Sentence-Level Multilingual Multi-modal Embedding for Natural Language Processing

Iacer Calixto
ADAPT Centre
Dublin City University
Glasnevin, Dublin 9
iacer.calixto@adaptcentre.ie

Qun Liu
ADAPT Centre
Dublin City University
Glasnevin, Dublin 9
qun.liu@adaptcentre.ie

Abstract

We propose a novel discriminative ranking model that learns embeddings from multilingual and multi-modal data, meaning that our model can take advantage of images and descriptions in multiple languages to improve embedding quality. To that end, we introduce an objective function that uses pairwise ranking adapted to the case of three or more input sources. We compare our model against different baselines, and evaluate the robustness of our embeddings on image–sentence ranking (ISR), semantic textual similarity (STS), and neural machine translation (NMT). We find that the additional multilingual signals lead to improvements on all three tasks, and we highlight that our model can be used to consistently improve the adequacy of translations generated with NMT models when re-ranking n-best lists.

1 Introduction

Distributional semantic models (DSMs) compute word vector representations from text based on word co-occurrence patterns. However, these models suffer from an obvious limitation since the meaning of a word is derived entirely from connections to other words, i.e. they do not take extra-linguistic modalities into account and thus lack *grounding* (Glenberg and Robertson, 2000). This is the case not only of widely adopted word-level DSMs, e.g. word2vec (Mikolov et al., 2013), but also of sentence-level DSMs, e.g. skip-thought vectors (Kiros et al., 2015).

In this work, we address this issue and expand on the idea of training sentence-level multi-modal embeddings (Kiros et al., 2014; Socher et al., 2014), introducing a model that can be trained not only on images and their monolingual descriptions but also on additional multilingual image descriptions when these are available. We believe that having multiple descriptions of one image, regardless of its language, is likely to increase the coverage and variability of ideas described in the image, which may lead to a better generalisation of the depicted scene semantics. Moreover, a similar description expressed in different languages may differ in subtle but meaningful ways.

To that end, we introduce an objective function that uses pairwise ranking (Cohen et al., 1999) adapted to the case of three or more input sources, i.e. an image and multilingual sentences (§3). Our objective function links images and multiple sentences in an arbitrary number of languages, and we validate our idea in experiments where we use the Multi30k data set (§4).

We evaluate our embeddings in three different tasks: an image–sentence ranking (ISR) task (§6), in both directions, where we find that multilingual signals improve ISR to a large extent, i.e. the median ranks for English are improved from 8 to 5 and for German from 11 to 6, although the impact on ranking sentences given images is less conclusive; two sentence textual similarity (STS) tasks (§7), finding consistent improvements over a comparable monolingual baseline and outperforming the best published SemEval results; a neural machine translation (NMT) task (§8), where we use our model to re-rank n-best lists generated by different NMT models and report consistent improvements. Our main contributions are:

- we introduce a novel ranking-based objective function to train a discriminative model that utilises not only *multi-modal* but also *multilingual* data;
- we compare our proposed multilingual multi-modal embedding (MLMME) to embeddings

Proceedings of Recent Advances in Natural Language Processing, pages 139–148,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_020

trained on only one language (Kiros et al., 2014) on three different tasks (ISR, STS and NMT), as well as the Skip-Thought vectors when applicable (Kiros et al., 2015), and find that our model consistently improves over comparable monolingual baselines in all tasks but ranking sentences given images, where results are mixed.

2 Background and Related work

Multi-modal distributional semantic models try to expand DSMs and include inputs from additional modalities other than text as a means to address the grounding problem (Glenberg and Robertson, 2000). At the word level, Bruni et al. (2014) propose deriving word and image vectors, where the word vector representations are based on co-occurrence counts in text corpora, and the images are represented using a bag-of-visual-words method with Scale-Invariant Feature Transform (SIFT) vectors (Lowe, 1999, 2004) extracted from a data set of tagged images. These two representations are concatenated and merged using Singular Value Decomposition. Silberer and Lapata (2014) use stacked auto-encoders to map words and images to one same shared multi-modal embedding space. Their image representation is obtained using attribute classifiers that predict visual attributes (e.g., `has wings`, `made of wood`) for given words, proposed in Farhadi et al. (2009). Lazaridou et al. (2015) expand the word2vec *skip-gram* (Mikolov et al., 2013) into a multi-modal *skip-gram* model by incorporating image features extracted from pre-trained Convolutional Neural Networks (CNNs). Visual features obtained with pre-trained CNNs are widely used in transfer learning scenarios, such as in visual question answering (Zhang et al., 2016), to train multi-modal word embeddings (Lazaridou et al., 2015) or in multi-modal neural machine translation (Calixto et al., 2017).

All these DSMs have in common that they learn models at the word-level. Nonetheless, there are many models that propose to learn sentence-level (Kiros et al., 2015; Arora et al., 2017) or even paragraph-level vector representations (Le and Mikolov, 2014). Similarly to their word-level counterparts, these models are trained based on text signals only.

At the sentence level, Kiros et al. (2014) propose a multi-modal embedding model trained to map sentences and images into one shared multi-modal embedding space, where the sentences are encoded using Recurrent Neural Networks (RNN). In a similar vein, Socher et al. (2014) utilised Recursive Neural Networks, i.e. RNNs that operate on parse trees, as their sentence encoder. They both utilised pre-trained CNNs to extract image features and a pairwise ranking function to train their multi-modal embeddings.

We build on previous work and extend the idea of training multi-modal sentence-level embeddings to the scenario where the training data is not only *multi-modal*, but also *multilingual*. We thus put forward a model that integrates images and an arbitrary number of descriptions in different languages.

3 Multilingual and multi-modal embeddings (MLMME)

Our model has two main components: one *textual* and one *visual*. In the textual component, we have K different languages L_k, $k \in [1, K]$, and for each language we use a recurrent neural network (RNN) with gated recurrent units (GRU) (Cho et al., 2014) as a sentence encoder. Let $S^k = \{w_1^k, \ldots, w_{N_k}^k\}$ denote sentences composed of word indices in a language L_k, and $X^k = (\boldsymbol{x}_1^k, \boldsymbol{x}_2^k, \cdots, \boldsymbol{x}_{N_k}^k)$ the corresponding word embeddings for these sentences, where N_k is the sentence length. An RNN Φ_{enc}^k reads X^k word by word, from left to right, and generates a sequence of annotation vectors $(\boldsymbol{h}_1^k, \boldsymbol{h}_2^k, \cdots, \boldsymbol{h}_{N_k}^k)$ for each embedding $\boldsymbol{x}_i^k$, $i \in [1, N_k]$. For any given input sentence, we use the corresponding encoder RNN's last annotation vector $\boldsymbol{h}_{N_k}^k$ for that language L_k as the sentence representation, henceforth $\boldsymbol{v}^k$.

In our visual component we use publicly available pre-trained models for image feature extraction. Simonyan and Zisserman (2014) trained deep CNNs for classifying images into one out of 1000 ImageNet classes (Russakovsky et al., 2015). We use their 19-layer VGG network (VGG19) to extract feature vectors for all images in our dataset. More specifically, we use *global* features extracted from the penultimate fully-connected layer of the VGG19 network, which consists of a $4,096$D feature vector, henceforth FC7.

Each training example consists of a tuple *(i)* sentences S^k in L_k, $\forall k \in [1, K]$, and *(ii)* the associated image these sentences describe. Given

a training instance, we retrieve the embeddings $X^k = \{\boldsymbol{x}_1^k, \ldots, \boldsymbol{x}_{N_k}^k\}$ for each sentence S^k using one separate word embedding matrix for each language k. A sentence embedding representation $\boldsymbol{v}^k$ is then obtained by applying the encoder Φ_{enc}^k onto each embedding $\boldsymbol{x}_{1:N_k}^k$ and using the last annotation vector $\boldsymbol{h}_{N_k}^k$ of each RNN, after it has consumed the last token in each sentence. An image feature vector $\boldsymbol{q} \in \mathbb{R}^{4096}$ is extracted using the VGG19 CNN so that $\boldsymbol{d} = W_I \cdot \boldsymbol{q}$ is an image embedding and W_I is a model parameter. Also, image embeddings $\boldsymbol{d}$ and sentence embeddings $\boldsymbol{v}^k$, $\forall k \in [1, K]$ are normalised to unit norm and have the same dimensionality. Finally, $s_i(\boldsymbol{d}, \boldsymbol{v}^k) = \boldsymbol{d}^\top \cdot \boldsymbol{v}^k, k \in [1, K]$ is a function that computes the similarity between images and sentences in any language, and $s_s(\boldsymbol{v}^k, \boldsymbol{v}^l) = (\boldsymbol{v}^k)^\top \cdot \boldsymbol{v}^l, \forall k, l \in [1, K], k \neq l$, computes the similarity between sentences in two different languages.[1]

We now describe two *pairwise ranking* functions used in our objective, one that scores sentences and images, and another one that scores sentences in two different languages. Our model takes into consideration not only the relation between sentences in a given language and images computed by the $s_i(\cdot,\cdot)$ function, but also sentences in different languages in relation to each other, computed by $s_s(\cdot,\cdot)$. Our sentence–image, *multi-modal* ranking function is given in (1):

$$\begin{aligned}\text{R}_{\text{MM}} = &\sum_d \sum_r \max\{0, \alpha - s_i(\boldsymbol{d}, \boldsymbol{v}^k) + s_i(\boldsymbol{d}, \boldsymbol{v}_r^k)\} + \\ &\sum_{v^k} \sum_r \max\{0, \alpha - s_i(\boldsymbol{v}^k, \boldsymbol{d}) + s_i(\boldsymbol{v}^k, \boldsymbol{d}_r)\}, \\ &k \in K, \end{aligned} \tag{1}$$

where $\boldsymbol{v}_r^k$ (subscript *r* for *random*) is a contrastive or non-descriptive sentence embedding in language L_k for image embedding $\boldsymbol{d}$ and vice-versa, and α is a model parameter, i.e. the *margin*. R_{MM} learns to rank a sentence embedding v^k in language L_k, $k \in K$, against an image embedding $\boldsymbol{d}$, and vice-versa. Our sentence–sentence, *multilingual* ranking function is (2):

$$\begin{aligned}\text{R}_{\text{ML}} = &\sum_{v^k} \sum_r \max\{0, \alpha - s_s(\boldsymbol{v}^k, \boldsymbol{v}^l) + s_s(\boldsymbol{v}^k, \boldsymbol{v}_r^l)\} + \\ &\sum_{v^l} \sum_r \max\{0, \alpha - s_s(\boldsymbol{v}^l, \boldsymbol{v}^k) + s_s(\boldsymbol{v}^l, \boldsymbol{v}_r^k)\}, \\ &k \in K, l \in K, l \neq k, \end{aligned} \tag{2}$$

where $\boldsymbol{v}_r^k$ is a contrastive or non-descriptive sentence embedding in language L_k for sentence $\boldsymbol{v}^l$ in language L_l, and vice-versa. In both R_{MM} and R_{ML}, contrastive terms are chosen randomly from the training set and resampled at every epoch.

Finally, our optimisation function in Equation (3) minimises the linearly weighted combination of R_{MM} and R_{ML}:

$$\begin{aligned}&\min_{\theta_k, W_I} \beta \text{R}_{\text{MM}} + (1 - \beta)\text{R}_{\text{ML}}, \forall k \in K, \\ &0 \geq \beta \geq 1, \end{aligned} \tag{3}$$

where θ_k includes all the encoder RNNs parameters for language L_k, and W_I is the image transformation matrix. β is a model hyperparameter that controls how much influence a particular similarity (*multi-modal* or *multilingual*) has in the overall cost. We illustrate the model in Figure 1.

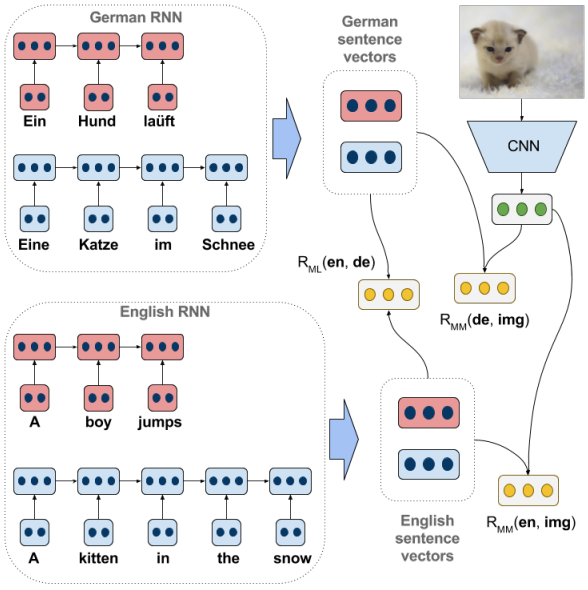

Figure 1: Multilingual multi-modal embedding trained with images and their English and German descriptions. The sentences in red denote contrastive examples, whereas the sentences in blue are descriptive of the image.

The two extreme scenarios are $\beta = 0$, in which case only the multilingual similarity is used, and $\beta = 1$, in which case only the multi-modal similarity is used. If the number of languages $K = 1$ and $\beta = 1$, our model computes the Visual Semantic Embedding (VSE) of Kiros et al. (2014).

4 Datasets

The original Flickr30k data set contains 30k images and 5 English sentence descriptions for each image (Young et al., 2014). We use the Multi30k

[1] In this work, both s_i and s_s are implemented as the dot product, but they could be any other suitable function.

	English							German				
	Skip-T.	VSE		Ours				VSE	Ours			
		paper	current	β=1	β=.75	$\beta = 0.5$	$\beta = 0.25$	current	β=1	β=.75	$\beta = 0.5$	$\beta = 0.25$
Sentence to image												
r@1	18.2	16.8	16.5	23.0 (+6.2)	**24.9** (+8.1)	22.3 (+5.5)	21.3 (+4.5)	13.5	**21.6** (+8.1)	20.3 (+6.8)	20.3 (+6.8)	19.5 (+6.0)
r@5	41.9	42.0	41.9	49.3 (+7.3)	**52.3** (+10.3)	48.3 (+6.3)	45.5 (+3.5)	36.6	**48.8** (+12.2)	45.0 (+8.4)	43.7 (+7.1)	43.0 (+6.4)
r@10	53.5	56.5	54.4	61.1 (+4.6)	**63.6** (+7.1)	58.4 (+1.9)	56.7 (+0.2)	49.0	**59.5** (+10.5)	56.6 (+7.6)	55.4 (+6.4)	54.4 (+5.4)
mrank	9	8	9	6	**5**	6	7	11	**6**	7	8	8
Image to sentence												
r@1	26.8	23.0	30.7	**33.1** (+2.4)	30.7 (+0.0)	27.4 (−3.3)	26.7 (−4.0)	30.5	**32.3** (+1.7)	24.9 (−5.6)	23.0 (−7.5)	21.8 (−8.7)
r@5	54.9	50.7	**57.8**	57.2 (−0.6)	55.4 (−2.4)	54.5 (−3.3)	51.4 (−6.4)	56.0	**58.6** (+2.6)	52.3 (−3.7)	48.4 (−7.6)	49.8 (−6.2)
r@10	67.5	62.9	**70.6**	68.7 (−1.9)	65.6 (−5.0)	64.0 (−6.6)	61.9 (−8.7)	**68.9**	68.1 (−0.8)	63.6 (−5.3)	62.8 (−6.1)	61.3 (−7.6)
mrank	5	5	**4**	**4**	**4**	**4**	5	**4**	**4**	5	6	6

Table 1: We show results for our MLMME model evaluated on the M30k$_C$ test set when trained using different values of β, and two monolingual baselines: the Skip-thought vectors (Skip-T.) of Kiros et al. (2015) and the VSE model of Kiros et al. (2014), where *paper* are the results reported in their paper and *current* were obtained when re-training their model. Best monolingual results are underlined and best overall results appear in bold. We show improvements over the best monolingual baseline in parenthesis.

data set (Elliott et al., 2016), which consists of two expansions of the Flickr30k.

To train NMT models (§8) we use the *translated Multi30k*, henceforth M30k$_T$, where for each of the 30k images in the original Flickr30k, one of its English descriptions is manually translated into German by a professional translator. Training, validation and test sets contain 29k, 1014 and 1k images respectively, each accompanied by one translated sentence pair in English and German. In all other experiments (§6 and §7), we use the *comparable Multi30k*, henceforth M30k$_C$, an expansion of the Flickr30k where 5 German descriptions were collected for each image in the original Flickr30k independently from the English descriptions. Training, validation and test sets contain 29k, 1014 and 1k images respectively, each accompanied by 5 English and 5 German sentences.

We split the M30k$_C$'s validation set in two and use the first 500 images and their corresponding bilingual sentences for model selection and the remaining 514 images and bilingual sentences for model evaluation. Source and target languages were estimated over the entire vocabulary, i.e. ∼22 English and ∼34 German tokens.

5 MLMME experimental setup

For each language we train a separate 1024D encoder RNN with GRU. Word embeddings are 620D and trained jointly with the model. All non-recurrent matrices are initialised by sampling from a Gaussian $\mathcal{N}(0, 0.01)$, recurrent matrices are random orthogonal and bias vectors are all initialised to zero. We apply dropout (Srivastava et al., 2014) with a probability of 0.5 in both text and image representations, which are in turn mapped onto a 2048D multi-modal embedding space. We set the margin $\alpha = 0.2$. Our models are trained using stochastic gradient descent with Adam (Kingma and Ba, 2015) with minibatches of 128 instances.

As our main baseline, we retrain Kiros et al. (2014) monolingual models separately on the M30k$_C$'s English and German sentences (+images), whereas model MLMME is trained on the entire M30k$_C$.

When processing English sentences and images, we additionally use the pre-trained Skip-Thought vectors (Kiros et al., 2015), more specifically the 4800D *combine-skip* vectors as a second baseline. We follow the authors description[2] on how to do it: *(i)* we use their pre-trained encoders to compute the English sentence representations, i.e. a 4800D vector; *(ii)* we train their model on the M30k$_C$ training set using their image–sentence ranking model; *(iii)* we select the model with the best performance of the M30k$_C$ validation set and use it to compute results in the test set.

6 Image↔Sentence Ranking

In Table 1, we show results for the monolingual English Skip-thought vectors of Kiros et al. (2015), the monolingual VSE English and German models of Kiros et al. (2014) and our MLMME models on the M30k$_C$ data set and evaluated on images and bilingual sentences. Recall-at-k ($r@k$) measures the mean number of times the correct result appear in the top-k retrieved entries and *mrank* is the median rank.

[2] https://github.com/ryankiros/skip-thoughts#image-sentence-ranking

First, we note that multilingual models show consistent improvements in ranking images given sentences. All our models, regardless of the value of the hyperparameter β ($= .25, .5, .75, 1$), show strong improvements in recall@k (up to $+12.2$) and median rank (in English, the mrank is reduced from 8 to 5 and in German from 11 to 6 in comparison to the best model by Kiros et al. (2014)). Nevertheless, when ranking sentences given images, results are less conclusive. The best results achieved by our multilingual models, for both languages, are observed when $\beta = 1$, with the recall@k slightly deteriorating as we include more multilingual similarity, i.e. $\beta = .75, .5, .25$, and the median rank also slightly increasing for English (from 4 to 5) and German (from 4 to 6). In short, model MLMME consistently improves over all baselines when ranking images given sentences, and applying model MLMME with β=1 to rank sentences given images performs comparably to the monolingual VSE baseline and clearly improves over using the Skip-Thought model on the same task.

6.1 Discussion

Using image features are crucial in *grounding* the sentence vector representations. We note that using β=0 in Equation 3 is equivalent to using only multilingual similarity scores (eq. 2), and no multi-modal similarities (eq. 1). However, in the training data there are multiple sentences describing one same image, and by not using the multimodal similarity the model loses the ability to generalise and project semantically similar sentences, i.e. sentences that describe one same image, close together. In other words, the model has no way of mapping the comparable sentences that describe one same image together, since the link between these sentences are the image they describe.

In practice, we noted that using $\beta = 0$ leads to a model that cannot learn to rank sentences given images and vice-versa, i.e. the results for median ranks in Table 1 when $\beta = 0$ drop to chance levels. For that reason, we do not include $\beta = 0$ in our hyperparameter search for the experiments we report in Sections 7 and 8.

7 Semantic Textual Similarity

In the semantic textual similarity task, we use our model to compute the distance between a pair of sentences (distances are equivalent to cosine similarity and therefore lie in the $[0, 1]$ interval). Gold standard scores for all tasks are given in the $[0, 5]$ interval, where 0 means complete dissimilarity and 5 complete similarity. We simply use the cosine similarity distance and scale it by 5, directly comparing it to the gold standard scores. There is no SemEval data set including the German language, therefore we only use our English encoders to compute embedding vectors for both sentences in each entry in the test set. We report results for the two in-domain similarity tasks in SemEval, specifically the image description similarity tasks from years 2014 (Agirre et al., 2014) and 2015 (Agirre et al., 2015).

Test set	VSE	Our model β=1	β=.75	β=.5	β=.25	SemEval best
in-domain data						
IMG_1	.791	.797	.819	**.826**	.817	.821
IMG_2	.834	.880	.882	.885	**.886**	.864

Table 2: Pearson rank correlation scores for semantic textual similarities in two different SemEval test sets. IMG_1: image descriptions (2014), IMG_2: image descriptions (2015).

In Table 2, we note that our MLMME model consistently improves on the monolingual baseline of Kiros et al. (2014) in the two in-domain similarity tasks, and our best models also outperform the best published SemEval results.

We note that we only use the English side of our models (+images) in these two evaluations, but we do not directly use our German encoders since there are no German sentences in the SemEval STS task. Nonetheless, training on additional German sentences—incorporated in our model via the German encoder—clearly increases the quality of the English encoder, specially for lower values of β as can be seen in Table 2. These are interesting results, showing that the additional multilingual data brings holistic effects to the entire model and makes the overall model better.

8 Neural Machine Translation (NMT)

In this set of experiments, we use model MLMME to re-rank n-best lists generated with baseline text-only NMT models. Arguably, the main advantage of using such discriminative models to re-rank n-best lists instead of directly training a multi-modal NMT model is the shorter training time . Whereas training the discriminative MLMME model on the Multi30k data set takes $\sim$6 hours, training a multi-

modal NMT model on the same data set usually takes many days.

In order to evaluate how VSE and MLMME models perform in n-best list re-ranking, we train NMT baselines based on the model of Bahdanau et al. (2015) using different hyper-parameter settings. All NMT models have an encoder bidirectional RNN with GRU (one 1024D single-layer forward RNN and one 1024D single-layer backward RNN). Source and target word embeddings are 620D each and both are trained jointly. All non-recurrent matrices are initialised by sampling from a Gaussian ($\mu = 0, \sigma = 0.01$), recurrent matrices are orthogonal and bias vectors are all initialised to zero. The decoder is an attention-based RNN with GRU and is a neural LM (Bengio et al., 2003).

NMT models are trained using stochastic gradient descent with Adadelta (Zeiler, 2012) and minibatches of size 40, where each training instance consists of one English sentence, one German sentence and one image. We apply early stopping for model selection based on BLEU scores. We evaluate our models' translation quality quantitatively in terms of BLEU4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), and TER (Snover et al., 2006) and we compute statistical significance using approximate randomisation computed with the MultEval toolkit (Clark et al., 2011).

8.1 NMT baselines

We train one *weak* model, one *regular* model and one *optimised* NMT model on the translated Multi30k training data set (without images) to translate from English into German. In order to train these three different models, we search for the best dropout and L2 regularisation weight combination by observing model performance on the validation set. The search space for the dropout hyper-parameter is the set $\{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}$, and for the L2 regularisation weight is the set $\{0.0, 1e{-}1, 1e{-}2, \ldots, 1e{-}9, 1e{-}10\}$.

Weak model Our *weak* model is the text-only NMT baseline model trained with no regularisation, i.e. L2 regularisation weight is 0.0 and dropout probability is 0.0. It corresponds to the model with the worst performance on the translated Multi30k validation set.

Regular model Our *regular* model is a text-only NMT baseline model with medium-performance regularisation. Specifically, from the hyper-parameter search on the translated Multi30k validation set, we use a weight of $1e{-}8$ to scale the L2 regularisation term and a dropout of 0.5;

Optimised model Our *optimised* model is the text-only NMT baseline model that has the best performance on the translated Multi30k validation set, according to our dropout and L2 regularisation hyper-parameters search. This corresponds to the model with no L2 regularisation, i.e. L2 weight is 0.0, and dropout with probability 0.2.

8.2 N-best re-ranker

We first use the three different NMT models we have just described to generate n-best lists ($n \in \{20, 50\}$) for each entry in the M30k$_\text{T}$ validation and test sets. Second, we use the monolingual VSE model (Kiros et al., 2014) trained on German sentences and images to compute the distance between translations into German and images, for all entries in the M30k$_\text{T}$ validation and test sets. We also use our MLMME models trained with $\beta \in \{.25, .5, .75, 1\}$ to compute the distance between German and English sentences with $s_s(\cdot, \cdot)$, and between a German sentence and an image using $s_i(\cdot, \cdot)$, for all entries in the M30k$_\text{T}$ validation and test sets.

We then train an n-best list re-ranker on the M30k$_\text{T}$ validation set's 20-best (50-best) lists with k-best MIRA (Crammer and Singer, 2003; Cherry and Foster, 2012), and use the new distances as additional features to the original MT log-likelihood $p(Y \mid X)$. We finally apply the optimised weights to re-rank the test set's 20-best (50-best) lists.

8.3 Results

In Tables 3 and 4, we show results obtained with the *weak*, the *regular*, and the *optimised* models when used to re-rank 20-best and 50-best lists, respectively. We compute 20- and 50-best lists to be able to observe whether the different models we use in re-ranking generate consistent results, regardless of the size of the n-best lists.

In order to measure of the quality of the n-best lists generated by the different models, we compute their oracle scores. The difference between the oracle scores for the n-best lists generated by the weak and the regular model is considerable: 8.8/10.4 BLEU, 7.9/8.0 METEOR, and

	BLEU		METEOR		TER	
Weak NMT model						
baseline	25.7		43.1		56.1	
+ VSE	25.8	(+0.1)	43.2	(+0.1)	56.1	(-0.0)
+ MLMME, $\beta = 1$	26.1	(+0.4)	**44.4**†‡	**(+1.3)**	55.5	(-0.6)
+ MLMME, $\beta = 0.75$	26.1	(+0.4)	44.3†‡	(+1.2)	55.9	(-0.2)
+ MLMME, $\beta = 0.5$	26.0	(+0.3)	43.9†‡	(+0.8)	55.9	(-0.2)
+ MLMME, $\beta = 0.25$	**26.3**†‡	**(+0.6)**	44.3†‡	(+1.2)	**55.2**†‡	**(-0.9)**
oracle	33.1		51.4		46.5	
Regular NMT model						
baseline	32.4		50.7		51.9	
+ VSE	32.2	(-0.2)	50.7	(+0.0)	52.6	(+0.7)
+ MLMME, $\beta = 1$	**33.8**†‡	**(+1.4)**	**51.4**†‡	**(+0.7)**	49.0‡	(-2.9)
+ MLMME, $\beta = 0.75$	33.5‡	(+1.1)	51.3†‡	(+0.6)	49.0‡	(-2.9)
+ MLMME, $\beta = 0.5$	**33.8**†‡	**(+1.4)**	**51.4**†‡	**(+0.7)**	**48.6**†‡	**(-3.3)**
! + MLMME, $\beta = 0.25$	33.7‡	(+1.3)	**51.4**†‡	**(+0.7)**	49.4‡	(-2.5)
oracle	41.9		59.3		41.2	
Optimised NMT model						
baseline	35.3		52.3		44.9	
+ VSE	32.3	(-3.0)	49.8	(-2.5)	46.5	(+1.6)
+ MLMME, $\beta = 1$	35.3‡	(+0.0)	**52.7**†‡	**(+0.4)**	**44.5**‡	**(-0.4)**
+ MLMME, $\beta = 0.75$	35.2‡	(-0.1)	52.6‡	(+0.3)	44.6‡	(-0.3)
+ MLMME, $\beta = 0.5$	35.1‡	(-0.2)	52.3‡	(+0.0)	44.9‡	(-0.0)
+ MLMME, $\beta = 0.25$	**35.7**‡	**(+0.4)**	**52.7**‡	**(+0.4)**	**44.5**‡	**(-0.4)**
oracle	43.2		59.7		37.8	

Table 3: MT evaluation metrics computed for 1-best translations generated with three NMT baselines, and for 20-best lists re-ranked using VSE and MLMME as discriminative features. Results improve significantly over the corresponding 1-best baseline (†) or over the translations obtained with the VSE re-ranker (‡) with $p = 0.05$.

	BLEU		METEOR		TER	
Weak NMT model						
baseline	25.7		43.1		56.1	
+ VSE	25.8	(+0.1)	43.5†	(+0.4)	56.1	(-0.0)
+ MLMME, $\beta = 1$	26.2	(+0.5)	**44.6**†‡	**(+1.5)**	55.4	(-0.7)
+ MLMME, $\beta = 0.75$	**26.4**†	**(+0.7)**	44.5†‡	(+1.4)	55.6	(-0.5)
+ MLMME, $\beta = 0.5$	25.9	(+0.2)	43.9†	(+0.8)	55.9	(-0.0)
+ MLMME, $\beta = 0.25$	**26.4**†‡	**(+0.7)**	44.5†‡	(+1.4)	**55.0**†‡	**(-1.1)**
oracle	36.2		53.8		43.4	
Regular NMT model						
baseline	32.4		50.7		51.9	
+ VSE	32.7	(-0.3)	50.8	(+0.1)	51.4	(-0.5)
+ MLMME, $\beta = 1$	**34.2**†‡	**(+1.8)**	**51.6**†‡	**(+0.9)**	48.3‡	(-3.6)
+ MLMME, $\beta = 0.75$	34.1†‡	(+1.7)	**51.6**†‡	**(+0.9)**	47.6†‡	(-4.3)
+ MLMME, $\beta = 0.5$	34.0†‡	(+1.6)	51.4†‡	(+0.7)	**47.3**†‡	**(-4.6)**
+ MLMME, $\beta = 0.25$	34.1†‡	(+1.7)	**51.6**†‡	**(+0.9)**	48.5‡	(-3.4)
oracle	46.6		61.8		34.1	
Optimised NMT model						
baseline	35.3		52.3		44.9	
+ VSE	30.7	(-4.6)	47.9	(-4.4)	48.6	(+3.7)
+ MLMME, $\beta = 1$	35.4‡	(+0.1)	**52.7**†‡	**(+0.4)**	**44.4**†‡	**(-0.5)**
+ MLMME, $\beta = 0.75$	35.2‡	(-0.1)	52.5‡	(+0.2)	44.7‡	(-0.2)
+ MLMME, $\beta = 0.5$	35.1‡	(-0.2)	52.3‡	(+0.0)	44.7‡	(-0.2)
+ MLMME, $\beta = 0.25$	**35.6**‡	**(+0.3)**	52.6‡	(+0.3)	**44.4**†‡	**(-0.5)**
oracle	46.3		61.9		34.9	

Table 4: MT evaluation metrics computed for 1-best translations generated with three NMT baselines, and for 50-best lists re-ranked using VSE and MLMME as discriminative features. Results improve significantly over the corresponding 1-best baseline (†) or over the translations obtained with the VSE re-ranker (‡) with $p = 0.05$.

5.3/9.3 TER, for the 20-best and 50-best lists respectively. Nevertheless, the difference between the oracle scores for the n-best lists generated by the regular and the optimised model is not nearly as high: 1.2/$-$0.3 BLEU, 0.0/0.1 METEOR, and 3.4/0.8 TER, again for the 20-best and 50-best lists respectively. However, when we analyse the metrics scores obtained by re-ranked models, we see a considerable difference between the improvements brought by VSE and MLMME features to the regular and optimised models.

Weak model First of all, using VSE features to re-rank n-best lists generated by the weak model practically does not change translations. MLMME features have a strong impact on METEOR scores, suggesting that they are making translations more adequate by improving their word-level recall. Using MLMME features to re-rank significantly improves METEOR in relation to the baseline and to the translations obtained with the VSE-features re-ranked model, for all values of β and for all n-best list sizes.

The model re-ranked with MLMME features with $\beta = 0.25$ performs best in this scenario. It is the only model that significantly improves on the three automatic metrics over both the 1-best baseline and the VSE-features re-ranked model, for all n-best lists sizes ($p = 0.05$).

Regular model Again, using VSE features to re-rank n-best lists generated by the regular model does not change translations in practice. Nevertheless, VSE re-ranked models are the only ones to show some small deterioration in relation to the baseline, even though these differences are not statistically significant. Models re-ranked with MLMME features are consistently better than the baseline, for all values of β and $n \in \{20, 50\}$. They also show strong improvements on METEOR scores in relation to both the baseline and to the translations obtained with the VSE-features re-ranked model, suggesting that they are still making translations more adequate by improving their word-level recall.

When applied to re-rank 50-best lists, MLMME features also significantly improve BLEU scores in relation to the baseline and to the translations obtained with the VSE-features re-ranked model, in spite of the values of β.

Optimised model First of all, we see that improving on the baseline using VSE or MLMME

features becomes harder when applied to n-best lists generated by the optimised model. From looking at the results, perhaps the most apparent outcome is the poor results obtained when using VSE features in this scenario. Using the additional VSE features to re-rank consistently and significantly deteriorate translations, for all n-best lists sizes ($n = \{20, 50\}$). The same does not happen when using MLMME features to re-rank n-best lists. MLMME features lead to translations that consistenly improve over those obtained with the VSE-features re-ranked model, for all different configurations of MLMME models ($\beta = \{0.25, 0.5, 0.75, 1.0\}$).

Model MLMME with $\beta = 0.25$ or $\beta = 1.0$ obtain the best results regardless of the n-best list sizes. These are the only two models that also significantly improve on the corresponding 1-best baseline according to at least one of the metrics.

8.4 Final Remarks

In this set of experiments we evaluated how well VSE (Kiros et al., 2014) and MLMME models perform when used to compute features to re-rank n-best lists. We found that VSE features do not affect translations when n-best lists are generated by less optimised NMT models, but they become less attractive as the baseline NMT models used to generate n-best lists gets better, getting to the point of significantly harming BLEU, METEOR and TER in the case of a highly optimised model.

In general, MLMME features outperformed VSE features across different scenarios, and seem to have a stronger impact on re-ranking n-best lists generated with the regular model compared to the weak and optimised models. Nonetheless, when applied to translations generated with the optimised model, MLMME models with $\beta = 0.25$ or $\beta = 1.0$ achieve the best results. They consistently and significantly increase METEOR scores, for all n-best lists sizes ($n \in \{20, 50\}$), which is an important finding since NMT models are known to suffer from adequacy issues (Tu et al., 2016).

Finally, MLMME models take considerably less time to train compared to a fully fledged multi-modal NMT model: training MLMME models take ∼3–6 hours, whereas training a text-only attention-based NMT model should take ∼3–4 days.[3] Likewise, using MLMME models to compute features at inference time is fast: it takes the time to encode the source and target sentences with the corresponding source- and target-language RNNs, the image with the pre-trained CNN, and then performing three dot products: source·target, target·image, and source·image.

Arguably, our results puts MLMME models as attractive candidates to be included in an NLP pipeline for processing image descriptions.

[3]This is the case of training an English–German translation model, evaluated in this work, on the Multi30k data set.

9 Conclusions

We propose a new discriminative ranking model that incorporates both multilingual and multimodal similarities, and obtain promising results in three different NLP tasks. We train our models using an objective function based on pairwise ranking. When applied to the task of image–sentence ranking, our model consistently outperforms all baselines when ranking images given sentences; our model when β=1 performs comparably to the monolingual VSE baseline, and the more weight we add to the multilingual similarity in the training objective, the worse the model ranks sentences given images. However, when applied to the task of Semantic Textual Similarity, our model outperforms the best published SemEval models in two image description similarity tasks, and when applied to re-rank n-best lists generated with different NMT models, they consistently improve translations as measured by three different MT metrics. We note that it has a consistent impact on METEOR, which is a recall-oriented metric that emphasises the *adequacy* of translations, which is precisely a problem that NMT models are known to suffer from (Tu et al., 2016). In the future we will train our model on a many-languages setting, with images and descriptions in ∼10 languages.

Acknowledgments

This project has received funding from Science Foundation Ireland in the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund and the European Union Horizon 2020 research and innovation programme under grant agreement 645452 (QT21).

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado, pages 252–263. http://www.aclweb.org/anthology/S15-2045.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland, pages 81–91. http://www.aclweb.org/anthology/S14-2010.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *International Conference on Learning Representations, ICLR 2017*. Toulon, France. https://openreview.net/pdf?id=SyK00v5xx.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations, ICLR 2015*. San Diego, California. http://arxiv.org/abs/1409.0473.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* 3:1137–1155. http://dl.acm.org/citation.cfm?id=944919.944966.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Int. Res.* 49(1):1–47. http://dl.acm.org/citation.cfm?id=2655713.2655714.

Iacer Calixto, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. 2017. Using images to improve machine-translating e-commerce product listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain, pages 637–643. http://www.aclweb.org/anthology/E17-2101.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montrèal, Canada, pages 427–436. http://aclweb.org/anthology/N12-1047.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pages 1724–1734. http://www.aclweb.org/anthology/D14-1179.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 176–181. http://dl.acm.org/citation.cfm?id=2002736.2002774.

W. W. Cohen, R. E. Schapire, and Y. Singer. 1999. Learning to order things. *Journal of Artificial Intelligence Research* 10:243–270.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research* 3:951–991. https://doi.org/10.1162/jmlr.2003.3.4-5.951.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA, pages 376–380. http://www.aclweb.org/anthology/W/W14/W14-3348.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language, VL@ACL 2016*. Berlin, Germany. http://aclweb.org/anthology/W/W16/W16-3210.pdf.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR*. Miami, Florida, USA, pages 1778–1785. https://doi.org/10.1109/CVPR.2009.5206772.

A. Glenberg and D. Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language* http://psych.wisc.edu/glenberg/.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR 2015*. San Diego, California.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR* abs/1411.2539. http://arxiv.org/abs/1411.2539.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought Vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, NIPS'15, pages 3294–3302. http://dl.acm.org/citation.cfm?id=2969442.2969607.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado, pages 153–163. http://www.aclweb.org/anthology/N15-1016.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*.

David G. Lowe. 1999. Object Recognition from Local Scale-Invariant Features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*. IEEE Computer Society, Washington, DC, USA, ICCV '99, pages 1150–. http://dl.acm.org/citation.cfm?id=850924.851523.

David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision* 60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS*. Lake Tahoe, Nevada, NIPS'13, pages 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y.

Carina Silberer and Mirella Lapata. 2014. Learning Grounded Meaning Representations with Autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland, pages 721–732. http://www.aclweb.org/anthology/P14-1068.

K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*. Cambridge, MA, pages 223–231.

Richard Socher, Karpathy Andrej, Q Le, Chris Manning, and Andrew Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics* 2:207–218.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pages 76–85. http://www.aclweb.org/anthology/P16-1008.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR* abs/1212.5701. http://arxiv.org/abs/1212.5701.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, Nevada, USA.