Learning to Scale Multilingual Representations for Vision-Language Tasks

Andrea Burns[1], Donghyun Kim[1], Derry Wijaya[1], Kate Saenko[1,2], and Bryan A. Plummer[1]

[1] Boston University, Boston MA 02215, USA
[2] MIT-IBM Watson AI Lab, Cambridge MA 02142, USA
{aburns4,donhk,wijaya,saenko,bplum}@bu.edu

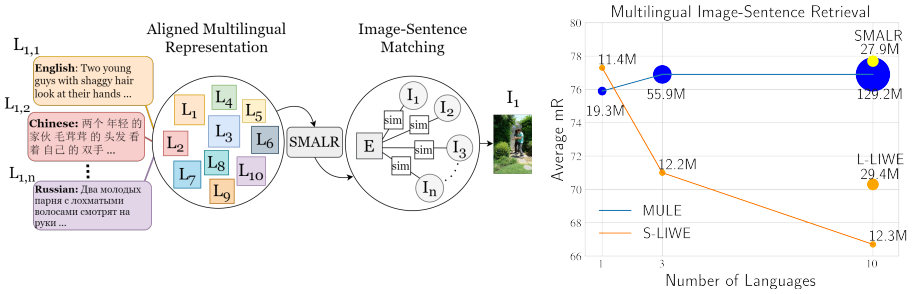
Abstract. Current multilingual vision-language models either require a large number of additional parameters for each supported language, or suffer performance degradation as languages are added. In this paper, we propose a Scalable Multilingual Aligned Language Representation (SMALR) that supports many languages with few model parameters without sacrificing downstream task performance. SMALR learns a fixed size language-agnostic representation for most words in a multilingual vocabulary, keeping language-specific features for just a few. We use a masked cross-language modeling loss to align features with context from other languages. Additionally, we propose a cross-lingual consistency module that ensures predictions made for a query and its machine translation are comparable. The effectiveness of SMALR is demonstrated with ten diverse languages, over twice the number supported in vision-language tasks to date. We evaluate on multilingual image-sentence retrieval and outperform prior work by 3-4% with less than 1/5th the training parameters compared to other word embedding methods.

Keywords: Scalable Vision-Language Models, Multilingual Word Embeddings, Image-Sentence Retrieval

1 Introduction

Learning a good language representation is a fundamental component of addressing a vision-language task, such as phrase grounding [22,34] or visual question answering [3,17]. Many recent methods have demonstrated that learning text representations aligned to images can boost performance across many vision-language tasks over traditional text-only trained representations [8,19,29,37,38]. This is often accomplished by using auxiliary vision-language tasks when learning the language representation (such as image-sentence retrieval, as shown in Figure 1(a)). However, these methods often only support a single language. Although some work has addressed a multilingual scenario (*e.g.*, [16,23,41]), these

Project page: http://ai.bu.edu/smalr



(a) Multilingual image-sentence retrieval (b) MSCOCO multilingual retrieval

Fig. 1: (a) presents multilingual bidirectional retrieval. We embed sentences in ten languages with SMALR, which is used to compute the highest scoring image. (b) shows the effect of the number of training languages on performance for prior work MULE [23] and LIWE [41]. LIWE is the original model, hereafter referred to as S-LIWE. The plot contains two points: L-LIWE, [41] trained with a larger embedding (120-D vs. 24-D) for fair comparison, in orange, and SMALR, in yellow. The points are scaled to the number of parameters, P; specifically, their area is $(\frac{P}{10^6})^{\frac{3}{2}}$. SMALR is able to outperform all prior work with few parameters

methods do not scale well to support many languages in terms of memory or performance (see Figure 1(b)). As the number of languages grows, methods like LIWE [41] that use character-based recognition systems can save memory but suffer from performance degradation. In contrast, methods that learn to align word embeddings across languages can maintain (or even improve) performance as languages are added (*e.g.*, [16,23]), but require additional parameters for the word embeddings that represent each new language's vocabulary. This becomes a challenge when scaling to support many languages, as an increasing majority of trainable parameters are required for representing each language (*e.g.* ∼93% of parameters of [23] with ten languages). While pretrained word embeddings could be used without fine-tuning, *e.g.* Multilingual BERT [13] or MUSE [11], this comes at a significant cost in downstream task performance [8,23].

To address this trade-off between multilingual capacity and performance, we propose a *Scalable Multilingual Aligned Language Representation (SMALR)* model, which we demonstrate achieves strong task performance while also being highly compact compared to state-of-the-art word embedding methods [13,24,26]. As seen in Figure 1, LIWE drops over 10% in performance going from supporting one to ten languages. MULE slightly increases performance with more languages, but requires 6x more parameters compared to its single language model. Our approach, SMALR, outperforms both with only 1/5th the parameters of MULE. We learn to efficiently represent each language by separating our language embedding into language-specific and language-agnostic token representations. As language follows a long-tailed distribution, only a few words occur often, with large portions of tokens occurring very rarely. For example, in the MSCOCO

dataset [28] there are 25,126 unique tokens, but 61% of them occur less than 4 times. This suggests that having unique representations for every token in the vocabulary is unnecessary, as only a subset would affect downstream task performance significantly. Thus, we use a Hybrid Embedding Model (HEM) that contains language-specific embeddings for the common tokens, thereby providing a good representation for each language, and a compact language-agnostic representation for rare and uncommon words. This results in a model that needs far fewer unique embeddings than prior work without sacrificing performance.

We learn how to assign tokens to the language-agnostic representation in a pretraining step, which uses monolingual FastText embeddings [7] to map similar words to the same token, *e.g.* mapping "double-decker" in English and "impériale" in French to the same shared token. Once we obtain our language embeddings, our goal is to align them so that semantically similar words, even those from other languages, are embedded nearby. To accomplish this, we use a multilingual masked language model, where we randomly mask words and then predict them based on context. Unlike similar masking approaches used to train models such as BERT [13], we mask words of sentences from any two languages, say German and Chinese, which are semantically similar sentences referring to the same image, and use the context from each to predict both masked tokens. To further encourage cross-language alignment, we also use an adversarial language classifier and neighborhood constraints that have been used in prior work [23]. These universal language embeddings are provided as input to a multimodal model that learns to relate them to images. Finally, we use a cross-lingual consistency module that uses machine translations to reason about the image-sentence similarity across multiple languages, which we show significantly boosts performance. Figure 2 contains an overview of our model.

We use bidirectional image-sentence retrieval as the primary evaluation of our multilingual language representation. In this task, the goal is to retrieve a relevant sentence from a database given an image or to retrieve a relevant image from a database given a sentence. We augment current multilingual datasets Multi30K [6,14,15,43] and MSCOCO [27,28,31] using machine translations so that every image has at least five sentences across ten diverse languages: English (En), German (De), French (Fr), Czech (Cs), Chinese (Cn), Japanese (Ja), Arabic (Ar), Afrikaans (Af), Korean (Ko), and Russian (Ru). See the supplementary for details on our data augmentation procedure. This constitutes the highest number of languages used in multilingual learning for vision-language tasks to date, supporting more than double the number of visually-semantically aligned languages compared to prior work [5,11,16,23,36,41].

We list the contributions of our work below:

- SMALR, a scalable multilingual model for training visually-semantically aligned word embeddings that outperforms the state-of-the-art on multilingual image-sentence retrieval while also requiring few model parameters.
- A comparison to four types of vocabulary reduction methods that serve as baselines to complement our evaluation against prior work.

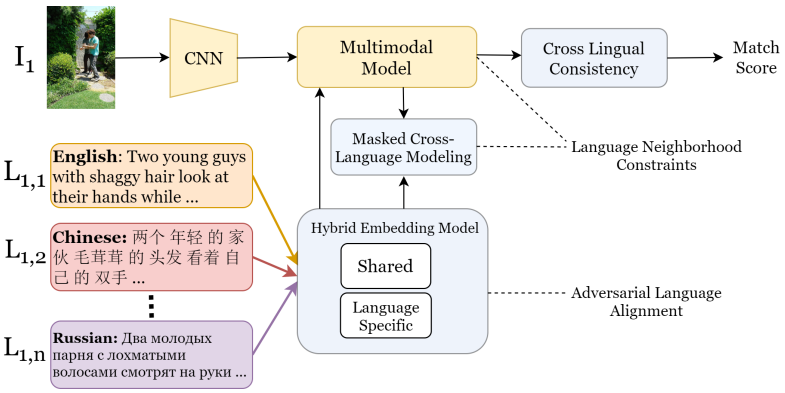


Fig. 2: The contributions of SMALR are in blue: a Hybrid Embedding Model (HEM), a Masked Cross-Language Model (MCLM), and a Cross-Lingual Consistency stage (CLC). HEM embeds input sentences as a mixture of language-specific and language-agnostic representations using a hard attention mechanism. The MCLM component provides an additional loss to enforce language alignment, while also augmenting the original dataset with masked sentences

- A Masked Cross-Language Modeling (MCLM) procedure that further aligns the multilingual embedding, stabilizing variance in performance over all languages, and serves as an additional data augmentation technique.
- A Cross-Lingual Consistency (CLC) module, the first of its kind, that learns how to aggregate an ensemble of predictions across languages made with machine translations, which, combined with our SMALR architecture, results in a total improvement over the state-of-the-art by 3-4%.

2 Related Work

Transformer [39] based representation learning models have become increasingly popular since the release of BERT [13]. BERT transfers surprisingly well to other languages, despite having no multilingual training data or explicit multilingual loss [42]. However, [33] demonstrates that there is an unequal transfer between different language pairs, notably those with typological differences to English. Both BERT and M-BERT, its multilingual extension, have been shown to be dependent on the number of parameters in the model, which reaches 110M parameters for the smaller base model [21]. Thus, as also shown in [1], a large number of additional parameters is needed to counter the performance degradation caused by training with many languages. Using the better performing large BERT model is impractical for many vision-language tasks as it contains 340M parameters, leaving little room in many GPUs memory for anything else.

Along with language-only BERT variants, a burst of multimodal BERT-like models have been designed specifically for vision-language tasks [26,29,37,38].

More traditional word embedding models have also been designed for multimodal tasks with the use of either visual-word co-occurrence frequencies [19], multi-task training [32], or both [8], and require significantly less training data to reach similar performance. While these efforts evaluate on many multimodal tasks such as Visual Question Answering [3], Visual Commonsense Reasoning [44], Phrase Grounding [34], and more, they only train and evaluate on a single language.

Recently, several multilingual methods have shown better performance on vision-language tasks than complicated transformer-based methods. LIWE [41] is a light-weight character embedding model that can represent many languages with few model parameters. LIWE uses a bidirectional gated recurrent unit (GRU) [9] to aggregate 24-D character embeddings for a text query that is encouraged to closely embed semantically similar images and sentences in other languages. Although LIWE represents a single language well, it suffers from significant performance loss when co-training on multiple languages as shown in Figure 1(b). Gella *et al*. [16] learns how to relate an image to language-specific representations and also constrain semantically similar sentences across languages to embed nearby each other. MULE [23] learns a universal language embedding so that it can use a single language branch in the multimodal model, significantly reducing the number of parameters required to represent each language compared to Gella *et al*. In addition, MULE combined the same cross-lingual constraints used in both Gella *et al*. and LIWE with an adversarial language classifier to further encourage alignment across languages. This results in a model that slightly improves performance as more languages are added as shown Figure 1(b). However, MULE learns a word-level embedding that requires significantly more parameters than LIWE (approximately 8x more with ten languages), and thus capacity concerns remain when scaling to many languages.

3 Scalable Multilingual Aligned Language Representation

In this section we describe how we train our Scalable Multilingual Aligned Language Representation (SMALR) to bridge the gap between scalability and downstream vision-language task performance. To accomplish this, we assume we are provided with an image and sentences that describe it in multiple languages. The intuition behind our model is to first learn a universal language embedding which represents all languages, and then learn to relate it to corresponding images using a multimodal model. In our experiments our multimodal model uses a modified version [23] of the Embedding Network architecture [40], although our approach can be easily adapted to use other multimodal models. After obtaining image and sentence features, the Embedding Network uses two branches, one for each modality, and projects them into a joint semantic space where distances are meaningful. The image branch consists of two fully connected layers, while the language branch obtains a sentence representation by passing the final hidden state of a GRU through a fully connected layer.

Our approach is architecturally similar to MULE [23], but with notable distinctions. First, MULE learned a unique word embedding for every word

in every language (*i.e.*, no shared tokens), whereas we learn an efficient universal embedding with our Hybrid Embedding Model (HEM) that consists of a mix of language-agnostic and language-specific word representations (Section 3.1). Then, we align our language representations using a novel Masked Cross-Language Model (MCLM) (Section 3.2) on both the input of the multimodal model and the final language representation of the multimodal model. This acts to supplement the neighborhood constraints, adversarial language classifier, and image-sentence matching losses used by MULE that we briefly review in Section 3.3. Finally, we also propose a Cross-Lingual Consistency (CLC) module that boosts model performance in downstream vision-language tasks using machine translation (Section 3.4). See Fig. 2 for an overview of our approach.

3.1 Efficient Multilingual Learning with a Hybrid Embedding Model

A significant challenge in multilingual representation learning is scaling to many languages, especially when there is a wide disparity in the available training data of different languages. This is more apparent for vision-language tasks where annotations are very expensive to collect, making it more difficult to learn a good visually-semantically aligned language representation than in monolingual settings [8,26]. Inspired by work in low-resource neural machine translation [18], we propose a Hybrid Embedding Model (HEM) which projects low-frequency words across languages into a shared latent vocabulary, while allowing the top-K most frequent words in each language to maintain their own unique (language-specific) representation. The output of the HEM is the universal language embedding that is used as input to the multimodal model in Fig. 2 and is also used in the language alignment losses (Section 3.2 and Section 3.3). The value of K can be determined experimentally for any targeted downstream task; we use $K = 5000$.

The language-specific word embeddings used for common words roughly follow the implementation used in prior work [18,23]. We begin by using a monolingual pretrained FastText embedding [11] that has been reduced from 300-D to 50-D using Principal Component Analysis (PCA) [30]. These reduced features are used as input to a fully connected layer that projects them into a 512-D universal embedding space that we align across languages; the alignment is applied with the language-agnostic representations as well (see Section 3.2 and 3.3 for details on our language alignment procedures).

While our language-agnostic representation is similar to Gu *et al.* [18], it has some key differences. Specifically, Gu *et al.* project all words into a universal embedding space with learned language-specific mappings. A soft-attention module is used over the universal embedding features (as it assumes an aligned cross-lingual input) to obtain mixing weights; these weights are then used to combine the language-agnostic features. While this does enable feature sharing across languages, it does not reduce the number of trainable parameters in the network, as a language-specific representation is still necessary for all words in the vocabulary. Additionally, aggregating the features in the latent vocabulary using soft-attention weights per-word is costly, especially for large vocabularies.

Instead, we perform a pretraining step where we learn both the initial representation of the latent vocabulary as well as how to assign the infrequent words to entries in it. We use a hard attention mechanism that is directly predicted from FastText features, in which each vocabulary word is mapped to only a single language-agnostic token, as opposed to an interpolation of many. This allows us to avoid both computing a language-specific representation for the uncommon words and aggregating the latent vocabulary features on a per-word basis.

To obtain our latent shared vocabulary in the pretraining step, we learn to embed semantically similar sentences near each other using a triplet loss. More formally, given a triplet of items (x, y^+, y^-) that can be decomposed into a positive pair (x, y^+) and a negative pair (x, y^-), a triplet loss is computed as:

$$L_{triplet}(x, y^+, y^-) = \max(0, m + d(x, y^+) - d(x, y^-)) \tag{1}$$

where $d(x, y)$ is a distance function, and m is a scalar parameter. We use cosine distance for all triplet losses and set $m = 0.05$. Following the methodology of [23,40], we construct minibatches by providing semantically similar sentence pairs as input and consider any non-paired sentence as a negative example. These negatives are randomly sampled from each minibatch. We enumerate all triplets in the minibatch and compute the loss over the top-N most violated constraints, where $N = 10$ in our experiments. Note that these sentences may not come from the same language, so sentences referring to the same image in different languages are also used as positive pairs. To predict which latent embedding we map a source word to, we use sentence representations obtained by feeding FastText embeddings into a fully connected layer. With this mapping, we average the latent embeddings of each word for use in Eq. (1) during the pretraining step, which has been shown to be an efficient, high-performing representation [4,8].

Instead of deterministically mapping to the latent token which achieves the best score, we randomly choose from the top M scoring tokens with probability p, which we refer to as exploration parameters. This helps ensure that spurious mappings are not learned, typically resulting in a 2% performance improvement (see supplementary for a detailed comparison). While we freeze the latent token assignments when training the full model, we allow the features themselves to be fine-tuned. Our experiments use a latent vocabulary size of $40K$ tokens, with exploration parameters $p = 0.2$, $M = 20$. In practice not all latent tokens are used at the end of pretraining; these are dropped when training the full model.

3.2 Masked Cross-Language Modeling (MCLM)

Masked Language Modeling has proven to be useful in training language representations by masking some tokens of an input sentence and then trying to predict the missing tokens [13]. We present a generalization of this approach to a multilingual scenario to encourage stronger cross-language alignment. In MCLM, we assume we have paired sentences across different languages. These sentences need not be direct translations of each other, but, as our experiments will show, they simply need to be semantically related. This is important as

vision-language datasets do not always have paired text queries that are direct translations, but are often independently generated instead (*e.g.* [15,31,27]).

Traditional Masked Language Modeling makes predictions about a single masked token using its surrounding words as context. The words immediately surrounding a token referring to the same entity between sentences in different languages may vary greatly due to differences in grammar. Thus, even using a dictionary between languages to identify word correspondences may not provide useful context. Instead, we use the intuition that semantically similar sentences should contain comparable information across languages, so a sentence in one language could be used as context to predict missing information in another. Conneau *et al.* [10] similarly use masking for improved language alignment. However, our approach does not require parallel data and may sample amongst any of the languages. Lastly, unlike [10] which computes its loss on the predicted word, our objective in Eq. (2) is computed on the fully reconstructed sentences.

More formally, for a pair of languages (i, j), we obtain sentences (S_i, S_j) such that both sentences describe the same image (*i.e.*, they are semantically similar to each other). Then, we randomly replace some portion of their words with a special MASK token to obtain masked representations (S_i^m, S_j^m). These are concatenated together and fed into a fully connected layer that is shared across language pairs to predict the missing information in both sentences $(S_i^{'}, S_j^{'})$. Our MCLM loss then compares this to the unmasked sentences, *i.e.*,

$$L_{mask} = ||\ell_2(S_i^m + S_i^{'}) - \ell_2(S_i)|| + ||\ell_2(S_j^m + S_j^{'}) - \ell_2(S_j)||, \tag{2}$$

where ℓ_2 identifies vectors forced to have unit norm. Both average embedding and LSTM representations are used; details can be found in the supplementary. We compute the masking loss in Eq. (2) for all unique pairs of languages in our experiments, and found masking 20% of the words in the sentences worked best.

3.3 Multilingual Visual-Semantic Alignment

In this section we briefly review the visual-semantic alignment constraints used by MULE [23] that we also employ. First, we use neighborhood constraints [40] that we shall refer to as L_{nc} to encourage similar sentences to embed nearby each other using a triplet loss (*i.e.*, Eq. (1)). Just as with the MCLM module described in Section 3.2, these neighborhood constraints are applied to both the universal language embedding (*i.e.*, the output of the HEM module) as well as the final language representation from the multimodal model as shown in Fig. 2. The second component of the MULE alignment constraint consists of an adversarial language classifier. We shall refer to this classifier loss as L_{adv}, using the approach of [23], whose goal is to ensure that the representations of the different languages in the universal embedding have similar feature distributions. The last component of the MULE constraint is used to train the multimodal model to embed the images and sentences near each other using a triplet loss. This uses a bidirectional triplet loss function, *i.e.*, for image I and paired sentences

(Q^+, Q^-) representing a positive and negative sentence pair, respectively, and sentence Q and its paired images (I^+, I^-), this multimodal loss would be,

$$L_{mm} = L_{triplet}(I, Q^+, Q^-) + \lambda_1 L_{triplet}(Q, I^+, I^-) \tag{3}$$

where λ_1 is a scalar parameter, which we set to 1.5 in our experiments. In addition to using the unmasked sentence representations for the multimodal loss, we observe that most sentences retain their overall semantic meaning if you remove just a few words at random. Using this intuition, we also compute Eq. (3) using the masked sentences (S_i^m, S_j^m) from the MCLM module, which we found provides a small, but consistent improvement to performance. As a reminder, all triplet losses use the implementation details (*e.g.* hyperparameter settings and hard-negative mining) as described in the first part of Section 3. Our total loss function to train SMALR is then,

$$L_{SMALR} = L_{mm} + \lambda_2 L_{mask} + \lambda_3 L_{adv} + \lambda_4 L_{nc} \tag{4}$$

where λ_{2-4} are scalar parameters that we set to (1e-4, 1e-6, 5e-2), respectively.

3.4 Cross-Lingual Consistency

Prior work on multilingual vision-language tasks has primarily focused on how to change training procedures or architectures to support multiple languages, and does not fully take advantage of this multilingual support at test time. In particular, we argue that semantically similar sentences in different languages may capture complementary information, and therefore, considering the predictions made in other languages may improve performance. We validate our intuition by obtaining machine translations of a query in the other languages supported by our model. More formally, suppose we have a set of languages L. Given a query q in language $l_i \in L$, we translate q to all other supported languages in $L \setminus \{l_i\}$ and use this as input to our Cross-Lingual Consistency (CLC) module.

We propose two variants of CLC: CLC-A and CLC-C. CLC-A simply averages matching scores over all languages, and does not require any additional parameters. CLC-C, on the other hand, uses a small Multilayer Perceptron (MLP) to aggregate the scores of each language, which enables us to consider the relative information present in each language's predictions. This MLP has two layers with input size $|L|$ and 32 hidden layer units (*i.e.*, it has 352 learnable parameters) and all parameters are initialized with uniform weight. We train the CLC-C module separately to SMALR using the validation set for 30 iterations. No mini-batches are employed (*i.e.*, it is trained with all image-sentence pairs at once) and it is trained using the multimodal triplet loss described in Eq. (3).

4 Experimental Setup

Datasets. SMALR is evaluated on bidirectional retrieval with Multi30K [6,14,15] and MSCOCO [27,28,31]. The Multi30K dataset is built off of Flickr30K [43],

which originally contained 31,783 images and five English descriptions per image. [6,14,15] obtained annotations in German, French, and Czech, resulting in a four-language dataset. Multi30K contains five descriptions per image in English and German, but only one per image in French and Czech; the latter two were collected as human-generated translations of the English annotations. We use the 29K/1K/1K train/test/val splits from the original dataset [43].

MSCOCO is approximately four times the size of Multi30K, with 123,287 images. There are five human-generated captions per image in English, but significantly fewer in Chinese and Japanese. YJ Captions [31] introduced Japanese annotations for MSCOCO, but only provides five captions per image for a subset of about 26K images. [27] extended MSCOCO with a total of 22,218 Chinese captions for 20,341 images. We use train/test/validation splits as defined in [23].

We augment both datasets with machine translations so every image contains at least five sentences for ten languages: English, German, Czech, French, Chinese, Japanese, Arabic, Afrikaans, Korean, and Russian. All models we compare to are trained using this augmented training set. For languages with no human-generated sentences, we use machine translations at test time as well. We found using translations at test time did not affect the relative performance of different methods in our experiments. See the supplementary for details.

Visual Features. We use ResNet-152 [20] features trained on ImageNet [12] as input to the Embedding Network (EmbN) [40]. As done in [23], we average visual features over ten 448x448 image crops. This generates an image embedding of size 2048, which is then passed through a pair of fully connected layers. The resulting 512-D embedding can be used in the shared image-sentence embedding space. The learning rate was set to $1e^{-3}$ for the HEM and LA models; remaining hyperparameters are consistent with those in [23].

Note that all LIWE [41] experiments use bottom-up Faster R-CNN [35] visual features trained on Visual Genome [25]. This represents a significant increase in the annotation cost compared to our approach, which doesn't use these annotations. Visual Genome also contains MSCOCO [28] images, which means that there is train/test contamination, as LIWE's features are extracted using the pretrained, publicly available model from [2].

Metrics. For our results, we report the mean Recall (mR) across Recall@K, with $K \in [1, 5, 10]$, for both the image-sentence and sentence-image directions per language. All recall values can be found in the supplementary. We also provide an average mR across all languages to serve as a global performance metric: "A" in Tables 1 and 2. The human average, "HA," refers to the average mR over the languages which have human-generated annotations (*i.e.* English, Chinese, and Japanese for MSCOCO, and English, German, French, and Czech for Multi30K).

Comparative Evaluation. We compare the following methods:

- **Frequency Thresholding:** We drop words that occur fewer than t times in the training set. Results are reported in Figure 3.
- **PCA Reduction:** We use PCA [30] to reduce the size of the initial 300-D FastText word embeddings. Results are reported in Figure 3.

- **Dictionary Mapping:** We map words that occur fewer than t times in non-English languages to English using dictionaries [11]. By mapping rare words in other languages to English, some information may be lost, but the token will still exist indirectly in the vocabulary. Results are reported in Figure 3.
- **Language-Agnostic (LA):** We compare to only using a latent vocabulary as described in Section 3.1 with 40K tokens, *i.e.* not using any language specific features. Results are in Tables 1 and 2.
- **HEM:** We evaluate our full hybrid embedding model (Section 3.1), which uses a mix of language-agnostic and language-specific representations. This baseline does not include MCLM nor CLC. Results are in Tables 1 and 2.
- **SMALR:** Our base SMALR is composed of the HEM (Section 3.1) and MCLM (Section 3.2) components of our model. We compare to our complete SMALR which makes use of CLC variants (CLC-A and CLC-C, described in Section 3.4). Results are in Tables 1 and 2.

Note that the first line of Tables 1 and 2, **Trans To En**, refers to using machine translation on non-English sentences, and then using an English-only trained Embedding Network [40], providing a strong baseline method to compare to.

5 Multilingual Image-Sentence Retrieval Results

We provide results for MSCOCO and Multi30K in Table 1 and Table 2, respectively, which contain comparisons to prior work on fewer languages **(a)**, adaptations of prior work to our setting **(b)**, and our model variants **(c)**. SMALR obtains consistent performance gains when evaluating on ten languages over the state-of-the-art (S-LIWE, line 3**(b)**) while also being more efficient than high-performing models like MULE (line 5**(b)**). SMALR outperforms S-LIWE by 11 points on MSCOCO and 5.8 points on Multi30K (line 3**(c)** versus 3**(b)**). A parameter comparison is later shown in Figure 3. SMALR's initial Language-Agnostic (LA) baseline alone is able to boost performance over previous scalable method LIWE by 2-7 points. The HEM, which combines language-agnostic and language-specific embeddings as described in Section 3.1, consistently improves upon the fully language-agnostic vocabulary, even though they share the same latent vocabulary size of 40K tokens. This points to the utility of our hybrid embedding space, which improves performance upon LA by 3.4 average mR on MSCOCO and 2.4 average mR on Multi30K while adding only a few parameters.

When MCLM losses are added, referred to as SMALR in Tables 1 and 2 (line 3**(c)**), mR improves for nearly all languages. This is significant, because we find more compact models like LIWE degrade with additional languages when using the same number of parameters (S-LIWE). The LA baseline is still able to outperform L-LIWE on MSCOCO and Multi30K, in which LIWE learns an embedding five fold larger to try to compensate for the increased number and diversity of languages (120-D instead of 24-D embedding). This suggests that the masking process may help regain some semantic information that is lost when tokens are mapped to the language-agnostic space.

Table 1: MSCOCO multilingual bidirectional retrieval results. (a) contains results from prior work, (b) contains reproductions of two state-of-the art methods evaluated for our scenario using their code, and (c) contains variants of our model

	Model	En	De[1]	Fr[1]	Cs[1]	Cn	Ja	Ar[1]	Af[1]	Ko[1]	Ru[1]	HA	A
(a)	Trans. to En [23]	75.6	–	–	–	72.2	66.1	–	–	–	–	71.3	–
	EmbN [40]	76.8	–	–	–	73.5	73.2	–	–	–	–	74.5	–
	PAR. EmbN [16]	78.3	–	–	–	73.5	76.0	–	–	–	–	75.9	–
	MULE [23]	79.5	–	–	–	74.8	76.3	–	–	–	–	76.9	–
(b)	(1) S-LIWE [41][2]	80.9	–	–	–	–	73.6	–	–	–	–	–	–
	(2) S-LIWE[2]	77.4	–	–	–	–	66.6	–	–	–	–	–	–
	(10) S-LIWE[2]	77.3	67.4	68.5	66.9	64.5	65.8	63.8	66.2	63.1	63.6	69.2	66.7
	(10) L-LIWE[2]	79.1	71.2	70.3	70.1	70.0	69.6	67.5	68.9	66.2	69.6	72.9	70.3
	MULE [23]	79.0	77.2	76.8	77.8	75.6	75.9	77.2	77.8	74.3	77.3	76.8	76.9
(c)	Language-Agnostic	75.0	74.3	74.1	73.4	72.3	72.1	74.4	74.7	71.6	72.7	73.1	73.5
	HEM	78.7	77.3	76.4	77.9	76.7	76.3	77.0	76.7	**75.5**	77.0	77.3	76.9
	SMALR	79.3	**78.4**	**77.8**	**78.6**	76.7	77.2	**77.9**	**78.2**	75.1	**78.0**	77.7	**77.7**
	SMALR-CLC-A	81.2	–	–	–	79.6	75.0	–	–	–	–	78.6	–
	SMALR-CLC-C	**81.5**	–	–	–	**80.1**	**77.5**	–	–	–	–	**79.7**	–

[1]uses translations from English for testing
[2]visual features trained using outside dataset that includes some test images

We next evaluate two CLC variants that use machine translations at test time (described in Section 3.4) on top of SMALR: an average ensemble over all languages (CLC-A), and a weighted ensemble which makes use of a simple classifier (CLC-C). CLC-A uses no additional test-time parameters, and increases the human average performance by 1-3 points, with a larger gain on Multi30K. This may be because more languages can be leveraged on Multi30K (four versus three, compared to MSCOCO). Surprisingly, English performance improves the most amongst CLC-A metrics on Multi30K, demonstrating that certain image-sentence pairs can be better retrieved from the queries in other languages, which may better capture the visual semantics of the same image. CLC-C further improves the human average over CLC-A by 0.9 points on MSCOCO and 0.5 points on Multi30K, using negligible additional parameters.

Parameter reduction method comparison. We present a comparison of baseline vocabulary reduction techniques, described in Section 4, against prior works LIWE and MULE, in addition to our method SMALR (consisting of only HEM and MCLM components in Figure 3). The frequency thresholding and dictionary mapping labels represent the threshold with which we drop infrequent words or map them to English (*e.g.* the blue 50 data point represents dropping words that occur fewer than 50 times). PCA point labels represent the dimensionality we reduce our input vectors to (*e.g.* 300D $\rightarrow$ 50D, 100D, or 200D).

In our comparison of vocabulary reduction methods, frequency thresholding with $t = 50$ and vanilla language-agnostic vocabularies (LA) obtain better per-

Table 2: Multi30K multilingual bidirectional retrieval results. (a) contains results from prior work, (b) contains reproductions of two state-of-the art methods evaluated for our scenario using their code, and (c) contains variants of our model

	Model	En	De	Fr	Cs	Cn[1]	Ja[1]	Ar[1]	Af[1]	Ko[1]	Ru[1]	HA	A
(a)	Trans. to En [23]	71.1	48.5	46.7	46.9	–	–	–	–	–	–	53.3	–
	EmbN [40]	72.0	60.3	54.8	46.3	–	–	–	–	–	–	58.4	–
	PAR. EmbN [16]	69.0	62.6	60.6	54.1	–	–	–	–	–	–	61.6	–
	MULE [23]	70.3	64.1	62.3	57.7	–	–	–	–	–	–	63.6	–
(b)	(1) S-LIWE [41][2]	**76.3**	**72.1**	–	–	–	–	–	–	–	–	–	–
	(2) S-LIWE[2]	75.6	66.1	–	–	–	–	–	–	–	–	–	–
	(10) S-LIWE[2]	75.2	65.2	51.8	50.0	54.1	56.2	62.7	62.8	54.5	63.1	60.6	59.6
	(10) L-LIWE[2]	75.9	66.7	53.3	51.3	56.9	56.3	65.0	63.7	57.1	65.4	61.9	61.2
	MULE [23]	70.7	63.6	63.4	59.4	**64.2**	**67.3**	65.8	67.3	63.6	65.4	64.3	65.1
(c)	Language-Agnostic	65.5	61.3	59.9	54.0	59.4	64.7	63.9	66.5	60.3	60.3	60.2	61.6
	HEM	69.2	62.8	63.3	60.0	62.4	66.3	64.5	66.8	62.3	62.6	63.8	64.0
	SMALR	69.6	64.7	64.5	61.1	64.0	66.7	**66.0**	**67.4**	**64.2**	**65.7**	65.0	**65.4**
	SMALR-CLC-A	74.1	68.9	65.2	64.5	–	–	–	–	–	–	68.2	–
	SMALR-CLC-C	74.5	69.8	**65.9**	**64.8**	–	–	–	–	–	–	**68.7**	–

[1]uses translations from English for testing
[2]visual features trained using outside dataset

formance than both LIWE variants on Multi30K, without adding significantly more parameters, as shown on the right of Figure 3. While more model parameters are needed for MSCOCO, due to the increased vocabulary size, all baselines and prior work MULE significantly outperform LIWE. This demonstrates that more-complex character-based models do not necessarily obtain competitive performance with few parameters when addressing a larger multilingual scenario.

SMALR outperforms all baselines for MSCOCO, as seen on the left of Figure 3, outperforming S-LIWE by over 10 points and using fewer parameters than L-LIWE. We also find that average mean recall performance on MSCOCO is more robust to vocabulary reduction, with a maximum range of about 1.5 average mR between the most extreme reduction and the least. We believe this may be due to the size discrepancy between the two datasets, as MSCOCO is approximately four times the size of Multi30K. PCA reduction appears to have a more linear effect as parameters increase on both datasets. Since Multi30K performance is more sensitive to the number of parameters, it is significant that our SMALR model, in green, (which does not yet make use of our cross-lingual consistency module in Figure 3) outperforms all other models while having less than 20M parameters, 1/5th the parameter count of high performing MULE.

In addition to SMALR outperforming MULE on both datasets while using significantly fewer trainable parameters, we find MULE even fails to outperform simple baselines such as dictionary mapping on MSCOCO. This exposes that

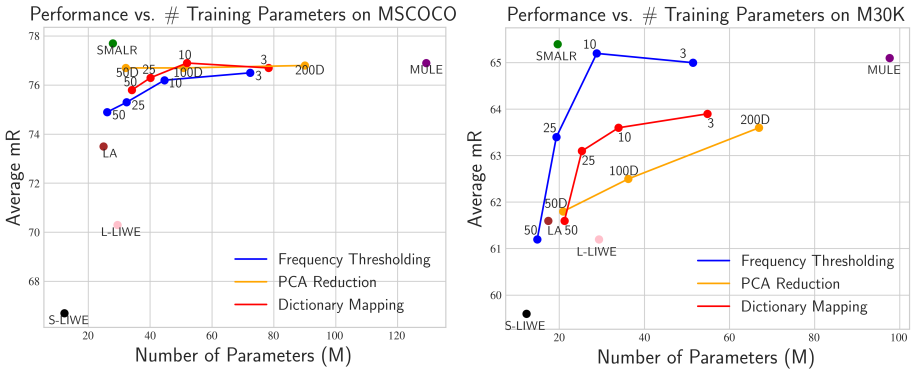


Fig. 3: We compare three types of vocabulary reduction: frequency thresholding, PCA dimensionality reduction, and mapping rare words to English with the use of dictionaries. The left-hand side evaluates on MSCOCO, the right on Multi30K. We have additional standalone points for the small LIWE (S-LIWE), large LIWE (L-LIWE), MULE, language agnostic vocabulary (LA), and our model, SMALR

the large number of parameters used in MULE are unnecessary for performance gains. While SMALR uses more parameters during training than S-LIWE, we have far fewer test-time parameters. We reduce the computation needed for evaluation by using precomputed language representations from training. This reduces the entire SMALR model to the image-sentence matching model with our CLC add-on, totaling only 7.1M parameters, now fewer than S-LIWE.

6 Conclusion

We have presented a Scalable Multilingual Aligned Representation (SMALR), which addresses the trade-off between multilingual model size and downstream vision-language task performance. Our approach is modular, and thus can be used as a drop-in language representation for any vision-language method/task. SMALR outperforms all prior work on the task of multilingual image-sentence retrieval on average across ten diverse languages, with the use of a hybrid embedding model, masked cross-language modeling loss, and cross-lingual consistency module. Our hybrid embedding model significantly reduces the input to a language model by mapping most tokens to a fixed size, shared vocabulary. The masking procedure aligns our diverse set of languages and uses the multimodal model to provide additional alignment with visual grounding. We find that both cross-lingual consistency modules better aggregates retrieved results, boosting performance with minimal additional parameters. This is all accomplished with less than 20M trainable parameters, significantly reducing oversized prior work by 1/5th, while improving performance over the state-of-the-art by 3-4%.

Acknowledgements This work is funded in part by the NSF, DARPA LwLL, and DARPA XAI grants, including NSF grant 1838193.

References

1. Aharoni, R., Johnson, M., Firat, O.: Massively multilingual neural machine translation. 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) (Jun 2019)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
4. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations (ICLR) (2017)
5. Artetxe, M., Labaka, G., Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 2289–2294 (2016)
6. Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., Frank, S.: Findings of the third shared task on multimodal machine translation. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 304–323 (2018)
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics (TACL) **5**, 135–146 (2017)
8. Burns, A., Tan, R., Saenko, K., Sclaroff, S., Plummer, B.A.: Language features matter: Effective language representations for vision-language tasks. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
9. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Empirical Methods in Natural Language Processing (EMNLP) (2014)
10. Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
11. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: International Conference on Learning Representations (ICLR) (2018)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: arXiv:1810.04805v1 (2018)
14. Elliott, D., Frank, S., Barrault, L., Bougares, F., Specia, L.: Findings of the second shared task on multimodal machine translation and multilingual image description. arXiv:1710.07177 (2017)
15. Elliott, D., Frank, S., Sima'an, K., Specia, L.: Multi30k: Multilingual english-german image descriptions. arXiv:1605.00459 (2016)
16. Gella, S., Sennrich, R., Keller, F., Lapata, M.: Image pivoting for learning multilingual multimodal representations. In: Empirical Methods in Natural Language Processing (EMNLP) (2017)

17. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
18. Gu, J., Hassan, H., Devlin, J., Li, V.O.: Universal neural machine translation for extremely low resource languages. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT) (2018)
19. Gupta, T., Schwing, A., Hoiem, D.: Vico: Word embeddings from visual co-occurrences. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv:1512.03385 (2015)
21. K, K., Wang, Z., Mayhew, S., Roth, D.: Cross-lingual ability of multilingual bert: An empirical study. arXiv:1912.07840 (2019)
22. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Empirical Methods in Natural Language Processing (EMNLP) (2014)
23. Kim, D., Saito, K., Saenko, K., Sclaroff, S., Plummer, B.A.: Mule: Multimodal universal language embedding. In: AAAI Conference on Artificial Intelligence (2020)
24. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
25. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV) (2017)
26. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv:1908.03557 (2019)
27. Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G., Xu, J.: Coco-cn for cross-lingual image tagging, captioning and retrieval. IEEE Transactions on Multimedia (2019)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: The European Conference on Computer Vision (ECCV) (2014)
29. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv:1908.02265 (2019)
30. Maćkiewicz, A., Ratajczak, W.: Principal components analysis (pca). Computers and Geosciences **19**(3), 303 – 342 (1993)
31. Miyazaki, T., Shimizu, N.: Cross-lingual image caption generation. In: Conference of the Association for Computational Linguistics (ACL) (2016)
32. Nguyen, D.K., Okatani, T.: Multi-task learning of hierarchical vision-language representation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
33. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual bert? arXiv:1906.01502 (2019)
34. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: The IEEE International Conference on Computer Vision (ICCV) (2015)

35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2015)
36. Smith, S.L., Turban, D.H.P., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv:1702.03859 (2017)
37. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. arXiv:1908.08530 (2019)
38. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2019)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017)
40. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **41**(2), 394–407 (2018)
41. Wehrmann, J., Souza, D.M., Lopes, M.A., Barros, R.C.: Language-agnostic visual-semantic embeddings. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
42. Wu, S., Dredze, M.: Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. arXiv:1904.09077 (2019)
43. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (TACL) **2**, 67–78 (2014)
44. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)