Language-Agnostic Visual-Semantic Embeddings

Jônatas Wehrmann, Douglas M. Souza, Mauricio A. Lopes, Rodrigo C. Barros
School of Technology
Pontifícia Universidade Católica do Rio Grande do Sul
{jonatas.wehrmann@edu,rodrigo.barros@}pucrs.br

Abstract

This paper proposes a framework for training language-invariant cross-modal retrieval models. We also introduce a novel character-based word-embedding approach, allowing the model to project similar words across languages into the same word-embedding space. In addition, by performing cross-modal retrieval at the character level, the storage requirements for a text encoder decrease substantially, allowing for lighter and more scalable retrieval architectures. The proposed language-invariant textual encoder based on characters is virtually unaffected in terms of storage requirements when novel languages are added to the system. Our contributions include new methods for building character-level-based word-embeddings, an improved loss function, and a novel cross-language alignment module that not only makes the architecture language-invariant, but also presents better predictive performance. We show that our models outperform the current state-of-the-art in both single and multi-language scenarios. This work can be seen as the basis of a new path on retrieval research, now allowing for the effective use of captions in multiple-language scenarios. Code is available at `https://github.com/jwehrmann/lavse`.

1. Introduction

This paper addresses the problem of cross-modal retrieval. The task consists in retrieving content from one modality given a query on a different modality, e.g., returning an image based on a textual description. Several important applications benefit from successful retrieval strategies, such as image and video retrieval, captioning [32, 37], and navigation for the blind, just to name a few.

One of the contributions of this paper is incorporating an important feature towards robustness over different retrieval domains: language-invariant behavior. Besides making the task language-invariant, we also propose a versatile strategy that relies solely on character-level learning of word embeddings. This means that our embedding approach is virtually not affected in terms of storage requirements when adding new languages for the retrieval task. Also, our method can be extended to learn novel languages without requiring extra machine translation models that are much more costly in terms of processing. For such, we present a novel training procedure that performs both cross-modal and cross-language alignment by enforcing similar sentences to have high-similarity in the embedding space, while projecting correlated image-caption pairs into the same space.

Our contributions also include better image and text encoding functions to explicitly leverage inner-attention maps, which allow for better semantic encoding of both modalities. We show that the use of region-based non-linear non-local modules provide a large improvement in predictive performance, capable of outperforming state-of-the-art approaches based on stacked attention layers [22]. We also provide experiments training the text encoder with distinct granularity of the text being learned. For instance, current state-of-the-art approaches [36, 7, 14, 38, 22] are based on networks trained over word-embeddings [25], whereas our proposed method can be trained in an end-to-end fashion for learning both word-level and character-level features from scratch without any preprocessing for the text encoder. More specifically, raw characters are mapped to a word-latent space that is learned during training, which allows the resulting model to project words from distinct languages into the very same word-based embedding space.

We perform a thorough set of experiments to evaluate multiple aspects of the proposed architecture. In summary, our contributions are as follows: (i) novel character-based word-embedding methods; (ii) a cross-language, cross-modal retrieval framework; (iii) an improved pairwise ranking loss function that enables training of word and character-level models in multilingual scenarios; (iv) an improved image representation strategy that maps object representations into the shared semantic space, discarding cross-modality attention layers, and (v) we provide a transliterated version of YJ Captions dataset with novel retrieval splits. The proposed approach outperforms the state-of-the-art methods in both image retrieval and image annotation tasks, while performing much faster when compared to the best baseline strategy.

2. Cross-Language Multimodal Retrieval

We propose a method for training language-invariant word embeddings that can be used for the retrieval of images and their respective captions written in multiple languages, namely CLMR. Formally, consider a set of images $\mathcal{X} = \{x^{(i)}\}_{i=1}^{|X|}$ and their respective captions $\mathcal{C} = \{c^{(i)}\}_{i=1}^{|C|}$. Additionally, let $\mathcal{L} = \{l_i\}_{i=1}^{|L|}$ be a set of languages, and $\mathcal{T}_{l_i} = \{t^{(i)}\}_{i=1}^{|T|}$ be a set of sentences from language l_i. Captions $c_{l_1}^{(i)}$ and $c_{l_2}^{(i)}$ present the same semantic content despite being written in distinct languages. Likewise, $t_{l_1}^{(i)}$ and $t_{l_2}^{(i)}$ are generic sentences containing the same semantics though in distinct languages.

Our approach for learning the cross-modal space follows the state-of-the-art methods [18, 10, 36, 39], in which two functions must be approximated, namely $\varphi(c_{l_i})$ and $\phi(x)$ in order to project images $\mathcal{X}$ and their respective captions $\mathcal{C}$ into the same latent space. Therefore, $\varphi(c_{l_i}) \in \mathbb{R}^d$ and $\phi(x) \in \mathbb{R}^d$ can be seen as feature vectors that represent the semantic content of captions and images in a shared d-dimensional space, in which correlated image-caption pairs become close to each other, and the distance of non-correlated pairs should necessarily be larger than the correlated ones. Therefore, we want to approximate both vectors so that a similarity measure $s(\phi(x), \varphi(c_{l_i})) \approx 1$.

Given that our goal is to train language-invariant cross-modal embedding models, the choice for function $\varphi(\cdot)$ is particularly important. Such a function should be capable of learning semantic textual information across distinct languages, which often requires a very large vocabulary. We ensure that by using the same similarity measure that is used for approximating images and captions, we can approximate two distinct sentences written in different languages though with the same semantics into the same joint embedded space. Therefore, we also want that $s(\varphi(t_{l_1}^{(i)}), \varphi(t_{l_2}^{(i)})) \approx 1$. The overall architecture of CLMR can be seen in Figure 1.

2.1. Text Encoder

Regarding the text encoding function $\varphi(.)$, it should ideally be capable of approximating high-level semantic concepts from images and captions while learning correlations between sentences across distinct languages.

Recent studies have mostly focused on encoding image captions through GRU [5] networks, handcrafted transformations over word-embeddings, or character-level convolutional networks. Most of those strategies encode text in a global manner by projecting them onto a high-dimensional semantic embedding. On the other hand, a recent state-of-the-art approach [22] makes use of the hidden states of the GRU network for computing a cross-modal attention between image regions and captions, a strategy that makes the test phase much slower when compared to global embedding methods. In our approach, word vectors are fed into a bidirectional GRU generating $|c^{(i)}|$ d-dimensional hidden-states for each direction. Those word vectors can be either traditional ones, as those from CLMR; or character-based generated, as those from LIWE. Following [22], the final representation is generated by averaging the textual representation from each direction.

2.2. Character-based Word Embeddings

Strategies based on word-embeddings and RNNs, or on handcrafted transformations for encoding sentences present many significant drawbacks: (i) they require training word embeddings [28, 25] and RNNs [20] in very large corpora (with millions or billions of words), consuming a lot of time and demanding high computational power; (ii) for encoding a single word or sentence, it is necessary to have at disposal the entire word-dictionary containing all the known words, largely increasing the memory requirements to store all data; (iii) for cross-language or informal domains, the number of words in the dictionary increases with the number of languages; (iv) a preprocessing step is required for correcting typos and standardizing words.

Bearing in mind the advantages of both character-encoding and word-embedding approaches, we have designed a novel strategy for the representation of the input captions that tries to leverage the advantages of both while avoiding their drawbacks. This strategy, hereby called LIWE (Language-Invariant Word Embeddings) learns to generate word embeddings from character-level inputs, which can be further processed by either GRUs or convolutional layers. Unlike previous work [29, 17, 16] that generate word embeddings using characters or similar sub-word information together with RNNs, our approach is simple to implement and allows for fast word-embedding computation.

Word-embedding vectors are typically generated by either pre-training them on a separate large corpus, or fully-training all word vectors via back-propagated gradients during the training of the target task. In both strategies, one must have at hand all known words and store them within a vocabulary $\mathcal{V} = \{w^{(i)}\}_{i=1}^{|\mathcal{V}|}$ so they can be retrieved at training and test time. Let $\varrho(i)$ be the function for retrieving the i^{th} word-embedding vector. Such a function is often implemented using one of two main approaches: (i) a binary vector $\mathbf{w} \in \{0,1\}^{|V|\times 1}$ so that $\mathbf{w}_i = 1$ and $\sum_j^{|V|} \mathbf{w}_j = 1$, which is then multiplied by the embedding weight matrix $\Omega_E \in \mathbb{R}^{|V|\times|\omega|}$, making $\varrho(i) = \mathbf{w}_i^T \Omega_E$; and (ii) $\varrho(i)$ is encapsulated as a look-up table function so that $\varrho(i) = \Omega_{E_i}$.

In LIWE, the $\varrho(i)$ function is implemented following a different strategy. We use the word's atomic components $w_j = \{a^{(i)}\}_{i=1}^{|w_j|}$, where $a^{(i)}$ is the i^{th} character token within word w_j. That character is represented as a dense vector, $\alpha \in \mathbb{R}^{24}$, so that $\varrho(i)$ can be implemented as a function that processes character vectors from each word independently,

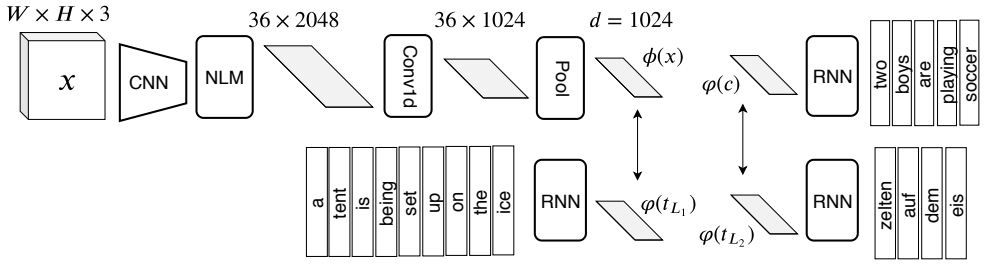

Figure 1. Overall architecture of CLMR.

and that ultimately results in vector $\mathbb{R}^{|w|}$. It is important for $\varrho(i)$ to be computed fast since it needs to be computed for each word for all captions $\in \mathcal{C}$ and sentences $\in \mathcal{T}$.

LIWE is computed by simply concatenating character embeddings instead of using convolutional layers applied over the character-level inputs, with the goal of optimizing the word-embedding space. Additionally, we project the concatenation of character-based vectors by using at least one batch-wise fully-connected layer. Here, the concatenation of the character-level vectors already works as a low-level word-embedding. Indeed, the first fully-connected layer applied over this input learns from the entire input at once (i.e., achieving the maximum receptive field over the input) projecting the low-level embedding into a higher-level embedding — which, in turn, achieves a higher degree of detachment from the syntax in favor of the semantics.

Let $C_t \times d_c$ be a matrix that encodes a sequence of character-level embeddings, and $d_c = 24$. That sequence is split into words, and then concatenated to build a primitive word-level representation of size $N_w \times C_w \times 24$, where N_w is the number of words in a given text, C_w is the number of characters in each word, and 24 is the size of the character embedding. In this strategy, the number of characters in each word is crucial, since we concatenate them to feed a fully-connected layer that does not accept variable-sized inputs. For handling this issue, we pad the words with a special token so all the words comprise the same number of characters. The fixed number of characters is based on the statistics of the word lengths within the corpora. We employ up to three fully-connected layers to project the padded $N_w \times C_w \times 24$ tensor into a $N_w \times D_w$ matrix that is ready to be processed by our text-encoding model.

LIWE is thus designed to replace the traditional word-embedding matrix by a learnable function ϱ that approximates the behavior of those embeddings without requiring the storage of many thousands of word-vectors. For instance, assume a LIWE incarnation of LIWE(128,256), that encodes character-level vectors $\in \mathbb{R}^{24}$ through fully-connected layers containing respectively $f \in \{128, 256\}$ neurons. The complexity in terms of parameters required for learning information from all words in a given vocabulary is given by: $(N_a \times d_c) + ([d_c \times C_w] \times f_1) + (f_1 \times f_2)$, where N_a is the number of characters in the alphabet, resulting in just $\approx$ 115k parameters. Finally, the corpus from Flickr30k and Multi30k together (Flickr30K translated to German) comprises roughly 20,000 words, requiring $20,000 \times 300 = 6,000,000$ parameters when using word-embeddings $\in \mathbb{R}^{300}$, and whose memory requirement is about 50-fold larger than LIWE(128,256). Given that the difference increases linearly, a corpus of $40,000$ words would be enough for that configuration of LIWE to run with two-orders of magnitude fewer parameters for embedding words.

2.3. Self-Attentive Image Encoder

The image encoding function $\phi(x)$ encapsulates three main steps: (i) a forward pass of an object detector network (Faster R-CNN [31]) trained on the Visual Genome dataset [21] for extracting the k most important regions within the image, which is inspired by [1, 22]; (ii) a region-based mapping into the cross-modal space that is weighted using a non-linear NLM module; and (iii) a one-dimensional convolutional layer to project regions into the shared space, followed by a global average pooling that generates the final vector representation of the original image. The last two steps substantially differ from the baseline approaches [18, 10], given that we project object-based features onto the semantic space instead of projecting the image feature vector generated by the last pooling layer within a given convolutional network. It is somewhat similar to [22], the difference being that we compute an inner attention map through the NLM module, in which all regions are used to compute the attention weights.

A generic NLM is denoted by:

$$NLM(\mathbf{x}) = \sigma\Big(\textsc{softmax}\left(q(\mathbf{x})^T k(\mathbf{x})\right) v(\mathbf{x})\Big) \quad (1)$$

which can be applied to map long-range spatial dependencies when applied to the regions of input $\mathbf{x}$. This module is particularly effective for mapping global relations once it learns a similarity function that compares value $\mathbf{x}_i$ to all remaining positions $\mathbf{x}_j$, which results in an affinity scalar value. In this NLM incarnation, the affinity scalar is given by a matrix multiplication between embeddings resulting from

$q(\cdot)$ and $k(\cdot)$ functions. $q(\cdot)$, $k(\cdot)$, and $v(\cdot)$ are implemented as one-dimensional convolutional layers, that reduce input dimensionality 8-fold. Different from [35], in this work the resulting weighted feature map is processed by a non-linearity function $\sigma(\mathbf{x}) = \max\{0, \mathbf{x}\} + 0.1 \times \min\{0, \mathbf{x}\}$.

Here, the NLM module learns relations across all regions in order to project them in a weighted fashion onto the cross-modal space. Therefore, the network is able to give more weight to important regions and words, while not being required to compute stacked attention layers across distinct modalities. This is important because the computation of stacked attention is rather slow in both training and test times. Instead, our models can leverage global information to project image features in a more representative semantic vector space, an approach that allows for fast search using very efficient matrix multiplication functions.

2.4. Loss Function

State-of-the-art retrieval frameworks employ the pairwise ranking loss as the objective function to compute $\phi(\cdot)$ and $\varphi(\cdot)$ gradients. The pairwise ranking loss pushes away instances with small violations from the query and approximate matching instances maintaining a minimum margin on the joint embedded space. The default incarnation does that by summing the computed similarities between the query and contrastive examples. This approach may suffer from small-violating negatives domination over hard contrastives [10]. Hard contrastives are those negative examples whose similarity to the query example is the largest with exception to the positive (matching) example. For a specific query, when the returned examples contain several negatives with small violations, a single negative example too close to the query might not be sufficiently taken into account. In that scenario, to move the hard contrastive away, such mapping might require an update step that would bring back the small-violating negatives, creating local minima in which the model might get trapped into.

A pairwise ranking loss based on hard contrastives – e.g., *Max of Hinges* loss – has proved to be more suitable for the ranking task. The drawback of such an approach is that it optimizes the loss function based on a single hard negative example for each query. Since we are trying to learn character-level embeddings from scratch, it becomes unfeasible for the optimization process to learn such deep layer representations from a single random value in the beginning of training. We overcome this issue by introducing a novel loss function that increases exponentially the relevance of the hard contrastives over time, as follows:

$$\mathcal{J} = \lambda(\epsilon) \cdot \mathcal{J}_m + (1 - \lambda(\epsilon)) \cdot \mathcal{J}_s \tag{2}$$

$$\lambda = 1 - \eta^{\epsilon} \tag{3}$$

where λ is the trade-off weight and ϵ is the number of iterations. The *Sum of Hinges* and *Max of Hinges* for the cross-modal alignment are given by:

$$\mathcal{J}_{M_s}(x, c_{l_1}) = \sum_{c'_{l_1}} [\alpha - s(\phi(x), \varphi(c_{l_1})) + s(\phi(x), \varphi(c'_{l_1}))]_+ + \sum_{x'} [\alpha - s(\varphi(c_{l_1}), \phi(x)) + s(\phi(c_{l_1}), \phi(x'))]_+ \tag{4}$$

$$\mathcal{J}_{M_m}(x, c_{l_1}) = \max_{c'_{l_1}} [\alpha - s(\phi(x), \varphi(c_{l_1})) + s(\phi(x), \varphi(c'_{l_1}))]_+ + \max_{x'} [\alpha - s(\varphi(c_{l_1}), \phi(x)) + s(\phi(c_{l_1}), \phi(x'))]_+ \tag{5}$$

where c_{l_1} is image x's description on the main language l_1. c'_{l_1} and x' denote the negative examples for the image and description queries, respectively. $s(x_i, x_j)$ is the computed similarity between x_i and x_j. To compute $s(x_i, x_j)$ we first scale x_i and x_j to have unit norm, so the inner product of both results become the cosine similarity.

Since we are also dealing with cross-language alignment, we denote the cross-language loss functions as:

$$\mathcal{J}_{L_s}(t_{l_i}, t_{t_j}) = \sum_{t'_{l_i}} [\alpha - s(\phi(t_{l_j}), \varphi(t_{l_i})) + s(\phi(t_{l_j}), \varphi(t'_{l_i}))] + \sum_{t'_{l_j}} [\alpha - s(\varphi(t_{l_i}), \phi(t_{l_j})) + s(\phi(t_{l_i}), \phi(t'_{l_j}))] \tag{6}$$

$$\mathcal{J}_{L_m}(t_{l_i}, t_{t_j}) = \max_{t'_{l_i}} [\alpha - s(\phi(t_{l_j}), \varphi(t_{l_i})) + s(\phi(t_{l_j}), \varphi(t'_{l_i}))] + \max_{t'_{l_j}} [\alpha - s(\varphi(t_{l_i}), \phi(t_{l_j})) + s(\phi(t_{l_i}), \phi(t'_{l_i}))] \tag{7}$$

where t_{l_1} and t_{l_2} denotes two semantically aligned sentences from two different languages. Note that t_{l_1} and t_{l_2} have no semantic relationship with the image captions from the cross-modal retrieval task and can be obtained from a completely different corpus.

The final loss function to optimize the multimodal cross-lingual latent space is given by

$$\min_{W} \mathcal{J}_M(x, c_{l_1}) + \frac{1}{|\mathcal{L}|} \sum_{j}^{|\mathcal{L}|} \mathcal{J}_L(t_{l_1}, t_{t_j}) \tag{8}$$

3. Experimental Setup

3.1. Datasets

We have performed several experiments using four large-scale datasets for cross-modal retrieval, namely MS COCO [23], Flickr30k [30], its multi-language version Multi30k [8], and YJ Captions 26K Dataset [26], the latter comprising captions in Japanese language for a subset of

Table 1. Cross-modal results on COCO test set. Cross-modal results on Flickr30k test set. Underlined values outperform the best published results. Bold values indicate current state-of-the-art results.

Method	Image to text			Text to image			
	R@1	R@5	R@10	R@1	R@5	R@10	Σ
Order [36]	49.3	78.5	89.4	39.5	75.0	86.2	417.9
CHAIN [39]	61.2	89.3	95.8	46.6	81.9	90.9	465.7
VSE++ [9]	64.6	-	95.7	52.0	-	92.0	-
DPC [43]	65.6	89.8	95.5	47.1	79.9	90.0	467.9
GXN [13]	68.5	-	97.9	56.6	-	94.5	-
SCO [15]	69.9	92.9	97.5	56.7	87.5	**94.8**	499.3
SCAN-t2i-avg [22]	70.9	94.5	97.8	56.4	87.0	93.9	500.5
SCAN-i2t-lse [22]	69.2	93.2	97.5	54.4	86.0	93.6	493.9
VSE++*	67.5	93.7	96.8	53.4	84.9	92.4	488.8
LIWE	69.6	93.9	98.0	55.5	87.3	94.2	498.6
CLMR	71.8	93.1	97.6	56.2	87.5	94.2	500.3
LIWE(+Glove)	**73.2**	**95.5**	**98.2**	**57.9**	**88.3**	94.5	**507.7**

Table 2. Cross-modal results on Flickr30k test set. Underlined values outperform the best published results. Bold values indicate current state-of-the-art results.

Method	Image to text			Text to image			
	R@1	R@5	R@10	R@1	R@5	R@10	Σ
VSE++ [9]	52.9		87.2	39.6		79.5	
DAN [27]	55.0	81.8	89.0	39.4	69.2	79.1	413.5
DPC [43]	55.6	81.9	89.5	39.1	69.2	80.9	416.2
SCO [15]	55.5	82.0	89.3	41.1	70.5	80.1	418.5
SCAN-i2t-avg [22]	67.9	89.0	94.4	43.9	74.2	82.8	452.2
SCAN-t2i-avg [22]	61.8	87.5	93.7	45.8	74.4	83.0	446.2
VSE++*	56.9	83.2	88.6	41.0	70.5	79.5	419.7
CLMR	64.0	88.3	93.3	46.8	76.4	84.5	453.2
LIWE	66.4	88.9	94.1	47.5	76.2	84.9	458.1
LIWE(+Glove)	**69.6**	**90.3**	**95.6**	**51.2**	**80.4**	**87.2**	**474.3**

26k images from COCO. COCO is largely used for training and evaluating systems for image-caption alignment, and it has become the standard benchmark to evaluate the predictive performance of state-of-the-art methods. It comprises 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. Flickr30k comprehends roughly 28,000 images for training and 1,000 for both validation and testing. Each image has 5 corresponding textual descriptions. Multi30k was originally developed for training visually-guided machine translation [8] models, though we use it as a multi-lingual corpus since it has German captions for each Flickr image. Finally, the YJ Captions dataset also comprises roughly 5 captions per image, which results in a corpus of $\approx$ 130k Japanese image descriptions aligned to English ones. Given that the original work is focused on the task of image captioning and does not provide splits for image retrieval evaluation, we generate validation and test splits by randomly sampling $\approx$ 1k images for each split while keeping the remaining of the images on the training set. A final remark regarding YJ Captions is that we perform experiments that make use of the transliterated version of the dataset, which allows training character-based word-embedding models.

3.2. Evaluation Measures

For evaluating the results, we use the same measures as those in [18, 36, 10]: $R@K$ (reads "Recall at K"), which is the percentage of queries in which the ground-truth is one of the first K retrieved results. The higher its value, the better.

4. Experimental Analysis

We first analyze the predictive performance of our models on COCO and Flickr30k datasets trained specifically for English captions. Our second analysis is regarding the results generated by using our framework for multiple-language cross-modal retrieval, where we can understand the impact of training cross-modal and multiple-language models altogether. Our models trained with the improved non-local image and text-encoding functions along with the proposed loss function are denoted as CLMR. Character-based word embedding models are depicted as LIWE. Hyper-parameters and training details are reported in the supplementary material.

4.1. Single-Language Results

In this section we present the results for both COCO and Flickr30k test sets in the English language. We first compare our methods trained only with the cross-modal retrieval loss function J_m with the state-of-the-art approaches. Results in Table 1 show that our methods perform on par to state-of-the-art approaches such as SCAN [22] despite being up to four-fold faster to train, and up to one order of magnitude faster in the test phase (depending on the number of instances to retrieve). Such difference in running time is due

Table 3. Single language cross-modal results on Multi30k and YJ Captions test sets.

Method	Image to text R@1	Image to text R@10	Text to image R@1	Text to image R@10	Text to image R@1	Text to image R@10	Σ
			MULTI30K				
VSE++	47.2	77.0	86.5	33.7	61.1	71.7	343.6
SCAN-t2i	44.5	76.8	86.4	35.7	60.9	71.0	339.6
SCAN-i2t	51.8	82.0	91.0	32.7	61.7	72.2	358.7
CLMR	51.6	79.7	88.9	34.5	63.5	73.6	357.3
LIWE	**59.9**	**87.5**	**93.7**	**42.3**	**71.1**	**79.8**	**392.0**
			YJ CAPTIONS				
VSE++	54.0	82.1	90.9	43.2	76.5	86.5	433.2
SCAN-i2t	51.2	83.0	91.8	39.8	74.6	85.8	426.2
SCAN-t2i	56.5	85.7	93.0	42.5	73.6	83.4	434.6
CLMR[Ours]	**57.4**	85.3	94.0	45.1	**80.1**	**89.6**	**451.4**
LIWE[Ours]	56.9	**86.1**	**94.1**	**45.1**	78.0	88.2	448.4

to the cross-attention mechanism used within SCAN, while in our approaches we use non-local inner-attention modules for building better vector representations. One can observe that CLMR outperforms all other approaches in $R@1$ for both image-to-text (71.8%) and text-to-image (57.9%). Models trained with LIWE, despite fully replacing word-embeddings by much more memory-efficient learned function, also present solid performance for all tasks and metrics, performing only slightly below its word-embedding competitor, namely CLMR. The best performing method on COCO is LIWE trained with character-based word-embeddings concatenated to pretrained Glove vectors, leading to 73.2% R@1 and 57.8% for image-to-text and text-to-image respectively, an absolute improvement of 7.7% when considering the recall sum (Σ).

Results on Flickr30k depicted in Table 2 show that LIWE presents the overall top score with a margin for the text-to-image retrieval, i.e., 47.5% of $R@1$, an absolute improvement of 1.7% when compared to SCAN-t2i-avg, though outperforming it by 4.6% in image-to-text R@1 metric. In addition, LIWE performs 3.6% higher on R@1 when compared to SCAN-i2t-avg. Once again, the use of character-based word-embedding methods presents state-of-the-art results, and CLMR is superior than all the baselines for image retrieval tasks considering all the metrics. This clearly shows that LIWE is quite effective despite approximating word-embeddings via a learned function over the input characters. In addition, we see more evidence that LIWE can be complemented with Glove pretrained word-embeddings. In this case, it seems to be quite effective given that Flickr30k is a medium-sized dataset, and using both approaches altogether can help avoiding overfitting. Hence, LIWE(+Glove) outperforms all the state-of-the-art approaches by large margins for most metrics ($\approx 12\%$ of relative improvement in text-to-image R@1).

Table 3 depicts results on Multi30k dataset, i.e., the German version of Flickr30k. For providing fair comparison with state-of-the-art approaches, we trained SCAN and VSE++ models using our loss function. In this experiment we observe that LIWE is able to outperform all other approaches, with significant margins. CLMR performed similar to the SCAN-i2t approach, though surpassing VSE++ and SCAN-t2i by a margin.

Results on YJ Captions are shown in Table 3. Recall that this dataset comprises roughly 30 thousand images from the MS COCO training set, aligned to Japanese captions, that we have transliterated for allowing training LIWE models. The best performing method are CLMR, LIWEand SCAN-t2i, respectively. VSE++ also provides strong performance, specially for text-to-image.

4.2. Cross-language Results

In this section, we report cross-language experiments as to evaluate the performance of all models for learning language-independent representations. In this case, we use to complete formulation of CLMR, optimizing the complete loss function $\mathcal{J}$ (Equation 2). Once again, all models were trained using the same loss function. In this case we also add a strong BERT-Multilingual baseline, which comprises 12-layers and $\approx 110M$ parameters. We use activation values from its last layer and use them as fixed word-embedding vectors, that are processed by a BiGRU layer. Table 4 shows results of bilingual models, trained for approximating images to English captions, while also approximating aligned English-German sentences from Multi30k. Note that LIWE is able to outperform all baselines in all metrics. BERT-Multilingual presents quite a strong performance, surpassing values from the baselines and CLMR. Though, note that it is much costly in terms of memory, parameters and running time.

Table 5 shows results for models trained in English and Japanese languages. LIWE shows strong performance (R@1 59.2% for image-to-text), closely followed by SCAN-t2i and CLMR. This is quite a notable result, specially when we consider how Japanese and English languages are structurally different. Nevertheless, LIWE was able to learn good representations and top state-of-the-art approaches.

Table 4. Cross-modal results on Multi30k German test set by co-training multi-language sentence embeddings.

Method	Image to text (ENGLISH)		Text to image (ENGLISH)		Image to text (GERMAN)		Text to image (GERMAN)	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
VSE++	58.9	91.8	43.9	81.5	49.8	84.6	33.6	70.6
SCAN-t2i	59.4	93.4	45.0	83.7	42.2	82.4	27.9	66.5
SCAN-i2t	58.9	91.8	37.1	79.3	44.4	83.6	26.0	65.5
BERT-Multilingual	62.0	92.1	42.7	82.5	50.9	86.4	33.2	73.5
CLMR	59.9	92.8	43.9	84.3	50.4	86.8	34.6	73.1
LIWE	**64.4**	**94.1**	**47.5**	**85.4**	**53.0**	**89.1**	**36.7**	**76.8**

Table 5. Cross-modal results on YJ Captions Transliterated test set by co-training multi-language sentence embeddings.

Method	Image to text (ENGLISH)		Text to image (ENGLISH)		Image to text (JAPANESE)		Text to image (JAPANESE)	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
VSE++	54.6	92.7	42.5	87.8	49.4	88.5	38.9	84.5
SCAN-i2t	52.9	92.7	36.4	84.8	42.7	86.9	28.5	79.8
SCAN-t2i	58.2	94.0	**47.4**	90.3	48.2	89.4	**39.6**	85.4
CLMR	56.9	92.9	43.2	89.1	**51.4**	**91.7**	38.6	**87.3**
LIWE	**59.2**	**94.7**	46.1	**90.4**	48.6	90.6	37.0	85.6

Table 6. Ablation study: cross-modal results on Flickr30k.

Method	Image to text		Text to image	
	R@1	R@10	R@1	R@10
LIWE (Complete)	66.4	92.3	47.5	84.6
LIWE (Without NLM)	61.8	92.3	44.6	82.2
LIWE (Linear NLM)	60.7	93.5	44.7	83.3
LIWE (NLM WR)	56.3	91.3	41.7	79.1
CMLR[$\mathcal{J}_{M_S}$ [19]]	60.4	92.2	43.8	83.3
CMLR[$\mathcal{J}_{M_M}$) [10]]	Diverges	-	-	-
CMLR[Ours, $\epsilon = 0.991$]	65.8	93.1	47.3	84.2
LIWE(256,256)[$\mathcal{J}_{M_S}$ [19]]	63.2	93.0	46.6	84.8
LIWE(256,256)[$\mathcal{J}_{M_M}$ [10]]	Diverges	-	-	-
LIWE(256,256)[Ours, $\epsilon = 0.991$]	65.3	92.5	47.4	84.6

4.3. Ablation Study on Flickr30k

In Table 6, we show the importance of each component within CLMR. First, we observe that the complete method in its default incarnation, denoted by CLMR(Complete) or simply CLMR, presents the best overall performance. It is also clear that the application of the Non-Local Module is quite important given that we can fully discard cross-modal attention strategies. In addition, we see that it is quite important to reduce the input dimnensionality to use it as query of the NLM module (NLM WR is trained without reducing the inputs). Finally, results show that the proposed loss function outperform those from [19], and present themselves as more stable option than [10].

Figure 2 depicts our proposed approaches along with the approaches VSE++ and SCAN-t2i. It is quite clear that the optimization of the proposed loss function leads to much better results. In this case VSE++ was trained with our loss function, allowing it to converge in this multilingual scenario. We also hightlight that after the 5^{th} epoch LIWE outperforms CLMR in all languages, becoming the best performing method across the remaining of the optimization. It is somewhat surprising that SCAN-t2i underperformed on validation set when compared to VSE++. Although, it achieved good predictive performance on test set.

4.4. Time Analysis

In order to demonstrate our approaches' efficiency, we have run the evaluation procedure ten times for our methods and SCAN (our strongest baseline). In average CLMR takes 3.1 seconds to encode data for 1,000 images and 5,000 captions and takes 0.15 seconds to calculate similarity between all pairs on CPU, and 0.07s on GPU. CLMR+LIWE takes 5.11s seconds for data encoding, 0.14s for building the similarity matrix on CPU, and 0.05s on GPU. On the other side, using the original SCAN code, it takes 10s to encode the very same data, and 180 seconds to build the similarity matrix on GPU. We have not evaluate their method on CPU. Our methods encode data up to $30\times$ faster, and calculates similarity matrices up to three orders of magnitude faster than the current state-of-the-art approach (on GPU). In addition, CLMR+LIWE is able to reduce the word embeddings to a fixed size number of parameters.

5. Related Work

There is recent work that employ a similar approach to approximate distinct languages in a semantic space by using image-caption pairs as pivot points [12, 33]. The work in [12] introduces such an idea by adapting a traditional pairwise loss function from [18], though the authors have only trained bilingual models. Their approach is limited in learning good language-invariant embeddings given limitations of both text and image encoders, as well as the loss function that is based on the sum of the hinges which often leads to local minima. The authors in [33] propose a multilingual embedding approach based on deep Partial Canonical Correlation Analysis, which is designed for handling two main semantic tasks, namely multilingual word similarity and cross-lingual image description retrieval.

In [24], the authors show that different languages have similar word embedding spaces. Based on this notion, sev-

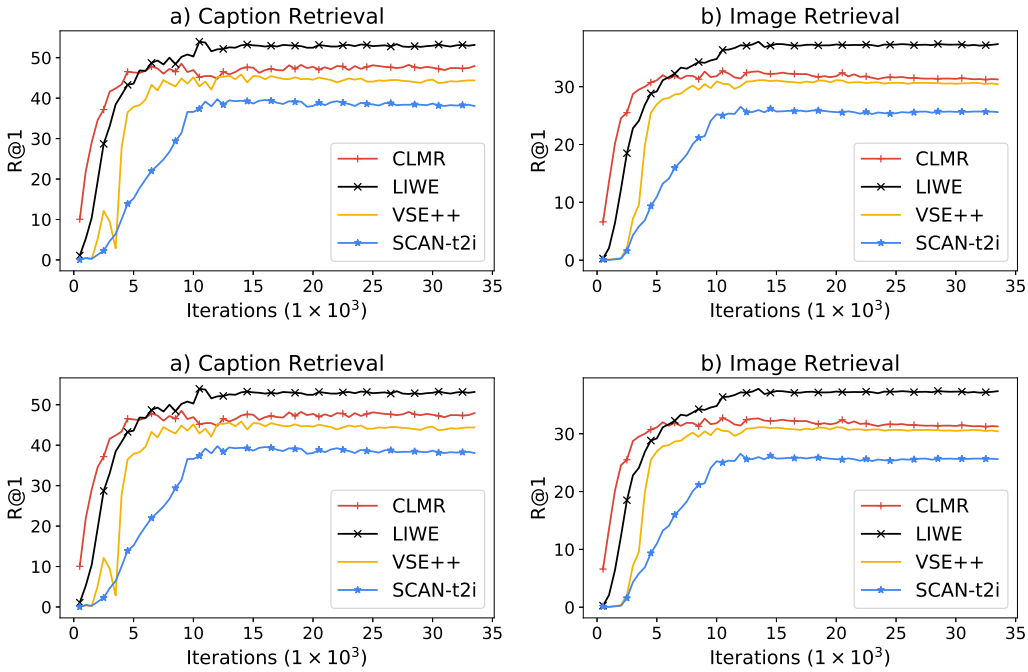

Figure 2. R@1 cross-modal language-invariant values for image-to-text and text-to-image across training epochs on Flickr30k (first row) and Multi30k (second row).

eral papers have proposed algorithms for cross-language alignment [11, 42, 2, 34, 3, 4, 6]. Our work follows the same assumption, though from the best of our knowledge, this is the first work that makes use of character-based inputs to improve multi-language cross-modal retrieval.

Previous work have extensively explored the cross-modal retrieval task relying on word-level features [36, 7, 14, 38, 22]. In [41, 39, 40], authors explored a character-level module designed to learn textual semantic embeddings by convolving raw characters with distinct granularity levels. Despite being conceptually much simpler and requiring fewer parameters, their methods outperformed state-of-the-art results.

Anderson et al. [1] proposed the use of an object detector to extract regions features from raw images instead of a single feature representation for the image. Lee et al. [22] have shown that such features could increase cross-modal retrieval performance with the aid of stacked attention layers, once it is capable of retain more detailed information and highlight more relevant content.

Recently, Elliott et al. [8] created the Multi30k dataset that extends the Flickr30K dataset with German translations created by professional translators over English descriptions. To the best of our knowledge, our work is the first to propose character-level embeddings that are language-invariant via a co-training strategy that leverages aligned multi-language corpora for helping in the task of cross-modal retrieval.

6. Conclusions

In this paper, we propose a novel approach for cross-modal retrieval that learns language-invariant multimodal embeddings. The proposed framework CLMR makes use of improved text and image encoding functions, along with a more robust loss function. We also introduce a novel data representation approach, in which we replace the traditional word-embedding matrix with a module that maps the character sequences to a word-level embedding space.

We have shown that our novel architecture outperforms state-of-the-art models for the image annotation task ($R@1$) in the widely used MS COCO and Flickr30k datasets, while not requiring costly computation of cross-modal attention mechanisms. Our models also present the best performance and overall suitability for learning language-invariant representations, as seen in the results for the Multi30k dataset.

As future work, we intend to explore several other languages within this framework, and also verify the potential of the proposed co-training strategy for other cross-modal tasks such as image captioning, visual question & answering and image synthesis.

Acknowledgments

We thank Google, and the Brazilian research CNPq, and FAPERGS for funding this research. This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001*.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017.

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, 2016.

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 451–462, 2017.

[4] Hailong Cao, Tiejun Zhao, Shu Zhang, and Yao Meng. A distribution-based model to learn bilingual word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1818–1827, 2016.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[6] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*, 2016.

[7] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Desmond Elliott and 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, Berlin, Germany*.

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. 2017.

[10] Fartash Faghri, David J. Fleet, Ryan Kiros, and Sanja Fidler. VSE++: improved visual-semantic embeddings. *CoRR*, abs/1707.05612, 2017.

[11] Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, 2014.

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

[13] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. *arXiv preprint arXiv:1711.06420*, 2017.

[14] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] Yan Huang, Qi Wu, and Liang Wang. Learning semantic concepts and order for image and sentence matching. *arXiv preprint arXiv:1712.02036*, 2017.

[16] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[17] Yeachan Kim, Kang-Min Kim, Ji-Min Lee, and SangKeun Lee. Learning to generate word representations using subword information. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2551–2561, 2018.

[18] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 595–603, 2014.

[19] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[20] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[22] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*, 2018.

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision*, 2014.

[24] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[26] Takashi Miyazaki and Nobuyuki Shimizu. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1780–1790, 2016.

[27] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016.

[28] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[29] Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. Mimicking word embeddings using subword rnns. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112, 2017.

[30] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[32] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017.

[33] Guy Rotman, Ivan Vulić, and Roi Reichart. Bridging languages through images with deep partial canonical correlation analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 910–921. Association for Computational Linguistics, 2018.

[34] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[36] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations (ICLR 2016)*, 2016.

[37] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *CVPR*, 2017.

[38] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013, 2016.

[39] Jônatas Wehrmann and Rodrigo C Barros. Bidirectional retrieval made simple. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7718–7726, 2018.

[40] Jônatas Wehrmann, Willian Becker, Henry EL Cagnini, and Rodrigo C Barros. A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 2384–2391. IEEE, 2017.

[41] Jônatas Wehrmann, Anderson Mattjie, and Rodrigo C Barros. Order embeddings and character-level convolutions for multimodal alignment. *Pattern Recognition Letters*, 102:15–22, 2018.

[42] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, 2015.

[43] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535*, 2017.