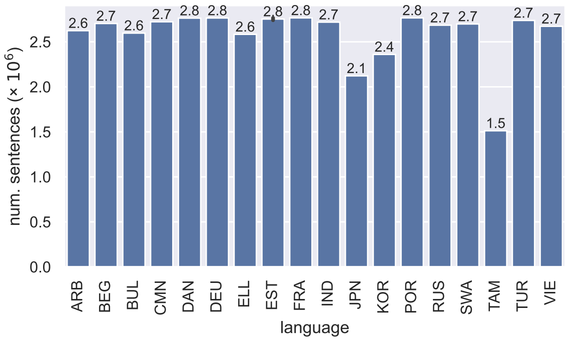


Figure 3: Number of sentences (in millions) used for pretraining in each of the nineteen IGLUE languages. The data is obtained by translating the Conceptual Captions dataset (2.77M sentences) and filtering out the poor translations. The total number of sentences in the translated dataset (including English) is 52M.

training process. Table 3 shows examples of translations that are filtered out by this procedure. The first two are rejected due to repeated words, the third because English words appear in the output.

The cumulative distribution of the scores and the corresponding thresholds are shown in Figure 2. For most languages over 95% of their translated sentences are kept, the most notable exceptions being Tamil, Japanese and Korean, for which only 54.6%, 76.6%, 85.2%, respectively, of the initial sentences are kept. Figure 3 shows the final distribution of training data across languages. The total number of sentences in the translated dataset for pretraining (including English) is 52M.

6 Model

The model we implement within our Translated Data for Multilingual Multimodal Learning framework, TD-MML, follows the single-stream cross-modal framework, such as UNITER (Chen et al., 2020) and xUNITER (Liu et al., 2021). It can be seen as a translate-train version of xUNITER, to which we add an additional pretraining task: visual translation language modelling (Zhou et al., 2021).

The TD-MML architecture consists of a series of Transformer blocks, which first concatenate the visual and language embeddings as the input, and then passes these into a multi-layer Transformer to encode the contextualized representations across image-text modalities and languages.

The input sequences are image-text pairs $(\mathbf{V}, \mathbf{X})$, where $\mathbf{V}$ are the visual features and $\mathbf{X}$ are the embedding sequence of the corresponding caption. The image features $\mathbf{v}$ in $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$ correspond to $N = 36$ object proposals extracted with Faster R-CNN (Ren et al., 2015). We extend the English pretraining text to also include the machine translated captions in m languages: $\left\{\mathbf{x}^{l_1}, \mathbf{x}^{l_2}, \ldots, \mathbf{x}^{l_m}\right\}$, The captions are processed by the same SentencePiece tokeniser regardless of their corresponding language.

6.1 Pretraining Tasks

To learn the contextualised representation across modalities and languages, we pretrain our model with three types of pretraining tasks introduced below. The goal of these pretraining tasks is to learn both the cross-modal alignment between each language and the visual modality, as well as alignments between the different languages.

Masked language and region modelling. In the V&L pretraining literature, Masked Language Modelling (MLM) and Masked Region Modelling (MRM) are two mainstream pretraining tasks, which have been demonstrated to be effective in UNITER. Given the visual features $\mathbf{V}$ and the corresponding text $\mathbf{x}^l$ in language l, MLM randomly masks tokens in the text with 15% probability and predicts the identity of the masked token using remaining textual and visual features. Analogously to MLM, MRM samples and masks image regions with 15% probability and replaces the region input with zeros. The MRM task is to classify the top-1 object label of the masked visual feature region. Please refer to Chen et al. (2020) for more details.

Image–Text Matching (ITM). This task attempts to determine whether an image–text pair is matched or not. As such, this enables the model to learn the alignment of the language and vision modalities. The matching score $s_\theta(\mathbf{x}^l, \mathbf{v})$ is computed based on the special [CLS] token which is passed through a fully-connected layer with sigmoid activation function. To train the ITM objective, we sample a positive or negative caption with equal probability for a given input image from the dataset D; the selected caption is sampled uniformly from one the twenty languages in the pretraining dataset. The loss is defined by the following equation, where we denote whether the sampled pair is a match or not by the binary label c:

$$\begin{aligned}\mathcal{L}_{\text{ITM}}(\theta) = -\,\mathbb{E}_{(\mathbf{x}^l,\mathbf{v})\sim D}\Big[&c\log s_\theta(\mathbf{x}^l,\mathbf{v})\\ &+(1-c)\log\Big(1-s_\theta(\mathbf{x}^l,\mathbf{v})\Big)\Big]\end{aligned} \tag{1}$$

Visual Translation Language Modeling. VTLM is a training objective adopted from UC^2 (Zhou et al., 2021) that combines both cross-language and cross-modal alignment learning. It takes a triple of an image $\mathbf{v}$, an English caption $\mathbf{x}^{\text{ENG}}$, and a corresponding caption $\mathbf{x}^l$ in a different language l. The task is to predict the masked caption tokens, using the multilingual textual input as well as the visual input. During pretraining, we use the same masking strategy as in MLM to randomly mask 15% of tokens from English caption and 15% of tokens from the other language caption.[5] The loss function is:

$$\begin{aligned}\mathcal{L}_{\text{VTLM}}(\theta) = -\mathbb{E}_{(\mathbf{x}^{\text{ENG}},\mathbf{x}^l,\mathbf{v})\sim D}&\\ \log P_\theta\left(\mathbf{x}_a^{\text{ENG}},\mathbf{x}_b^l\,|\,\mathbf{x}_{\backslash a}^{\text{ENG}},\mathbf{x}_{\backslash b}^l,\mathbf{v}\right)&\end{aligned} \tag{2}$$

7 Experimental setup

The implementation of TD-MML is built in the VOLTA framework (Bugliarello et al., 2021). TD-MML uses the same model configuration as the xUNITER architecture (Liu et al., 2021), which is initialised from the XLM-R cross-lingual language model (Conneau et al., 2020).

Pretraining. The dataset used for pretraining is Conceptual Captions (Sharma et al., 2018), which consists of 3.3M images with their English alt-text descriptions, of which we only have access to 2.7M instances due to linkrot. The images are represented using ResNet-101 features extracted from $N = 36$ object proposals from the Faster R-CNN (Ren et al., 2015) object detector trained on the Visual Genome dataset (Anderson et al., 2018).

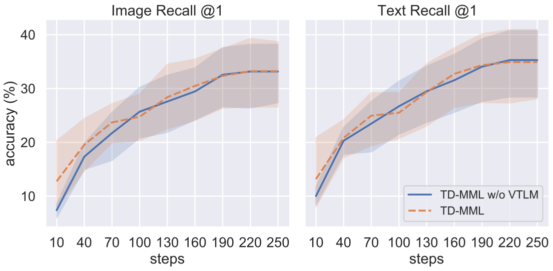


Figure 4: Average pretraining accuracy—image retrieval (left) and text retrieval (right)—as a function of the number of pretraining steps. Accuracy is calculated on 5K validation samples across five languages in the Conceptual Captions machine translated dataset.

As described in Section 5.1, we translate the captions in the 19 non-English IGLUE languages with the 1.2B-parameter M2M-100-large model (Fan et al., 2021); the translations are then filtered to eliminate likely translation errors. During training, we iterate over the images and uniformly at random sample the corresponding caption in one of the twenty languages, i.e. a batch contains samples from multiple languages. We reuse the training hyperparameters of Bugliarello et al. (2021). Specifically, xUNITER is trained over 2.77M image–English caption pairs, while TD-MML is pretrained on 52M image–multilingual caption pairs on 3×24GB TITAN RTX for 250,000 steps, which takes 5 days.

Figure 4 shows the Recall@1 curves for image retrieval and text retrieval on 5K validation samples from the Conceptual Captions dataset as a function of the number of pretraining steps. The samples are chosen from languages that represent the original MaRVL languages, i.e. ENG, IND, JPN, SWA and CMN, where the non-English data comes from the machine translation process. Model performance continues to increase in both metrics until 220,000 updates. Therefore, we use this checkpoint to fine-tune on the downstream tasks.

Fine-tuning on downstream tasks. We employ a translated data procedure at fine-tuning as well, using the same data filtering steps. Similar to the initial experiments in Section 4, we match the number of parameter updates between the experiments with English-only data and the ones with translated multilingual data. We use the same setup as in IGLUE when fine-tuning on downstream tasks.

[5] This is different from Zhou et al. (2021), who attempt to match the tokens across languages for co-masking. Caglayan et al. (2021) used the same setup as ours, where randomly mask instead of co-masking across languages.

Model	NLI	QA	Reasoning	Retrieval			
	XVNLI	xGQA	MaRVL	xFlickr&CO		WIT	
				IR	TR	IR	TR
mUNITER	53.69	9.97	53.72	8.06	8.86	9.16	10.48
xUNITER	58.48	21.72	54.59	14.04	13.51	8.72	9.81
UC^2	62.05	29.35	57.28	20.31	17.89	7.83	9.09
M^{3}P	58.25	28.17	56.00	12.91	11.90	8.12	9.98
TD-MML	64.84	**35.95**	**59.67**	**21.30**	**26.35**	**9.76**	10.35
- w/o VTLM	**66.28**	33.01	58.14	20.90	24.61	9.14	**10.61**

Table 4: Average zero-shot performance on non-English languages of multimodal models for the V&L evaluation tasks in IGLUE. Best results are marked in bold. The performance measure is accuracy for all the tasks except the cross-modal retrival tasks, which use Recall@1.

Type	Method	ENG	IND	SWA	TAM	TUR	CMN	avg
Fine-tune with English-only data (zero-shot)								
—	xUNITER	**71.55**	55.14	55.51	53.06	56.19	53.06	54.59
	TD-MML	69.00	59.04	61.01	56.44	61.95	59.88	59.67
Fine-tune with machine translated data								
Full	xUNITER	67.92	59.57	61.37	60.39	64.32	59.39	61.01
	TD-MML	67.52	59.40	**62.18**	60.55	**66.27**	59.59	**61.60**
Filtered	xUNITER	67.52	**60.82**	61.55	60.63	63.48	59.88	61.27
	TD-MML	67.09	57.62	61.91	**61.35**	64.58	**60.28**	61.15

Table 5: MaRVL accuracy results for zero-shot cross-lingual evaluation, i.e. English-only NLVR2 fine-tuning, and multilingual fine-tuning using machine translated NLVR2 data (either with the full or filtered translated data). All of the models are fine-tuned for a similar number of updates. The average results exclude ENG accuracy.

8 Results

8.1 TD Pretraining and English Fine-tuning

Here, we evaluate the zero-shot language understanding abilities of the TD-MML model that has been pretrained with multiple languages, but fine-tuned on English task-specific data only (e.g. NLVR2 for MaRVL, GQA for xGQA, etc.). The averaged zero-shot results across languages are shown in Table 4. The full zero-shot per-language results on each task are detailed in Appendix C.

We see a substantial improvement for TD-MML across all tasks compared to xUNITER and the state-of-the-art UC^2. The improvement between our TD-MML and the best baseline models for each task reaches 4.23 points for XVNLI, 6.6 points for xGQA, 2.39 points for MaRVL, 0.99 (IR) and 8.46 points (TR) for xFlickr&CO, and 0.6 (IR) and 0.13 points (TR) for WIT. The results from the other tasks show the clear benefit of pretraining on multilingual multimodal data on a diverse array of multimodal tasks across many languages.

Gains from VTLM. Table 4 also shows the effectiveness of pretraining with the additional visual translation language modeling (VTLM) objective: adding VTLM boosts the performance for five out of the seven tasks. Improving cross-lingual alignment during pretraining thus manifests itself in better multi-lingual understanding ability.

8.2 MT Fine-tuning TD-MML

We now ask whether combining the machine translated pretraining strategy of TD-MML with additional machine translated fine-tuning can provide further gains, compared to both English-only fine-tuning (zero-shot) and the MT-fine-tuning strategy applied to xUNITER in Section 4. The results for MaRVL and xGQA are shown in Table 5 and Table 6 respectively. Perhaps surprisingly, the performance for xUNITER and TD-MML are very

Type	Method	ENG	BEN	DEU	IND	KOR	POR	RUS	CMN	avg
Fine-tune with English-only data (zero-shot)										
—	xUNITER	**54.83**	10.80	34.83	33.73	12.12	22.13	18.84	19.55	21.70
	TD-MML	53.60	23.62	42.29	38.23	32.44	41.36	38.08	35.61	35.95
Fine-tune with machine translated data										
Full	xUNITER	48.08	41.76	**46.47**	**45.68**	**44.76**	**46.77**	**46.19**	**45.66**	**45.33**
	TD-MML	47.38	42.27	46.31	44.24	44.53	46.24	45.78	44.68	44.86
Filtered	xUNITER	48.40	42.10	46.13	45.30	44.70	46.33	45.95	45.50	45.14
	TD-MML	48.04	**42.86**	46.45	44.78	44.60	46.67	46.02	45.07	45.21

Table 6: xGQA accuracy results for English-only fine-tuning (zero-shot evaluation) and multilingual fine-tuning using machine translated GQA data (either with the full or filtered translated data). All of the models are fine-tuned for a similar number of updates. The average results exclude ENG accuracy.

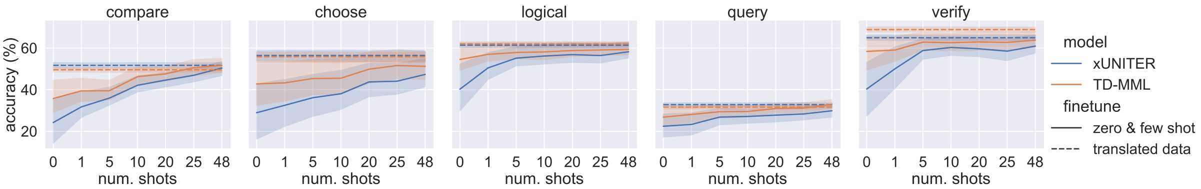


Figure 5: xGQA average accuracy (with confidence intervals) across the languages on the five question types. xUNITER and TD-MML are evaluated in the zero-shot, few-shot and machine translation fine-tuning settings. The error bars represent 95% confidence intervals and are obtained by bootstraping (1000 repeats).

similar after multilingual MT fine-tuning, with TD-MML slightly outperforming xUNITER on the MaRVL dataset and with xUNITER generally yielding better performance on xGQA. That is, in this setting, multilingual multimodal pretraining in TD-MML does not convey any added benefit. A potential explanation is that fine-tuning on enough multilingual data (as it is the case of machine translated data) compensates for the lack of multilingual multimodal pre-training. The performance of TD-MML on the English splits after English-only fine-tuning supports this hypothesis.

The results in Tables 5 and 6 further indicate that the filtering strategies offer mixed results when applied to fine-tuning. We believe that this outcome happens because the corresponding datasets, GQA for xGQA and NLVR2 for MaRVL, are typically much cleaner datasets than the Conceptual Captions dataset, resulting in less filtering and, consequently, less representative results.

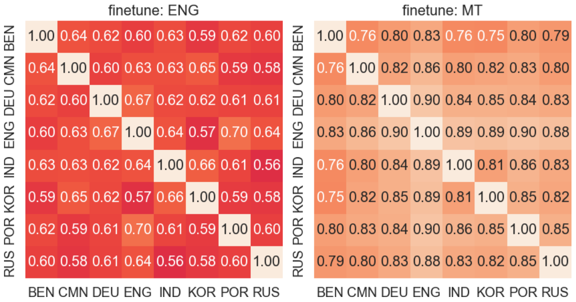
8.3 Few-Shot vs Machine Translated Data

We also ask whether it is better for downstream performance to train on a limited number of clean language-specific and task-specific samples (few-shot learning) or to simply machine translate the task-specific English data, which is likely to result in noisier data. So, in addition to the previously introduced fine-tuning setups—fine-tuning on task-specific English data for a zero-shot evaluation setting, and fine-tune on machine translated task-specific English data—we also consider the few-shot learning setup, in which we continue fine-tuning of the zero-shot (English fine-tuned) model, using a few human-authored language-specific and downstream-specific samples from the IGLUE benchmark (i.e., 1, 5, 10, 20, 25, 48 shots).

Few shot results for xGQA are presented in Figure 5, broken down by question type. We observe that, as expected, the performance generally improves with the quantity of training data (number of shots). More interestingly, the machine translated fine-tuning upper bounds the performance of the few-shot approach for both of our multimodal multilingual models. Across the five question categories, the variation in performance can be largely explained by the cardinality of the set of plausible answers: it is easier to answer correctly a verification or logical question, whose answer is usually either "yes" or "no", than a querying question, whose answer usually involves a broader set of words.

	Q1		Q2		Q3		Q4	
Q	What is the sign behind the young person?		Is the black and white cat unhappy or happy?		What is covering the man that is wearing jeans?		How is this cooking utensil called?	
A	traffic sign		unhappy		jacket		baking pan	
	fine-tune: ENG	fine-tune: MT	fine-tune: ENG	fine-tune: MT	fine-tune: ENG	fine-tune: MT	fine-tune: ENG	fine-tune: MT
BEN	car	pole	no	lush	bag	umbrella	paper	pretzel
CMN	pole	stop sign	white	happy	coat	umbrella	book	stove
DEU	fire hydrant	stop sign	unhappy	happy	jacket	umbrella	mirror	tea kettle
ENG	stop sign	stop sign	unhappy	happy	towel	umbrella	pan	tea kettle
IND	street sign	stop sign	unhappy	happy	blanket	umbrella	yes	yes
KOR	pole	stop sign	gray	happy	backpack	umbrella	drum	drum
POR	car	car	happy	happy	suitcase	umbrella	table	tea kettle
RUS	stop sign	stop sign	happy	happy	umbrella	umbrella	shelf	pan

Figure 6: Qualitative results on the xGQA dataset. Given an image and a question (denoted by Q), we show the corresponding groundtruth answer (denoted by A), together with the predictions in each of the eight languages for the model finetuned on either English-only sentences (left column of answers) or on machine translated sentences (right column of answers).



finetune: ENG

	BEN	CMN	DEU	ENG	IND	KOR	POR	RUS
BEN	1.00	0.64	0.62	0.60	0.63	0.59	0.62	0.60
CMN	0.64	1.00	0.60	0.63	0.63	0.65	0.59	0.58
DEU	0.62	0.60	1.00	0.67	0.62	0.62	0.61	0.61
ENG	0.60	0.63	0.67	1.00	0.64	0.57	0.70	0.64
IND	0.63	0.63	0.62	0.64	1.00	0.66	0.61	0.56
KOR	0.59	0.65	0.62	0.57	0.66	1.00	0.59	0.58
POR	0.62	0.59	0.61	0.70	0.61	0.59	1.00	0.60
RUS	0.60	0.58	0.61	0.64	0.56	0.58	0.60	1.00

finetune: MT

	BEN	CMN	DEU	ENG	IND	KOR	POR	RUS
BEN	1.00	0.76	0.80	0.83	0.76	0.75	0.80	0.79
CMN	0.76	1.00	0.82	0.86	0.80	0.82	0.83	0.80
DEU	0.80	0.82	1.00	0.90	0.84	0.85	0.84	0.83
ENG	0.83	0.86	0.90	1.00	0.89	0.89	0.90	0.88
IND	0.76	0.80	0.84	0.89	1.00	0.81	0.86	0.83
KOR	0.75	0.82	0.85	0.89	0.81	1.00	0.85	0.82
POR	0.80	0.83	0.84	0.90	0.86	0.85	1.00	0.85
RUS	0.79	0.80	0.83	0.88	0.83	0.82	0.85	1.00

Figure 7: Cross-language prediction correlations on the test split of xGQA for two of the proposed models: TD-MML fine-tuned on English-only data (left) and TD-MML fine-tuned on translated data (right).

8.4 Cross-Language Correlation Analysis

Are the same questions easy (answered correctly) or difficult (answered incorrectly) across languages? We use Cohen's kappa coefficient κ to measure agreement between languages on the xGQA dataset (see Appendix B for details). The results are presented in Figure 7. We show results for two variants of our pretrained TD-MML model: either fine-tuned on English-only data (ENG) or on machine translated data (MT).

We see that the MT fine-tuned results show much higher agreement across languages, compared to English-only fine-tuning. On the one hand, this could be considered counter intuitive: in the MT fine-tuned setting, there is more language-specific data. However, the MT fine-tuned results have higher accuracy overall, suggesting that high agreement across languages is due to the model confronting inherent item difficulty (as judged by all languages), rather than language-specific issues.

Examples of increased cross-language agreement in the MT fine-tuned model (MT) are presented in Figure 6. Across the eight languages, we find the predictions of the MT fine-tuned model are more consistent than those of the ENG-finetuned model (Q1, Q2, Q3). However, for more ambiguous and difficult samples, the model fine-tuned with translated data still gives varied, but arguably more plausible, predictions across languages (Q4).

9 Conclusion

In this paper, we investigate the role of machine translated (MT) data in multilingual multimodal learning in a controlled setup. We consider two applications of MT data, namely for augmenting the pretraining and the fine-tuning data. We find that both convey a clear immediate benefit on downstream performance for nearly all tasks in the IGLUE benchmark; however, we do not find an additive benefit of combining it for both pretraining and fine-tuning together. When using machine translated text, filtering out bad translations in a quick and reliable way is crucial, and we develop a simple and effective strategy for doing this. Our results shed light in the importance of explicitly grounding multilingual text in the visual modality in both pretraining and fine-tuning stages.

Limitations

Our paper investigates the benefits and limitations of machine translated data towards multilingual multimodal learning. In doing so, we solely rely on the M2M-100 model (Fan et al., 2021). This is a large, multi-to-multi translation system, which proved to be easy to use. Our analyses and results are based on the performance of this model. It would be instructive to investigate how the expected performance of translation systems[6] affects (i) the proportion of sentences with high 'badness' scores, and (ii) the resulting performance of the multilingual multimodal systems. Moreover, while machine translating a large corpus is a cheaper effort than manually translating the data or scraping it from the web, there is still a one-time effort required to translate the data before using it for training new models. Therefore, we release our multilingual pretraining and fine-tuning datasets.

From an experimental angle, although the proposed framework can be applied to any existing architecture, we only evaluate a single model due to computational constraints.

We would also like to stress the importance of using target-language originating evaluation data in multimodal setups, rather than translated data. Fitting to translationese is a risk when using translation data at training time, and can only be identified if the evaluation data does not also contain translations, especially automatically generated ones.

Finally, a core limitation of the overall translate data framework is that it centers English as the source language. For example, this means only concepts mentioned in English captions can be grounded across languages (Liu et al., 2021), and hence some non-English concepts might never be modelled. However, we show that machine translating data provides a strong starting point that can effortlessly be integrated in a pipeline, upon which language-specific annotations can be added.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. This work was supported in part by a grant of the Romanian Ministry of Education and Research No PN-III-P1-1.1-PD-2019-0918 (Dan Oneață). Emanuele Bugliarello was supported by the funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199.

References

Aishwarya Agrawal, Ivana Kajić, Emanuele Bugliarello, Elnaz Davoodi, Anita Gergely, Phil Blunsom, and Aida Nematzadeh. 2022. Rethinking evaluation practices in visual question answering: A case study on out-of-distribution generalization. *arXiv preprint arXiv:2205.12191*.

Ákos Kádár, Desmond Elliott, Marc-Alexandre Côté, Grzegorz Chrupala, and Afra Alishahi. 2018. Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 402–412, Brussels, Belgium. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, Salt Lake City, USA. Computer Vision Foundation / IEEE Computer Society.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 2425–2433, Santiago, Chile. IEEE Computer Society.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565. The Association for Computer Linguistics.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

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

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pretraining for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

[6] Despite our best efforts, we were unable to find the per-language translation performance for the M2M-100 model.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Proceedings of the Computer Vision - ECCV 2020 - 16th European Conference*, pages 104–120.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

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. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845, Copenhagen, Denmark. Association for Computational Linguistics.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617, Salt Lake City, USA. Computer Vision Foundation / IEEE Computer Society.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, Long Beach, CA, USA. Computer Vision Foundation / IEEE.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5583–5594, Virtual Event. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

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. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems*, pages 9694–9705, virtual.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of Computer Vision - ECCV 2020 - 16th European Conference*, volume 12375, pages 121–137, Glasgow, UK. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, Las Vegas, NV, USA. IEEE Computer Society.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3977–3986.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 91–99. Curran Associates, Inc.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650.

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*, pages 2443–2449, Virtual Event, Canada. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv:1901.06706*.

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. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4155–4165, virtual. Computer Vision Foundation / IEEE.

A Language information

The machine translated fine-tuning data and pretraining data cover 20 languages, spanning 11 language families and 9 scripts. The scripts are Arabic, Bengali-Assamese, Chinese, Cyrillic, Greek, Huangual, Kanji, Latin, and Tamil. Table 7 summarizes this information together with a listing of the language code abbreviations and details regarding the use of the languages in each of the five tasks from the IGLUE benchmark.

B Kappa Setting

Cohen's kappa corrects for random agreement, but has an upper bound based on the difference in marginal probabilities (here, equivalent to accuracy): if one system assigns more correct answers than the other, then the maximum achievable κ is less than one. Since we compare across languages with different accuracies, we normalise the usual coefficient by the maximum achievable value, resulting in κ-ratio, the proportion of agreement given system accuracies (label rates).

C Performance per Target Language

Tables 8 to 12 provide language-specific performance for zero-shot evaluation (i.e., English-only fine-tuning) on the five IGLUE tasks (MaRVL, XVNLI, xGQA, xFlickrCO, WIT) for two variants of our TD-MML approach (with and without VTML loss) against four state-of-the-art approaches (mUNITER, xUNITER, UC^2, M^{3}P).

The experimental results show that our TD-MML usually achieves better performance than the competing models on the non-English languages. The closest competitor is UC^2, which is also a method that is pretrained on machine translated data, but only in five languages (CMN, DEU, JPN, FRA, CZE). This partially explains UC^2 strong performance in some of these instances: for example, on FRA in XVNLI (Table 9) or on DEU in xGQA (Table 10).

On the WIT retrieval tasks, we notice that even if TD-MML still performs better or on par with previous approaches, the results are poor across all languages and methods. A possible explanation is that the distribution of images and captions on Wikipedia is distinct from other datasets.

Among the two variants of TD-MML, we generally observe a benefit of incorporating the VTML loss, the largest gains manifesting for BEN, CMN, KOR, RUS languages on the xGQA task.

D Analysis over Question Types in xGQA

Figure 8 shows the accuracy results for three evaluation setups (zero-shot, few shot learning and machine translation fine-tuning) on xGQA over the five different question types. The language-specific sub-figures show the gap between xUNITER and TD-MML. The largest differences between of them are the KOR, POR, RUS languages and for the *compare*, *logical*, and *verify* question types.

Language				**NLI**	**QA**	**Reasoning**	**Retrieval**	
Name	Code	Family	Script	XVNLI	xGQA	MaRVL	xFlickr&CO	WIT
English	ENG	Indo-E	Latin	✓✓	✓✓	✓✓	✓✓	✓✓
Arabic	ARB	Afro-A	Arabic	✓✓				✓
Bengali	BEN	Indo-E	Bengali		✓✓			
Bulgarian	BUL	Indo-E	Cyrillic					✓
Danish	DAN	Indo-E	Latin					✓
Estonian	EST	Uralic	Latin					✓
German	DEU	Indo-E	Latin		✓✓		✓✓	
Greek	ELL	Indo-E	Greek					✓
French	FRA	Indo-E	Latin	✓✓				
Indonesian	IND	Austron	Latin		✓✓	✓✓	✓✓	✓
Japanese	JPN	Japonic	Kanji				○ ✓✓	✓
Korean	KOR	Koreanic	Hangul		✓✓			✓
Mandarin	CMN	Sino-T	Hanzi		✓✓	✓✓	✓✓	
Portuguese	POR	Indo-E	Latin		✓✓			
Russian	RUS	Indo-E	Cyrillic	✓✓	✓✓		✓✓	
Spanish	SPA	Indo-E	Latin	✓✓			✓✓	
Swahili	SWA	Niger-C	Latin			✓		
Tamil	TAM	Dravidian	Tamil			✓		
Turkish	TUR	Turkic	Latin			✓✓	✓✓	✓
Vietnamese	VIE	Austro-A	Latin					✓

Table 7: IGLUE details: Table replicated from Table 2 in Bugliarello et al. (2022). Tasks legend: ✓✓few-shot train and test splits; ✓test-only; ○ Japanese captions in xFilickr&CO are manual translations of the English ones.

Method	**ENG**	**IND**	**SWA**	**TAM**	**TUR**	**CMN**	**avg**
mUNITER	**71.91**	54.79	51.17	52.66	54.66	55.34	56.76
xUNITER	71.55	55.14	55.51	53.06	56.19	53.06	57.42
UC^2	70.56	56.74	52.62	**60.47**	56.70	**59.88**	59.50
M^{3}P	68.22	56.47	55.69	56.04	56.78	55.04	58.04
TD-MML w/o VTLM	68.45	58.25	59.30	56.28	61.02	55.83	59.85
TD-MML	69.00	**59.04**	**61.01**	56.44	**61.95**	**59.88**	**61.22**

Table 8: Accuracy on the MaRVL task for zero-shot evaluation (i.e. English-only fine-tuning).

Method	**ENG**	**ARB**	**SPA**	**FRA**	**RUS**	**avg**
mUNITER	76.38	46.73	56.96	59.36	51.72	58.23
xUNITER	75.77	51.98	58.94	63.32	59.71	61.94
UC^2	76.38	56.19	57.47	**69.67**	64.86	64.91
M^{3}P	**76.89**	55.24	58.85	56.36	62.54	61.98
TD-MML w/o VTLM	76.12	**62.20**	**66.75**	68.39	**67.78**	**68.25**
TD-MML	75.52	61.25	65.89	67.44	64.78	66.98

Table 9: Accuracy on the XVNLI task for zero-shot evaluation (i.e. English-only fine-tuning).

Method	**ENG**	**BEN**	**DEU**	**IND**	**KOR**	**POR**	**RUS**	**CMN**	**avg**
mUNITER	54.68	3.06	23.95	9.36	4.21	13.67	8.49	7.30	16.77
xUNITER	54.83	10.80	34.83	33.73	12.12	22.13	18.84	19.55	26.75
UC^2	**55.19**	19.99	**42.85**	28.67	21.36	30.42	31.00	31.16	32.78
M^{3}P	53.75	18.64	33.42	32.48	25.11	31.40	27.50	28.65	31.76
TD-MML w/o VTLM	54.37	16.20	39.98	36.28	29.96	**41.79**	37.00	29.88	35.68
TD-MML	53.60	**23.62**	42.29	**38.23**	**32.44**	41.36	**38.08**	**35.61**	**38.15**

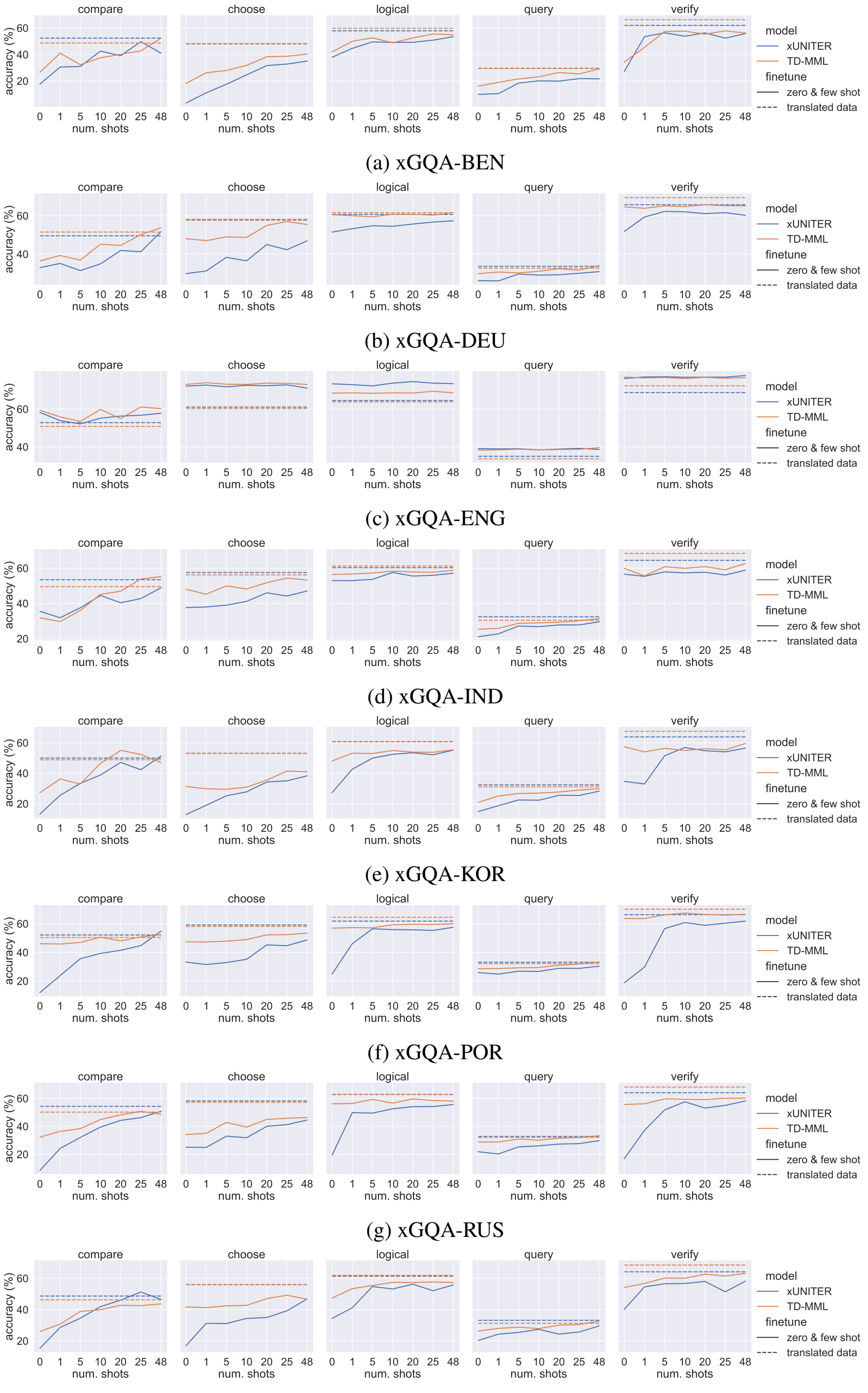
Table 10: Accuracy on the xGQA task for zero-shot evaluation (i.e. English-only fine-tuning).

Type	Method	ENG	DEU	SPA	IDN	JPN	RUS	TUR	CMN	avg
IR	mUNITER	**44.50**	12.05	13.15	5.95	6.30	5.85	1.75	11.35	12.61
	xUNITER	38.45	14.55	16.10	16.50	10.25	15.90	9.05	15.95	17.09
	UC^2	37.40	**28.60**	15.95	14.60	**24.25**	20.00	7.15	**31.60**	**22.44**
	M^{3}P	31.35	13.35	13.40	13.20	10.30	15.95	7.15	16.45	15.14
	TD-MML w/o VTLM	29.45	19.70	**23.20**	21.05	16.65	**24.15**	19.65	21.90	21.97
	TD-MML	28.15	19.95	22.20	**22.35**	18.35	23.30	**19.75**	23.20	22.16
TR	mUNITER	**40.90**	11.85	13.05	7.55	7.70	6.80	3.25	11.85	12.87
	xUNITER	32.05	13.25	15.10	16.75	9.85	14.80	10.05	14.80	15.83
	UC^2	34.55	23.90	15.30	13.60	22.40	16.75	6.95	26.30	19.97
	M^{3}P	24.60	11.85	12.15	12.10	9.65	14.45	8.35	14.75	13.49
	TD-MML w/o VTLM	34.35	22.30	27.80	24.35	21.00	27.60	22.30	26.90	25.83
	TD-MML	33.70	**24.70**	**29.40**	**25.55**	**22.90**	**29.95**	**23.65**	**28.30**	**27.27**

Table 11: Experimental results on the xFlickr&CO task (image retrieval, IR, top, and text retrieval, TR, bottom) for zero-shot evaluation (i.e. English-only fine-tuning).

Type	Method	ARB	BUL	DAN	ELL	ENG	EST	IND	JPN	KOR	TUR	VIE	avg
IR	mUNITER	7.74	8.26	10.66	8.95	**19.90**	7.67	10.88	**9.00**	5.91	9.57	**13.00**	10.14
	xUNITER	7.63	8.49	10.32	11.23	16.70	6.41	10.21	7.30	**6.34**	9.57	9.72	9.45
	UC^2	6.62	8.84	9.43	8.77	17.90	4.69	9.88	9.80	4.30	7.49	8.46	8.74
	M^{3}P	8.87	8.84	9.43	9.65	15.50	5.38	8.66	7.00	6.12	6.52	10.78	8.95
	TD-MML w/o VTLM	**9.20**	**9.19**	9.43	11.05	14.70	7.90	10.43	8.10	5.69	9.29	11.10	9.64
	TD-MML	8.64	8.02	**11.00**	**11.58**	16.10	**8.12**	**11.77**	**9.00**	6.02	**11.79**	11.63	**10.33**
TR	mUNITER	9.21	10.17	12.16	10.54	**22.34**	8.33	**12.88**	8.79	6.75	10.87	**15.07**	**11.56**
	xUNITER	9.08	10.30	9.34	12.38	18.54	7.82	10.66	10.10	6.97	9.69	11.74	10.60
	UC^2	8.32	7.69	10.44	11.64	19.71	6.03	11.47	**10.81**	5.74	8.81	9.90	10.05
	M^{3}P	8.32	9.80	11.79	12.02	15.33	8.21	10.89	8.43	**7.09**	10.57	12.66	10.46
	TD-MML w/o VTLM	**10.47**	**11.04**	10.69	**13.86**	18.54	**9.49**	11.59	8.91	6.79	11.01	13.02	11.33
	TD-MML	9.21	10.17	**13.39**	12.20	17.81	7.56	10.77	9.50	6.41	**11.16**	13.58	11.02

Table 12: Experimental results on the WIT tasks (image retrieval, IR, top, and text retrieval, TR, bottom) for zero-shot evaluation (i.e. English-only fine-tuning).



(a) xGQA-BEN

(b) xGQA-DEU

(c) xGQA-ENG

(d) xGQA-IND

(e) xGQA-KOR

(f) xGQA-POR

(g) xGQA-RUS

(h) xGQA-CMN

Figure 8: xGQA accuracy on zero-shot, few shot learning and machine translation fine-tuning evaluation for each of the five questions types (compare, choose, logical, query, verify) and eight languages (Bengali, German, English, Indonesian, Korean, Portuguese, Russian, Mandarin).