Cross-Lingual Transfer of Large Language Model by Visually-Derived Supervision Toward Low-Resource Languages

Masayasu Muraoka
IBM Research
mmuraoka@jp.ibm.com

Bishwaranjan Bhattacharjee
IBM Research
bhatta@us.ibm.com

Michele Merler
IBM Research
mimerler@us.ibm.com

Graeme Blackwood
IBM Research
blackwood@us.ibm.com

Yulong Li
IBM Research
yulongl@us.ibm.com

Yang Zhao
IBM Research
yangzhao@ibm.com

ABSTRACT

Recent progress on vision and language research has shown that visual supervision improves the performance of large language models (LLMs) in various natural language processing (NLP) tasks. In particular, the Vokenization approach [65] initiated a new way of incorporating visual information into LLM training, demonstrating the potential of visual supervision for NLP tasks in a monolingual (i.e., English) setting. Given the effectiveness of visual information in human communication among people who speak different languages, we tackle an ambitious question in this paper; can we expect that visual supervision contributes to cross-lingual transfer learning from a high-resource language to low-resource languages in NLP tasks? To study this hypothesis, we build a cross-lingual Vokenization model and train a cross-lingual LLM on three languages, English, Urdu, and Swahili, in which the last two are considered low-resource languages. The experimental results demonstrate that our visually-supervised cross-lingual transfer learning method significantly improves the LLM performance in multiple cross-lingual NLP tasks such as XNLI, NER, and TyDiQA tasks for low-resource languages. We also qualitatively and quantitatively demonstrate that the benefit of our approach increases as the linguistic distance between low- and high-resource languages grows larger.

CCS CONCEPTS

• **Information systems** → **Language models**; **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Transfer learning**; Natural language processing.

KEYWORDS

Visual supervision; Large language model training; Cross-lingual transfer; Low-resource languages

ACM Reference Format:
Masayasu Muraoka, Bishwaranjan Bhattacharjee, Michele Merler, Graeme Blackwood, Yulong Li, and Yang Zhao. 2023. Cross-Lingual Transfer of Large Language Model by Visually-Derived Supervision Toward Low-Resource Languages. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611992

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3611992

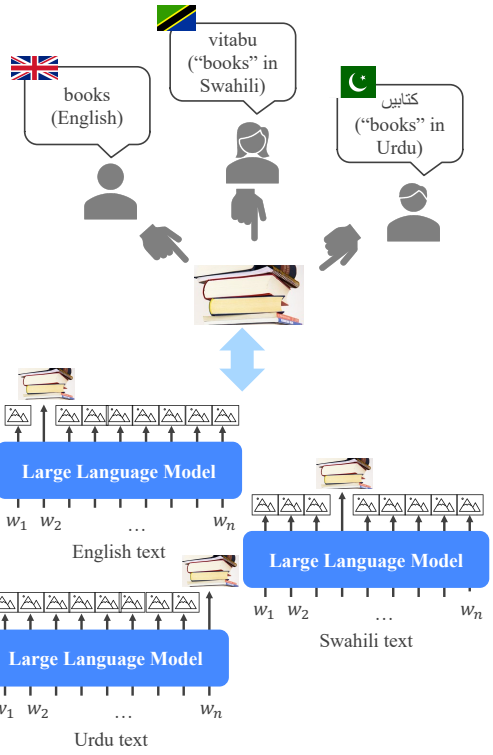

Figure 1: Visual information encourages cross-lingual natural language understanding.

1 INTRODUCTION

Language acquisition by human beings involves language itself as well as visual and other perceptions [1, 4, 15]. Prior works have proven that visual information can enhance various natural language processing (NLP) technologies [8, 21, 36, 37, 61, 79]. Recently, a number of works have been focusing more on improving the performance of large language models (LLMs) and vision and language (V+L) models with visual information [25, 41, 49, 65, 66, 70, 76, 80].

In particular, the Vokenization approach [65] enables the LLMs to learn better language representations from textual as well as visual

supervision during LLM training, leading to improved performance in NLP tasks in a monolingual (i.e., English) setting.

In addition to language acquisition, visual perception also helps human communication among people who have different native languages through, for example, gestures, visual pointing, and body language [22, 55, 62] as shown at the top of Figure 1. Lots of works have shown that visual supervision could improve the performance of multi-modal models in cross-lingual multi-modal tasks such as image captioning, image retrieval, machine translation, etc. [6, 7, 19, 29, 40, 59, 64, 81]. However, attempts to improve the performance of LLMs in *cross-lingual NLP tasks* using visual information have been less thoroughly explored, even though if successful, they could provide new ways of using visual information to further augment the cross-lingual transfer learning of LLMs.

To this end, we study whether visually-derived supervision can contribute to the cross-lingual transfer learning of an LLM, especially in low-resource languages. We extend the Vokenization technique [65] by building a cross-lingual LLM on three languages, namely English, Urdu, and Swahili, where Urdu and Swahili are considered low-resource languages. In addition to the standard masked language modeling, the cross-lingual LLM is jointly trained to predict the visually-derived supervision and some of these supervisions are shared across languages as intuitively shown at the bottom of Figure 1, from which we expect to encourage the cross-lingual transfer learning.

Our experimental results demonstrate that a visually-supervised cross-lingual LLM significantly improves the performance in multiple cross-lingual NLP tasks such as cross-lingual natural language inference (XNLI) [13], named entity recognition (NER) [54], and Typologically Diverse QA (TyDiQA) on not only the low-resource but also high-resource languages. We also qualitatively and quantitatively demonstrate that the benefit of our approach increases as the linguistic distance between low- and high-resource languages grows larger. Specifically, we show that the performance gain over a vanilla LLM on Urdu is larger than that on Swahili even though Urdu is less similar linguistically to English than Swahili.

Our contribution is threefold:

- Study whether visual supervision contributes to cross-lingual transfer learning of an LLM by extending the Vokenization technique to a cross-lingual setting to build a visually-supervised cross-lingual LLM,
- Experimentally verify the above hypothesis, demonstrating significant performance improvements over a vanilla LLM (up to 17.78 points) in multiple cross-lingual NLP tasks for low-resource languages,
- Qualitatively and quantitatively reveal that the performance gain by visual supervision can be larger even if a low-resource language is linguistically less similar to a high-resource one.

2 RELATED WORK

2.1 Visually-guided NLP methods and tasks

Before the emergence of LLMs, most works focused on incorporating visual information into specific NLP methods and tasks to improve language understanding, such as semantic parsing [8, 36, 37, 61], natural language inference [74], information retrieval [21, 23, 24], machine translation [18, 30, 73, 79], evaluation for natural language generation [82], spatial commonsense reasoning [45], bilingual lexicon learning [33, 68], and so on. Since then, a number of language representation learning methods have been proposed, which are applicable to multiple NLP tasks [5, 10, 32, 39, 52, 78]. Recent works have proven that visual information can enhance LLMs [25, 49, 65, 66, 70, 76], V+L models [41, 80], and both [42].

In particular, Vokenization [65] improved LLMs by using token-level image retrieval while iACE [49] did so by sentence-level image generation with a text-to-image model. Z-LaVI [76] combined sentence-level image retrieval and image generation. VaLM [70] appended image features to the attention layer in LLMs while VAWI [25] inserted image features to the embedding layer in LLMs.

We choose to extend the Vokenization work [65] from the viewpoint of the computational efficiency and invariance of the model architecture. The other works described above either use sentence-level image generation/retrieval to obtain visual supervisions or modify the model architecture of LLMs to insert visual information. The former would become computationally more expensive and possibly infeasible to generate/retrieve desirable images for whole input texts as the input texts become semantically more complex. As for the latter, we want to keep the broader applicability of LLMs to NLP tasks without modifying the model architecture. The Vokenization work [65] keeps the LLM architecture and works as a token-level image retriever, which is computationally efficient.

2.2 Cross-lingual transfer of multi-modal models with visual data

Instead of LLMs, utilizing visual data also improves performance of multi-modal models across languages in multi-modal tasks such as image captioning, visual question answering, multi-modal machine translation, text-image and speech-image retrieval etc. [6, 7, 19, 29, 40, 59, 64, 81]. Specifically, similar to our work, [59] investigated whether visual information contributes to cross-lingual transfer learning of a grounded speech model in the low-resource setting.

While most of these works rely on paired image-text or image-speech datasets where an image and multilingual texts/speeches are all aligned, we use a multilingual multi-modal dataset whose texts are not explicitly aligned across languages.

2.3 Cross-lingual transfer of LLMs with texts

LLMs trained on a large corpus of multilingual text data have been shown to perform better in cross-lingual NLP tasks [11, 12, 16, 17]. These LLMs use a single Transformer-based architecture [67] that has a unified vocabulary over multiple languages and can process texts in arbitrary languages without language-specific computations. There also exist probing works that attempt to investigate why cross-lingual LLMs work so well despite such a simple training that trains a single LLM on a multilingual corpus without any explicit alignment of texts across languages [14, 57].

In this paper, we propose to improve the performance of a cross-lingual LLM with *visual data* in cross-lingual NLP tasks.

3 METHOD

We describe the training procedure of the visually-supervised language model (VLM) [65] (§3.1). The visually-derived supervision

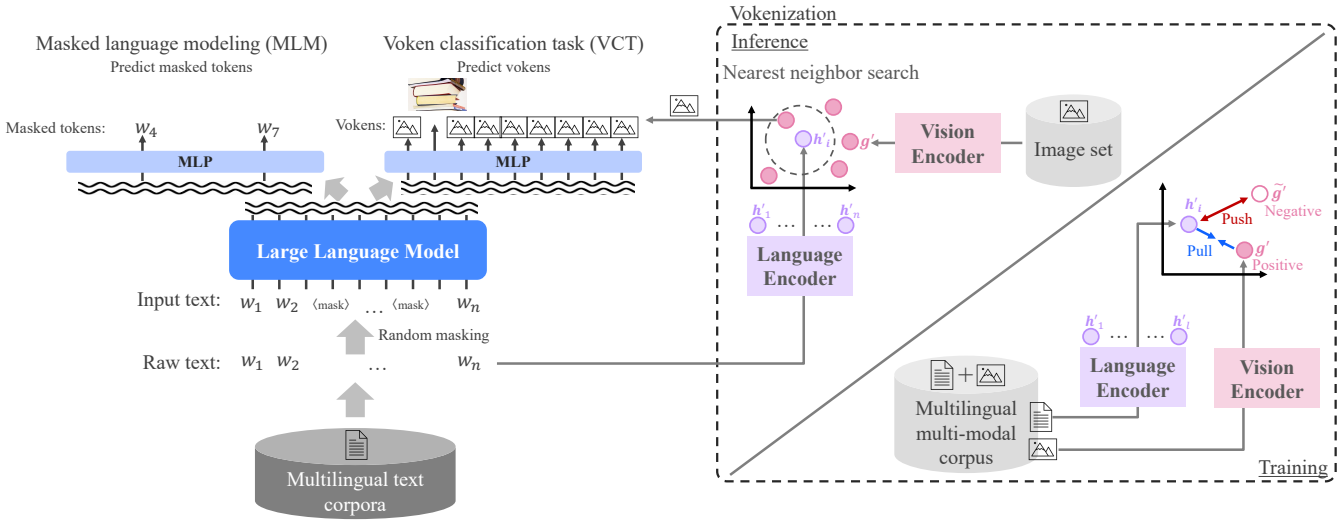

Figure 2: Method overview. Left: Training of visually-supervised language model (VLM). Right: Training and inference of Vokenization.

used in the VLM training is generated by Vokenization (§3.2). Figure 2 illustrates an overview of the method.

3.1 Visually-supervised language model (VLM)

A VLM is an LLM jointly trained with two pre-training tasks: masked language modeling (MLM) and voken classification task (VCT). Since the VLM is an LLM, the VLM is trained on a text corpus and applicable to NLP tasks taking a text as an input while it is not applicable to multi-modal tasks that requires to process images. In this paper, we choose a masked language model as our VLM architecture following previous work [65], but replace a monolingual model [16] with a cross-lingual model, specifically XLM-R [11]. Since the two pre-training tasks described below are independent of each other, we can apply the same training method to other model architectures (e.g., causal language models [58]) wherever possible by replacing the MLM with the causal language modeling (CLM) task, but we leave the investigation for future work.

3.1.1 Masked language modeling (MLM). In MLM, the VLM predicts masked tokens given a masked sentence. By utilizing the context of the masked sentence, the model learns a good representation of tokens that can be used in various NLP tasks [16].

Formally, a VLM $\mathcal{M}$ predicts a conditional probability of masked tokens $\boldsymbol{w}_{\text{mask}}$ given a masked sentence $\boldsymbol{s}_{\setminus \boldsymbol{w}_{\text{mask}}}$, in which the original sentence $\boldsymbol{s}$ from a text corpus $\mathcal{T}_{\text{txt}}$ consists of n tokens, $\boldsymbol{s} = w_1 w_2 \dots w_n = \{w_i\}_{i=1}^n$, and m% of the tokens in $\boldsymbol{s}$ are randomly masked, i.e., replaced with a special token <mask>.[1] For each masked token $w_i \in \boldsymbol{w}_{\text{mask}}$, we compute a probability distribution $P(w_i \mid \boldsymbol{s}_{\setminus \boldsymbol{w}_{\text{mask}}}; \mathcal{M})$ over the vocabulary $\mathcal{V}_{\text{token}}$ from the last hidden states $\boldsymbol{h}_i \in \mathbb{R}^d$ of tokens in the VLM.

$$P(w_i|\boldsymbol{s}_{\setminus \boldsymbol{w}_{\text{mask}}}; \mathcal{M}) = f_{\text{token}}(\boldsymbol{h}_i), \tag{1}$$

in which $f_{\text{token}}(\cdot)$ is a multi-layer perceptron (MLP) that projects the hidden states onto a continuous space with dimension equal to the vocabulary size $|\mathcal{V}_{\text{token}}|$ using an activation function. Based on the output, we compute the negative log likelihood as an objective function for MLM.

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|\boldsymbol{w}_{\text{mask}}|} \sum_{w_i \in \boldsymbol{w}_{\text{mask}}} \log P(w_i|\boldsymbol{s}_{\setminus \boldsymbol{w}_{\text{mask}}}; \mathcal{M}). \tag{2}$$

3.1.2 Voken classification task (VCT). The VCT enables the VLM to capture visually-induced knowledge by learning to predict vokens ('visualized tokens') assigned to tokens given a sentence. The vokens are natural images assigned by our Vokenizer (described in §3.2) considering the context of the input sentence. The use of the context means that even the same token in different contexts can be associated with different vokens. For example, the plausible voken for the token "dog" could be different in "a small white dog" and "a tall brown dog" due to the different contexts. By contrast, the voken for "white" in "a small white dog" might also be different from the one for "white clouds." Thus, the VLM should learn such visually-enriched representations of tokens through VCT. This type of learning cannot be achieved only by MLM because the supervision in MLM is always identical to the corresponding masked tokens even in different contexts.

Similar to the MLM formulation, the VLM predicts a conditional probability of vokens $\{v_j\}_{j=1}^n$ given a masked sentence $\boldsymbol{s}_{\setminus \boldsymbol{w}_{\text{mask}}}$.[2] We compute a probability distribution of the voken, $P(v_j \mid \boldsymbol{s}_{\setminus \boldsymbol{w}_{\text{mask}}}; \mathcal{M})$, over the voken vocabulary $\mathcal{V}_{\text{voken}}$ from the last hidden states $\boldsymbol{h}_j \in \mathbb{R}^d$ of tokens in the VLM.

$$P(v_j|\boldsymbol{s}_{\setminus \boldsymbol{w}_{\text{mask}}}; \mathcal{M}) = f_{\text{voken}}(\boldsymbol{h}_j), \tag{3}$$

in which $f_{\text{voken}}(\cdot)$ is an MLP that projects the hidden states onto a continuous space with dimension equal to the vocabulary size

[1] We follow the masking strategy proposed in [16].

[2] While we could give the original complete sentence s to the VLM for VCT, we follow the work [65] to keep the training consistent.

$|\mathcal{V}_{\text{voken}}|$ using an activation function. We do not share the model parameters between the two MLPs f_{token} and f_{voken} while we do share the model parameters of the VLM in MLM and VCT. Based on this output, we compute the negative log likelihood for VCT,

$$\mathcal{L}_{\text{VCT}} = -\frac{1}{n}\sum_{j=1}^{n} \log P(v_j | \boldsymbol{s}_{\setminus \boldsymbol{w}_{\text{mask}}}; \mathcal{M}). \tag{4}$$

Combining with $\mathcal{L}_{\text{MLM}}$, we minimize the joint objective function throughout the VLM training,

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{VCT}}. \tag{5}$$

3.2 Vokenization

We need a corpus where vokens are assigned to tokens for VCT training. However, no such corpus exists and annotating vokens is very costly and might not be feasible because some tokens may not have a visual correspondence (e.g., abstract or function words). To this end, we automatically annotate vokens using pseudo-labels produced by a Vokenizer at the expense of correctness.[3] Acting as a token-level image retriever, the Vokenizer retrieves vokens (images) related to tokens from an image set given a sentence.

3.2.1 Training. To train the Vokenizer, we leverage a multi-modal corpus, in which an image is paired with textual descriptions that describe or are related to the image. The previous work [65] used COCO [43] and Visual Genome [38], but the descriptions accompanied by images in these corpora are written in English. Indeed, we use WIT [63], a multilingual multi-modal corpus constructed from Wikipedia, to study the capability of the cross-lingual transfer of the VLM. WIT contains descriptions written in around 100 languages in total and images can be linked from multiple Wikipedia articles written in different languages, which means that some descriptions in different languages may be loosely coupled via images. However, it is not always ensured that these descriptions are the direct translations to each other since the images can be used in different contexts in different articles. [4]

The Vokenizer is composed of a language encoder and vision encoder and trained by contrastive learning [26, 60]. We treat the existing image-text (description) pair in WIT as a positive instance while we randomly create a negative instance by replacing an image with a different image taken from another image-text pair. Through contrastive learning, the Vokenizer learns representations of texts and images so that they are close to each other for positive instances while more distant for negative instances.

More concretely, we adopt a pre-trained XLM-R [11] and ResNeXt [75] as the language and vision encoders, respectively. Given an image v and text t as an input instance, we separately encode them by the respective encoders to obtain the feature representations.

$$\{\boldsymbol{h}_i\}_{i=1}^{l} = \text{Encoder}_{\text{lan}}(\{w_i\}_{i=1}^{l}), \tag{6}$$

$$\boldsymbol{g} = \text{Encoder}_{\text{vis}}(v), \tag{7}$$

where $\text{Encoder}_{\text{lan}}$ and $\text{Encoder}_{\text{vis}}$ are the language and vision encoders, respectively, and we apply a tokenizer to the input text t to split it into subword tokens $\{w_i\}_{i=1}^{l}$ before encoding. Note that we utilize the feature representations of each token[5] for Vokenization and also that we feed all the tokens to the language encoder at a time so that the encoder can consider the surrounding context of each token in the input text t when computing the feature representation of each token.

We then apply an MLP to both representations to project them onto the same feature space.

$$\boldsymbol{h}'_i = \text{Normalize}(f_{\text{lan}}(\boldsymbol{h}_i)), \tag{8}$$

$$\boldsymbol{g}' = \text{Normalize}(f_{\text{vis}}(\boldsymbol{g})). \tag{9}$$

$\text{Normalize}(\cdot)$ applies L2 normalization to the input vector to be of a length of 1. Based on the outputs $\boldsymbol{h}'_i, \boldsymbol{g}' \in \mathbb{R}^{d'}$ of the MLPs f_{lan}, f_{vis}, a scoring function $r(v, w_i)$ is calculated by a dot product of the feature representations.

$$r(v, w_i) = \boldsymbol{g}'^{\top} \cdot \boldsymbol{h}'_i. \tag{10}$$

Since the score is equivalent to the cosine similarity between the two normalized representations, the score represents the relevance between the image v and the token w_i in the text t.

Considering a positive and negative instance, we compute a triplet loss:

$$\mathcal{L}_{\text{Voken}} = \frac{1}{l}\sum_{i=1}^{l} \max\{0, 0.5 - r(v, w_i) + r(\tilde{v}, w_i)\}, \tag{11}$$

in which $\tilde{v}$ is a negative image randomly taken from another instance. To minimize this loss, we update only the model parameters of the MLPs, fixing those of the language and vision encoders.

While this might be a sub-optimal way to learn the token-level exact correspondence between vision and language, this can actually work quite well for VLM training. We hypothesize about this as follows. End-to-end neural machine translation models implicitly and automatically learn the token- or phrase-level alignment between a source and target language from a sentence-level parallel corpus (it is not perfect, though) even though the ground truth annotation of the token- or phrase-level alignment is not available [3, 77]. Similarly, in an image-text pair used for training the Vokenizer, there should also be an implicit correspondence between visual objects in the image and textual expressions in the text. After training the Vokenizer on a large multi-modal corpus, the Vokenizer should globally learn this correspondence. More specifically, if some visual objects (e.g., "dog") frequently co-occur with the corresponding textual expressions (e.g., "dog" or "puppy"), the visual feature $\boldsymbol{g}'$ and the textual feature $\boldsymbol{h}'$ should be accordingly updated to be close to each other during the training. Thus, arbitrary textual expressions can statistically be associated with some specific images based on the learned feature space.

3.2.2 Inference. We apply the trained Vokenizer to a text corpus $\mathcal{T}_{\text{txt}}$ used for VLM training to annotate vokens. We conduct a nearest neighbor search [35] over an image datastore to find an image (voken) relevant to each token in an input sentence.

First, we construct an image datastore $\mathcal{D}$ from an image corpus $\mathcal{T}_{\text{img}}$. We utilize a subset of the images in the WIT dataset. Given a set of images $\mathcal{T}_{\text{img}} = \{v_1, \ldots v_{|\mathcal{V}_{\text{voken}}|}\}$, we extract the visual features

[3] This is the same setting as the original work [65].

[4] For example, the image found at https://en.wikipedia.org/wiki/File:Classical_spectacular_laser_effects.jpg is used in different contexts in different articles such as https://en.wikipedia.org/wiki/Multimedia, https://en.wikipedia.org/wiki/Lighting, and https://it.wikipedia.org/wiki/Laser, and they are not the translations to each other.

[5] We interchangeably use 'token' instead of 'subword token' throughout the paper for brevity unless otherwise explicitly stated.

Table 1: Hyper-parameter settings of our evaluation. lr: learning rate, bs: batch size, #Epochs: number of epochs, Warm-ups: warm-up steps, and #Seeds: number of seeds.

Task	lr ($\times 10^{-5}$)	bs	#Epochs	Warm-ups	#Seeds
XNLI	{1, 2, 3}	{4, 8}	5	0	5
NER	{1, 2, 3}	{2, 4}	{6, 7, 8}	0	5
TyDiQA	{3, 4, 5}	{2, 4}	{1, 2, 3}	{500, 1000}	10
GLUE	2.5	32	3	0	10

Table 2: Evaluation results (accuracy) on XNLI. en, sw, and ur are the two-letter language codes defined by ISO 639-1 and denote English, Swahili, and Urdu, respectively. Better average scores are marked in bold. * indicates that the performance gap is greater than two sigmas (i.e., standard deviations), which are denoted by subscripts.

Model	en	sw	ur
LLM	$77.50_{\pm 0.25}$	$39.21_{\pm 0.65}$	$41.37_{\pm 0.37}$
VLM (ours)	$\mathbf{80.22}^{*}_{\pm 0.43}$	$\mathbf{44.57}^{*}_{\pm 1.19}$	$\mathbf{59.15}^{*}_{\pm 0.86}$

using the vision encoder in the Vokenizer. We couple the visual features with the corresponding input images to form an instance of the image datastore.

$$\mathcal{D} = \{(\boldsymbol{g}', v_j) | v_j \in \mathcal{T}_{\text{img}}\}, \tag{12}$$

where we obtain $\boldsymbol{g}'$ from Equations (7) and (9).

We then apply the nearest neighbor search for each token in an input sentence. Given an input sentence s from a text corpus $\mathcal{T}_{\text{txt}}$, we tokenize it, $s = \{w_i\}_{i=1}^{n}$, and encode it with the language encoder in the Vokenizer to obtain the language features $\{\boldsymbol{h}'_i\}_{i=1}^{n}$ (Equations (6) and (8)). Again, note that these features are contextualized in the input sentence. We can utilize the relevance score $r(v, w_i)$ for the nearest neighbor search.

$$v^{*} = \text{argmax}_{(\boldsymbol{g}', v_j) \in \mathcal{D}} r(v_j, w_i). \tag{13}$$

These vokens constitute the visually-derived supervision in VCT training, i.e., Equation (4).

4 EXPERIMENTS

4.1 Experimental setup

4.1.1 Datasets. We utilized WIT [63][6] to train our cross-lingual Vokenizer. We obtained the metadata and descriptions from the official site while we manually collected the images from Wikipedia using the URLs specified in the metadata. We stored the images with a size of 256 × 256 and in RGB format. The resultant WIT dataset consisted of 11,355,430 images and 16,335,736 descriptions in 106 languages. Out of 11M images, 3,238,941 images were paired with English descriptions while only 36,315 and 21,924 images were paired with Urdu and Swahili, respectively.

[6]https://github.com/google-research-datasets/wit

Table 3: Evaluation results (F1) on NER. See caption of Table 2 for additional details.

Model	en	sw	ur
LLM	$81.28_{\pm 0.13}$	$56.85_{\pm 0.94}$	$33.08_{\pm 1.68}$
VLM (ours)	$\mathbf{82.85}^{*}_{\pm 0.23}$	$\mathbf{60.97}^{*}_{\pm 2.06}$	$\mathbf{45.85}^{*}_{\pm 3.24}$

Table 4: Evaluation results (F1) on TyDiQA. See caption of Table 2 for additional details.

Model	en	sw
LLM	$49.26_{\pm 1.69}$	$30.76_{\pm 3.26}$
VLM (ours)	$\mathbf{54.17}^{*}_{\pm 1.44}$	$\mathbf{34.00}_{\pm 2.85}$

We trained our cross-lingual VLM on a mixture of Wikipedia[7] and CC-100[8] [71], in which English (en) texts were from Wikipedia (same as [65]) while we took Swahili (sw) and Urdu (ur) texts from CC-100 (considered as low-resource languages in [11]). After tokenization, we obtained 2,776,717,588 tokens for English, 275,268,340 tokens for Swahili, and 894,500,384 tokens for Urdu.

We evaluated our VLM on four downstream tasks. We considered three cross-lingual tasks: natural language inference (NLI), named entity recognition (NER), and extractive span-based question answering (QA). For each task, we used the XNLI [13], panx [13], and TyDiQA [9] datasets, respectively, all of which were taken from the XTREME benchmark [28]. For these cross-lingual tasks, the training set is always English while the test set is English and the target language, such as Swahili and Urdu. In addition, we evaluated our VLM on GLUE [69], which is a natural language understanding benchmark of English consisting of various NLP tasks, to study the performance of our VLM on a high-resource language.

4.1.2 Implementation. We built our cross-lingual Vokenizer and VLM by extending the implementation of [65][9], which is based on PyTorch [56] and Hugging Face [72].

As for Vokenizer, we used the pre-trained XLM-R [11] (`xlm-roberta-base`) provided by Hugging Face[10] and the pre-trained ResNeXt [75] (`resnext101_32x8d`) provided by TorchVision [50] as the language and vision encoders, respectively. We concatenated the hidden states extracted from the last four layers of XLM-R and considered the resultant 3072-dimension vector as the language feature representation while we treated the output vector (of 2048-dimension) of the last fully-connected layer in ResNeXt as the vision feature representation. We built the two MLPs (in Equations (8) and (9)) with a linear layer that projects the respective feature representation onto the intermediate feature space of 256-dimension, ReLU activation [53], dropout with a probability of 0.3, and another linear layer that projects the intermediate feature onto a 64-dimension vector space ($d' = 64$). We trained our cross-lingual Vokenizer on WIT with an effective batch size of 5120 for 20 epochs, using Adam [34] with a learning rate of 10^{-3}. The training took 19 hours using nine

[7]https://dumps.wikimedia.org/
[8]https://data.statmt.org/cc-100/
[9]https://github.com/airsplay/vokenization
[10]https://huggingface.co/xlm-roberta-base

Table 5: Evaluation results on GLUE. Train size indicates the number of examples in the train set. Acc.: accuracy, MCC: Matthews correlation coefficient, Corr: average of Pearson and Spearman correlations, and Average: macro average on eight tasks. See caption of Table 2 for the other details.

Task	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
Train size	393K	364K	105K	67K	8.5K	5.7K	3.7K	2.5K	-
Metric	Acc.	F1	Acc.	Acc.	MCC	Corr.	F1	Acc.	-
LLM	$77.76_{\pm 0.17}$	$85.60_{\pm 0.10}$	$85.87_{\pm 0.30}$	$89.22_{\pm 0.53}$	$31.80_{\pm 1.91}$	$84.24_{\pm 0.33}$	$87.34_{\pm 0.82}$	$\mathbf{60.07}_{\pm 1.60}$	$75.24_{\pm 0.72}$
VLM (ours)	$\mathbf{80.64}^{*}_{\pm 0.25}$	$\mathbf{86.61}^{*}_{\pm 0.13}$	$\mathbf{88.07}^{*}_{\pm 0.37}$	$\mathbf{90.77}^{*}_{\pm 0.33}$	$\mathbf{37.07}^{*}_{\pm 1.45}$	$\mathbf{85.72}^{*}_{\pm 0.34}$	$\mathbf{88.62}_{\pm 0.88}$	$58.34_{\pm 3.07}$	$\mathbf{76.98}^{*}_{\pm 0.85}$

NVIDIA Tesla V100 GPUs. To conduct the nearest neighbor search, we used FAISS [31], a fast and efficient similarity search library. We constructed the image datastore with the prototypical 50K images from WIT using K-means clustering [20, 47] ($K = 50,000$) based on the visual features encoded by ResNeXt.

For our VLM, we adopted XLM-R with the base architecture, which has 12 layers and 768 dimensions for the hidden states ($d = 768$). The MLPs used for MLM and VCT (in Equations (1) and (3)) consisted of a linear layer that projects the output hidden states of XLM-R to an intermediate feature space of the same dimension as that of the hidden state, GeLU activation [27], layer normalization [2], and another linear layer that projects the intermediate feature to a vector space of the respective vocabulary size. We trained our VLM on our three language corpora described in §4.1.1 from scratch, with an effective batch size of 256 for 200K steps, using AdamW [48] and a learning rate of 10^{-4} with a warm-up of 10K steps. The training took 5.6 days with four NVIDIA Tesla V100 GPUs.

To evaluate our VLM on the downstream tasks, we fine-tune our VLM on each downstream language task by adding an MLP on top of the VLM. Note that the fine-tuning no longer involves any visual data. Following the standard practice [16], we composed the MLP for sequence classification tasks (i.e., XNLI and GLUE) with a linear layer that projects the output hidden states of the `[CLS]` tokens in the VLM onto an intermediate feature space of the same dimension as that of the hidden state, Tanh activation, dropout with a probability of 0.1, and another linear layer that projects the intermediate feature to a vector space whose dimension corresponds to the number of the target labels. We composed the MLP for NER with a dropout with a probability of 0.1, and a linear layer that projects the intermediate feature to a vector space whose dimension corresponds to the number of the target labels. As for TyDiQA, the MLP consisted of a linear layer that projects the output hidden states from the VLM onto a vector space of two dimensions, which corresponds to the start and end positions of the answer span, respectively. Table 1 summarizes additional configuration of the hyper-parameters used in our evaluation. We used one NVIDIA A100 GPU for each evaluation and the elapsed time including fine-tuning varied from one minute to 3.8 hours depending on the task and hyper-parameters.

4.2 Main results

Tables 2 through 5 report the experimental results on four downstream tasks.[11] Since Urdu is not included in TyDiQA, we report the results for English and Swahili. We computed the average scores and standard deviations for multiple random seeds as shown in Table 1 (#Seeds). For the cross-lingual tasks, we conducted a hyper-parameter sweep using grid search and report the best average scores (averaged on random seeds for the best grid point). We decided the grid search space based on the characteristics of each task; we considered more grid points and random seeds for unstable tasks that tend to result in higher standard deviations. We compared our VLM with an LLM baseline that was trained on exactly the same text corpora (i.e., English Wikipedia and CC-100 Swahili and Urdu) using the same configuration but excluding the VCT loss, which means that the LLM was trained with the MLM loss only (Equation (2) instead of Equation (5)).

Our VLM significantly outperformed the LLM baseline by margins larger than two standard deviations[12] in most tasks except for TyDiQA (sw) and the smaller tasks in GLUE, i.e., MRPC and RTE.[13] In particular, our VLM that combines the VCT loss with the MLM loss led to large improvements in cross-lingual transfer learning, such as a 17.78 point improvement on Urdu (ur) in the XNLI task (Table 2), and a 12.77 point improvement on Urdu in NER (Table 3). It is also noteworthy that these large improvements were achieved by visually-derived supervision in VCT that was obtained by our cross-lingual Vokenizer even though the WIT datasets used to train the Vokenizer contained very few images relevant to the low-resource languages (36K for Urdu and 22K for Swahili) compared to those for English (3M). Furthermore, our VLM performed better in the monolingual evaluation setting, in which the model is fine-tuned and tested on the same language, that is, English (en) in XNLI, NER, and TyDiQA, as well as GLUE. Overall, these improvements clearly demonstrate the effectiveness of visually-derived supervision using Vokenization both for cross-lingual transfer learning and to improve the performance of a large language model in NLP tasks.

5 ANALYSES

While we demonstrated a clear performance improvement for the VLM in cross-lingual tasks in the previous section, it raised two new questions: (1) how does the cross-lingual transfer learning by the Vokenizer compare to that of cross-lingual texts? and (2) can visual information complement the cross-lingual transfer from texts? We

[11] Following the existing work [49, 65, 66], we report the results on the dev set for GLUE and TyDiQA since the ground truth labels are not published for the test set.

[12] This means that there is no overlap on the standard deviations from both models. In the case of XNLI (en) in Table 2, the score of our VLM is larger than that of the LLM plus two standard deviations: 80.22-0.43 > 77.50+0.25.

[13] One may notice the absence of the WNLI task in the GLUE results, but existing works [16, 46, 49, 65, 66] have also been reporting the scores on a subset of GLUE, excluding unstable tasks. Since we observed exceptionally higher standard deviations of both models for WNLI (i.e., greater than 7), we considered WNLI as abnormal and excluded it from the evaluation.

Table 6: Comparison between monolingual and cross-lingual models in a monolingual setting in XNLI. See caption of Table 2 for additional details.

Model	sw	ur
LLM-sw	$58.00_{\pm 0.60}$	N/A
VLM-sw	$\mathbf{59.92}^{*}_{\pm 0.32}$	N/A
LLM-ur	N/A	$57.80_{\pm 0.27}$
VLM-ur	N/A	$\mathbf{58.90}^{*}_{\pm 0.36}$
LLM	$53.43_{\pm 1.23}$	$55.40_{\pm 0.54}$
VLM (ours)	$\mathbf{56.39}^{*}_{\pm 0.36}$	$\mathbf{57.90}^{*}_{\pm 0.64}$

Table 7: Lexical diversity.

Language	Lexical diversity
sw	0.001771
ur	0.002249

Table 8: Subword overlap.

Language pair	Subword overlap
(sw, en)	0.3732
(ur, en)	0.3202

Table 9: Shared voken usage on training corpora.

Language pair	Top-10	Top-100	Top-1000
(sw, en)	0.08952	0.02024	0.003945
(ur, en)	0.10250	0.02224	0.005408

seek to answer these questions with the following analyses. Our first analysis (§5.1) disentangles the effects of visual information and cross-lingual transfer from texts. Our subsequent analyses (§5.2 and §5.3) verify that visual information encourages cross-lingual transfer even between a linguistically-distant pair.

5.1 Monolingual experiment

We trained the monolingual version of our VLM and the LLM baseline with the same configuration described in §4.1 except for a difference in the training data; we used either only Urdu or only Swahili texts from CC-100. We then fine-tuned the trained monolingual models on the low-resource languages in the XNLI task (using the dev set due to no training data being available in these languages), and tested them on the same languages.

Table 6 shows the results. The first section (*-sw and *-ur) shows the results of the monolingual versions of the models trained on that language. The second section (LLM and VLM (ours)) presents the results of the cross-lingual versions evaluated in the monolingual setting. Regardless of the training setting (i.e., monolingual or cross-lingual), our VLMs outperformed the corresponding LLMs. The outperformance of the monolingual VLMs over the monolingual LLMs indicates the effectiveness of Vokenization since the only difference between the VLM and LLM is the loss function. The outperformance of the cross-lingual VLM over the cross-lingual LLM in the monolingual setting suggests that the VLM achieved better language representations derived from the corresponding token embeddings of the model in the low-resource languages. We also note that the absolute values of the scores in the monolingual setting (Table 6) were typically higher than those in the cross-lingual setting (Table 2). This is natural because the monolingual setting uses the same language both for fine-tuning and evaluation, and it is therefore easier for the models to adjust themselves to that language and data. Moreover, the performance gap between our VLM and the LLM counterpart is larger in the cross-lingual setting (Table 2) than that in the monolingual setting (Table 6): 5 to 17 points in the cross-lingual setting and 2.5 to 3 points in the monolingual setting. This implies that the VLM acquired better cross-lingual representations and the performance improvement in the cross-lingual setting was boosted by the cross-lingual transfer learning of visual supervision.

5.2 Language similarity and voken usage

Next, we analyzed the training data used for VLM training from a language perspective. Prior work [44] proposed to select source languages useful for cross-lingual transfer learning in NLP tasks based on multiple linguistic features. Among these features, we measured the lexical diversity and subword overlap to inspect the characteristics of the data. The lexical diversity is the ratio of the number of unique token types, i.e., vocabulary, on a dataset to the total number of tokens in the dataset. A higher lexical diversity suggests that training will be difficult because the language model needs to capture a more diverse set of linguistic expressions. The subword overlap is the ratio of the number of unique token types commonly used in two languages to the sum of the numbers of the vocabulary for the two languages. Intuitively, two languages that are linguistically similar will have a higher subword overlap.

Tables 7 and 8 show the results. From Table 7, we observe that Urdu (ur) is more lexically diverse than Swahili (sw), which indicates that it is more difficult for the language model to learn good representations for Urdu in the low-resource scenario. Furthermore, Table 8 indicates that Urdu has less subword overlap. This suggests that it is less likely that the cross-lingual language model can take advantage of cross-lingual transfer from English during training. However, this is somewhat not consistent with the fact that the performance gap is larger in Urdu than Swahili in Tables 2 and 3.

Thus, we hypothesize that these larger performance gains in Urdu were a result of the visually-induced supervision through VCT. To reveal specifically which part of the training contributed to these improvements, we took a closer look at its training mechanism. In VCT (Equation (3) and the left side of Figure 2), our VLM is trained to predict vokens using its language representations, i.e., the last hidden states. There could be vokens visually similar to each other in the training data, and when predicting these vokens, the VLM should compute similar language representations to each other even if the corresponding tokens are different or even in different languages. This could encourage cross-lingual transfer learning,

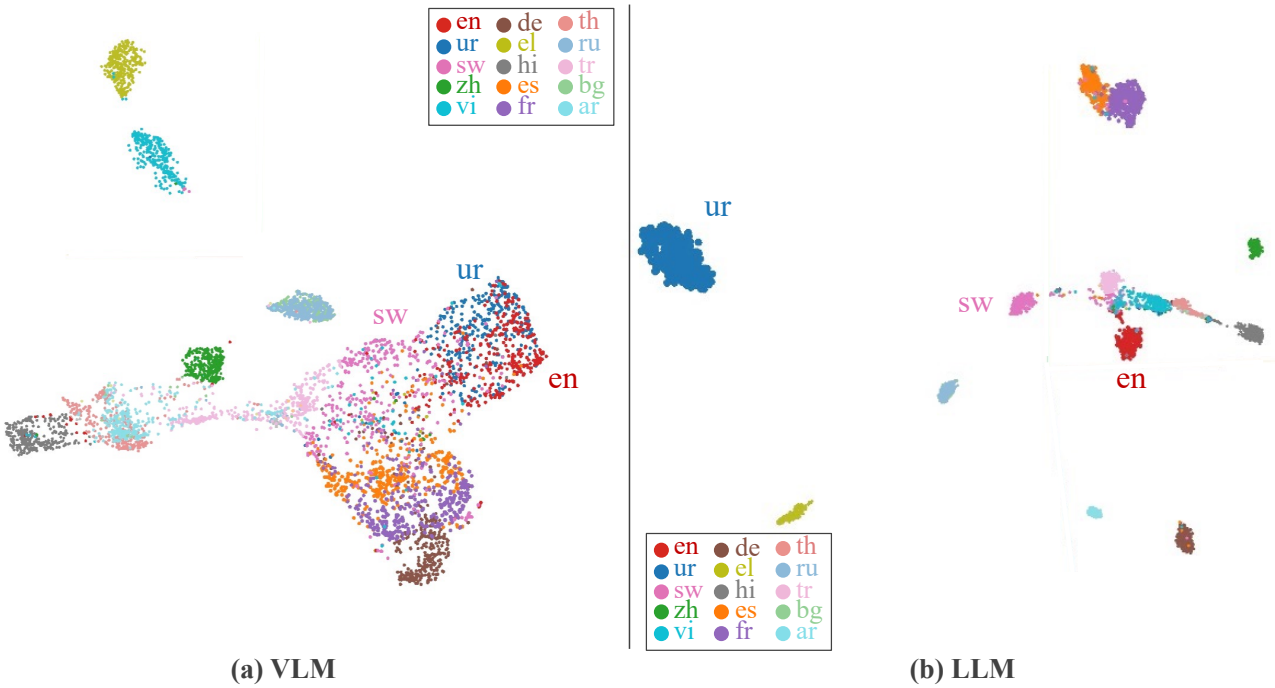

Figure 3: Visualization of language representations encoded by our VLM and LLM. The best view can be seen in color.

especially when more vokens assigned to English tokens are also assigned to tokens in low-resource languages.

We adopted the mean reciprocal rank (MRR) to measure the voken usage across languages considering its frequency.

$$\mathrm{MRR}(\mathcal{V}_{\mathrm{trg}}, \mathcal{V}_{\mathrm{src}}, k) = \frac{1}{k} \sum_{v_i \in \mathrm{top}(\mathcal{V}_{\mathrm{trg}}, k)} \frac{1}{\mathrm{rank}(v_i; \mathcal{V}_{\mathrm{src}})}, \tag{14}$$

in which $\mathrm{top}(\mathcal{V}_{\mathrm{trg}}, k)$ returns the most frequent top-k vokens in the voken vocabulary $\mathcal{V}_{\mathrm{trg}}$ in the target language while $\mathrm{rank}(v_i; \mathcal{V}_{\mathrm{src}})$ returns the rank of the frequency of the voken v_i in the voken vocabulary $\mathcal{V}_{\mathrm{src}}$ in the source language. For example, $\mathrm{MRR}(\mathcal{V}_{\mathrm{sw}}, \mathcal{V}_{\mathrm{en}}, 10)$ is the MRR between Swahili and English.

We measured the MRR for vokens used in the corpora for the low-resource languages ($k \in \{10, 100, 1000\}$) and report the results in Table 9. Clearly, the vokens assigned to the Urdu tokens are more frequently used in the English tokens. This means that our VLM had more chances for cross-lingual transfer in Urdu than Swahili through VCT by predicting more vokens shared with English texts. Combined with the results for the language similarity, we conclude that although Urdu is linguistically distant from English, visually-derived supervision compensated for this limitation.

5.3 Feature visualization

To intuitively understand the learned language representations, we visualize what the representations look like in their feature space. For this purpose, we obtained the language representations (of the `[CLS]` token) from the last hidden layers in both our VLM and the LLM counterpart for each instance in the dev set of the XNLI dataset. We visualized the language representations in Figure 3 using a dimension reduction technique [51]. While the XNLI dataset contains 15 languages, we focus on the three languages used in our experiments, that is, English (en), Swahili (sw), and Urdu (ur).

In Figure 3a, the English and Urdu representations are partially mixed and the Swahili ones are located next to them. This implies that our VLM learned better cross-lingual representations as shown in the previous sections (§4.2 and §5.2) especially between English and Urdu. In contrast, Figure 3b shows that the representations of the three (and more) languages are distant from each other, suggesting that the LLM struggled with cross-lingual transfer learning.

6 CONCLUSION

In this work, we have studied an interesting question: does visually-derived supervision contribute to cross-lingual transfer learning for low-resource languages in NLP tasks. In particular, we extended the Vokenization approach to the cross-lingual setting using multilingual and multi-modal datasets. Our experimental results demonstrated that the cross-lingual VLM significantly outperformed the LLM baseline by large margins in multiple cross-lingual NLP tasks. Our detailed analysis has shown that visually-derived supervision can help even more for linguistically-distant languages. We hope this work sheds light on the line of research for the cross-lingual transfer of LLMs using visual information.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable comments. We also thank our colleagues for their thoughtful discussions and suggestions that greatly improved our work.

REFERENCES

[1] Gerry TM Altmann and Yuki Kamide. 2004. Now you see it, now you don't: mediating the mapping between language and the visual world. In *The interface of language, vision, and action: eye movements and the visual world*. 347–368.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. arXiv:1607.06450 [stat.ML]

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*. http://arxiv.org/abs/1409.0473

[4] Paul Bloom. 2000. How children learn the meanings of words. (2000).

[5] Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2019. Incorporating Visual Semantics into Sentence Representations within a Grounded Space. In *EMNLP-IJCNLP*. 696–707. https://doi.org/10.18653/v1/D19-1064

[6] Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual Visual Pre-training for Multimodal Machine Translation. In *EACL*. 1317–1324. https://doi.org/10.18653/v1/2021.eacl-main.112

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *ICLR*. https://openreview.net/forum?id=mWVoBz4W0u

[8] Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. 2016. Resolving Language and Vision Ambiguities Together: Joint Segmentation & Prepositional Attachment Resolution in Captioned Scenes. In *EMNLP*. 1493–1503. https://doi.org/10.18653/v1/D16-1156

[9] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *TACL* 8 (2020), 454–470. https://doi.org/10.1162/tacl_a_00317

[10] Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined Visual Representations as Multimodal Embeddings. In *AAAI*, Vol. 31. https://doi.org/10.1609/aaai.v31i1.11155

[11] 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 *ACL*. 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[12] Alexis Conneau and Guillaume Lample. 2019. Cross-Lingual Language Model Pretraining. In *NeurIPS*. https://dl.acm.org/doi/abs/10.5555/3454287.3454921

[13] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *EMNLP*. 2475–2485. https://doi.org/10.18653/v1/D18-1269

[14] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging Cross-lingual Structure in Pretrained Language Models. In *ACL*. 6022–6034. https://doi.org/10.18653/v1/2020.acl-main.536

[15] Banchiamlack Dessalegn and Barbara Landau. 2013. Interaction between language and vision: It's momentary, abstract, and it develops. *Cognition* 127, 3 (2013), 331–344. https://doi.org/10.1016/j.cognition.2013.02.003

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Multilingual BERT readme document*. Retrieved April 1, 2023 from https://github.com/google-research/bert/blob/master/multilingual.md

[18] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*. 70–74. https://doi.org/10.18653/v1/W16-3210

[19] Hongliang Fei, Tan Yu, and Ping Li. 2021. Cross-lingual Cross-modal Pretraining for Multimodal Retrieval. In *NAACL*. 3644–3650. https://doi.org/10.18653/v1/2021.naacl-main.285

[20] Edward W Forgy. 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *biometrics* 21 (1965), 768–769.

[21] Ruka Funaki and Hideki Nakayama. 2015. Image-Mediated Learning for Zero-Shot Cross-Lingual Document Retrieval. In *EMNLP*. 585–590. https://doi.org/10.18653/v1/D15-1070

[22] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in Cognitive Sciences* 3, 11 (1999), 419–429. https://doi.org/10.1016/S1364-6613(99)01397-2

[23] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *OntoImage 2006 Workshop on Language Resources for Content-based Image Retrieval during LREC 2006 Final Programme*.

[24] Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. 2018. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval With Generative Models. In *CVPR*. https://openaccess.thecvf.com/content_cvpr_2018/html/Gu_Look_Imagine_and_CVPR_2018_paper.html

[25] Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Qinyu Zhang, and Ji-Rong Wen. 2023. Visually-augmented pretrained language models for NLP tasks without images. In *ACL*. 14912–14929. https://aclanthology.org/2023.acl-long.833

[26] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *CVPR*. 1735–1742. https://doi.org/10.1109/CVPR.2006.100

[27] Dan Hendrycks and Kevin Gimpel. 2020. Gaussian Error Linear Units (GELUs). arXiv:1606.08415 [cs.LG]

[28] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *ICML*, Vol. 119. 4411–4421. https://proceedings.mlr.press/v119/hu20b.html

[29] Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models. In *NAACL*. 2443–2459. https://doi.org/10.18653/v1/2021.naacl-main.195

[30] Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling Translations with Visual Awareness. In *ACL*. 6525–6538. https://doi.org/10.18653/v1/P19-1653

[31] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2021), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[32] Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning Visually Grounded Sentence Representations. In *NAACL*. 408–418. https://doi.org/10.18653/v1/N18-1038

[33] Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual Bilingual Lexicon Induction with Transferred ConvNet Features. In *EMNLP*. 148–158. https://doi.org/10.18653/v1/D15-1015

[34] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*. http://arxiv.org/abs/1412.6980

[35] Donald Ervin Knuth. 1998. *The Art of Computer Programming*. Vol. 3: Sorting and Searching. Addison-Wesley.

[36] Noriyuki Kojima, Hadar Averbuch-Elor, Alexander Rush, and Yoav Artzi. 2020. What is Learned in Visually Grounded Neural Syntax Acquisition. In *ACL*. 2615–2635. https://doi.org/10.18653/v1/2020.acl-main.234

[37] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are You Talking About? Text-to-Image Coreference. In *CVPR*. https://openaccess.thecvf.com/content_cvpr_2014/html/Kong_What_are_You_2014_CVPR_paper.html

[38] 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. *IJCV* 123, 1 (may 2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

[39] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *NAACL*. 153–163. https://doi.org/10.3115/v1/N15-1016

[40] Jialu Li, Hao Tan, and Mohit Bansal. 2022. CLEAR: Improving Vision-Language Navigation with Cross-Lingual, Environment-Agnostic Representations. In *Findings of NAACL*. 633–649. https://doi.org/10.18653/v1/2022.findings-naacl.48

[41] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2021. Unsupervised Vision-and-Language Pre-training Without Parallel Images and Captions. In *NAACL*. 5339–5350. https://doi.org/10.18653/v1/2021.naacl-main.420

[42] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *ACL*. 2592–2607. https://doi.org/10.18653/v1/2021.acl-long.202

[43] 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 *ECCV*. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[44] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *ACL*. 3125–3135. https://doi.org/10.18653/v1/P19-1301

[45] Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not Written in Text: Exploring Spatial Commonsense from Visual Signals. In *ACL*. 2365–2376. https://doi.org/10.18653/v1/2022.acl-long.168

[46] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[47] S. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (1982), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[48] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*. https://openreview.net/forum?id=Bkg6RiCqY7

[49] Yujie Lu, Wanrong Zhu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-Augmented Natural Language Understanding. In *NAACL*. 4392–4402. https://doi.org/10.18653/v1/2022.naacl-main.326

[50] TorchVision maintainers and contributors. 2016. *TorchVision: PyTorch's Computer Vision library*. https://github.com/pytorch/vision

[51] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *JOSS* 3, 29 (2018), 861. https://doi.org/10.21105/joss.00861

[52] Masayasu Muraoka, Tetsuya Nasukawa, and Bishwaranjan Bhattacharjee. 2020. Visual Objects As Context: Exploiting Visual Objects for Lexical Entailment. In *Findings of EMNLP*. 2723–2735. https://doi.org/10.18653/v1/2020.findings-emnlp.246

[53] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*. 807–814. https://dl.acm.org/doi/10.5555/3104322.3104425

[54] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *ACL*. 1946–1958. https://doi.org/10.18653/v1/P17-1178

[55] Johanne Paradis, Fred Genesee, and Martha B Crago. 2011. Dual Language Development and Disorders: A Handbook on Bilingualism and Second Language Learning. *Brookes Publishing Company* (2011).

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[57] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT?. In *ACL*. 4996–5001. https://doi.org/10.18653/v1/P19-1493

[58] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019). https://openai.com/research/better-language-models

[59] Hyeonggon Ryu, Arda Senocak, In So Kweon, and Joon Son Chung. 2023. Hindi as a Second Language: Improving Visually Grounded Speech with Semantically Similar Samples. arXiv:2303.17517 [cs.CL] https://arxiv.org/abs/2303.17517

[60] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *CVPR*. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html

[61] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually Grounded Neural Syntax Acquisition. In *ACL*. 1842–1861. https://doi.org/10.18653/v1/P19-1180

[62] Richard Sinatra. 1981. Using Visuals to Help the Second Language Learner. *The Reading Teacher* 34, 5 (1981), 539–546. http://www.jstor.org/stable/20195283

[63] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning. In *SIGIR*. 2443–2449. https://doi.org/10.1145/3404835.3463257

[64] Dídac Surís, Dave Epstein, and Carl Vondrick. 2022. Globetrotter: Connecting Languages by Connecting Images. In *CVPR*. 16474–16484. https://openaccess.thecvf.com/content/CVPR2022/html/Suris_Globetrotter_Connecting_Languages_by_Connecting_Images_CVPR_2022_paper.html

[65] Hao Tan and Mohit Bansal. 2020. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. In *EMNLP*. 2066–2080. https://doi.org/10.18653/v1/2020.emnlp-main.162

[66] Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. VidLanKD: Improving Language Understanding via Video-Distilled Knowledge Transfer. In *NeurIPS*, Vol. 34. 24468–24481. https://proceedings.neurips.cc/paper_files/paper/2021/file/ccdf3864e2fa9089f9eca4fc7a48ea0a-Paper.pdf

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*, Vol. 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[68] Ivan Vulić, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. 2016. Multi-Modal Representations for Improved Bilingual Lexicon Learning. In *ACL*. 188–194. https://doi.org/10.18653/v1/P16-2031

[69] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *ICLR*. https://openreview.net/forum?id=rJ4km2R5t7

[70] Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Visually-Augmented Language Modeling. In *ICLR*. https://openreview.net/forum?id=8IN-qLkl215

[71] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *LREC*. 4003–4012. https://aclanthology.org/2020.lrec-1.494

[72] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP: System Demonstrations*. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[73] Zixiu Wu, Julia Ive, Josiah Wang, Pranava Madhyastha, and Lucia Specia. 2019. Predicting Actions to Help Predict Translations. In *ICML Workshop on The How2 Challenge: New Tasks for Vision and Language*. https://srvk.github.io/how2-challenge/assets/authors/1908.01665.pdf

[74] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. arXiv:1901.06706 [cs.CV]

[75] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *CVPR*. https://openaccess.thecvf.com/content_cvpr_2017/html/Xie_Aggregated_Residual_Transformations_CVPR_2017_paper.html

[76] Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2022. Z-LaVI: Zero-Shot Language Solver Fueled by Visual Imagination. In *EMNLP*. 1186–1203. https://aclanthology.org/2022.emnlp-main.78

[77] Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-End Neural Word Alignment Outperforms GIZA++. In *ACL*. 1605–1617. https://doi.org/10.18653/v1/2020.acl-main.146

[78] Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. 2022. MCSE: Multimodal Contrastive Learning of Sentence Embeddings. In *NAACL*. 5959–5969. https://doi.org/10.18653/v1/2022.naacl-main.436

[79] Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural Machine Translation with Universal Visual Representation. In *ICLR*. https://openreview.net/forum?id=Byl8hhNYPS

[80] Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, and Ning Zhang. 2022. Unsupervised Vision-and-Language Pre-Training via Retrieval-Based Multi-Granular Alignment. In *CVPR*. 16485–16494. https://openaccess.thecvf.com/content/CVPR2022/html/Zhou_Unsupervised_Vision-and-Language_Pre-Training_via_Retrieval-Based_Multi-Granular_Alignment_CVPR_2022_paper.html

[81] 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 *CVPR*. 4155–4165. https://openaccess.thecvf.com/content/CVPR2021/html/Zhou_UC2_Universal_Cross-Lingual_Cross-Modal_Vision-and-Language_Pre-Training_CVPR_2021_paper.html

[82] Wanrong Zhu, Xin Eric Wang, An Yan, Miguel Eckstein, and William Yang Wang. 2023. ImaginE: An Imagination-Based Automatic Evaluation Metric for Natural Language Generation. arXiv:2106.05970 [cs.CL]