If Sentences Could See: Investigating Visual Information for Semantic Textual Similarity

Goran Glavaš[1], Ivan Vulić[2], and Simone Paolo Ponzetto[1]

[1]Data and Web Science Group
University of Mannheim
{goran, simone}@informatik.uni-mannheim.de

[2]Language Technology Lab
University of Cambridge
iv250@cam.ac.uk

Abstract

We investigate the effects of incorporating visual signal from images into unsupervised Semantic Textual Similarity (STS) measures. STS measures exploiting visual signal alone are shown to outperform, in some settings, linguistic-only measures by a wide margin, whereas multi-modal measures yield further performance gains. We also show that selective inclusion of visual information may further boost performance in the multi-modal setup.

1 Introduction

Semantic textual similarity (Agirre et al., 2012, 2015, *inter alia*) measures the degree of semantic equivalence between short texts, usually pairs of sentences. Despite the obvious applicability to sentence alignment for machine translation (MT) (Resnik and Smith, 2003; Aziz and Specia, 2011) or plagiarism detection (Potthast et al., 2011; Franco-Salvador et al., 2013), cross-lingual STS models were proposed only recently (Agirre et al., 2016; Brychcín and Svoboda, 2016; Jimenez, 2016). These are, however, essentially monolingual STS models coupled with full-blown MT systems that translate sentences to English.

Although research in cognitive science (e.g., Lakoff and Johnson (1999); Louwerse (2011)) shows that our meaning representations are grounded in perceptual system, the existing STS models (monolingual and cross-lingual alike) exploit only linguistic signals, despite the fact that models using perceptual information outperform uni-modal linguistic models on tasks like detecting conceptual association and word similarity (Silberer and Lapata, 2012; Bruni et al., 2014; Kiela and Bottou, 2014), predicting phrase compositionality (Roller and Schulte Im Walde, 2013), recognizing lexical entailment (Kiela et al., 2015), and metaphor detection (Shutova et al., 2016). While still predominantly applied in monolingual settings, representations originating from the visual modality are inherently language-invariable (Bergsma and Durme, 2011; Kiela et al., 2015). As such, they could serve as a natural cross-language bridge in cross-lingual STS.

In this work, we investigate unsupervised multi-modal and cross-lingual STS models that leverage visual information from images alongside linguistic information from textual corpora. We feed images retrieved for textual queries to the deep convolutional network (CNN) for image classification and use the CNN's abstract image features as visual semantic representations of words and sentences.

We implement models that combine linguistic and visual information at different levels of granularity – *early fusion* (word level), *middle fusion* (sentence level), and *late fusion* (fusing similarity scores). Results of an evaluation on two cross-lingual STS datasets (mutually very different in terms of text genre and average sentence length) show that (1) the proposed multi-modal STS models outperform uni-modal models relying only on visual or linguistic input and (2) in several experimental runs, purely visual STS models outperform purely linguistic STS models. We obtain further performance gains by selectively exploiting visual information, conditioned on the dispersion of retrieved images.

2 Related Work

We provide an overview of two different lines of research: (1) existing STS methods, which exploit linguistic information only and (2) multi-modal semantic representations used in other applications.

Semantic Textual Similarity. Despite the existence of earlier models (Islam and Inkpen, 2008; Oliva et al., 2011), the true explosion of STS research efforts is credited to the SemEval-2012 Pilot on Semantic Textual Similarity (Agirre et al., 2012). The most successful systems (Bär et al., 2012; Šarić et al., 2012) were methodologically similar – they employed a supervised regression model to learn the optimal combination of many different sentence-comparison features.

Subsequent STS shared tasks witnessed successful unsupervised STS models, based on aligning words between sentences and counting the number of aligned pairs. Han et al. (2013) use LSA-based and WordNet-based measures of word similarity to find the pairs of semantically aligned words. Sultan et al. (2014) further employ NER and dependency parsing to better align the words between the sentences. These models depend on language-specific resources and tools, which are fairly expensive to build and exist only for a handful of languages.

The cross-lingual STS has been tackled only in the most recent edition of the SemEval STS shared task (Agirre et al., 2016). The best performing systems (Brychcín and Svoboda, 2016; Jimenez, 2016), with over 90% correlation with human similarity scores, directly employ full-blown MT systems and next apply monolingual STS measures. Besides being as resource-intensive as monolingual

STS models, the applicability of this methodology is limited to language pairs for which a robust MT model exists.

The multi-modal STS measures proposed in this work are resource-light and do not require any language-specific resources and tools. For a given language, our models require only (1) reasonably large corpora to obtain linguistic representations (i.e., word embeddings) and (2) an image-retrieval system to obtain visual representations (i.e., image embeddings). In the cross-lingual STS setting, the models additionally require a reasonably small set of word translation pairs to learn a shared cross-lingual vector space (Mikolov et al., 2013).

Multi-modal semantics. While research in cognitive science clearly suggests that human meaning representations are grounded in our perceptual system and sensori-motor experience (Harnad, 1990; Lakoff and Johnson, 1999; Louwerse, 2011, *inter alia*), previous STS models relied exclusively on linguistic processing and textual information. To the best of our knowledge, there has not yet been an STS method that leveraged visual information and combined linguistic and visual input into a visually-informed multi-modal STS system. However, such visually-informed models have been successfully used in other tasks such as selectional preferences (Bergsma and Goebel, 2011), detecting semantic similarity and relatedness (Silberer and Lapata, 2012; Bruni et al., 2014; Kiela and Bottou, 2014), recognizing lexical entailment (Kiela et al., 2015), and metaphor detection (Shutova et al., 2016), to name only a few.

Another important property of visual data is their expected language invariance,[1] exploited in recent work on multi-modal modeling in cross-lingual settings (Bergsma and Durme, 2011; Kiela et al., 2015; Vulić et al., 2016; Specia et al., 2016). Supported by these findings, in this work we show that our multi-modal STS framework may be straightforwardly extended to cross-lingual settings.

3 Multi-Modal Concept Representations

Our multi-modal STS measures combine – at different fusion levels – linguistic and visual concept representations. We obtain linguistic and visual representations for unigrams and then derive sentence representations by aggregating unigram representations. This was a pragmatic decision, as we were unable to consistently retrieve images for whole sentences as queries.

3.1 Linguistic Representations

We use the ubiquitous word embeddings as the linguistic representations of words. Aiming to make our approach language-independent, we opted for embedding models that require nothing but the large corpora as input. Due to the common

[1] Using a simple example from Vulić et al. (2016), bicycles resemble each other irrespective of whether we call them *bicycle*, *vélo*, *fiets*, *bicicletta*, or *Fahrrad*; see also Fig. 1

usage, we chose the Skip-Gram (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) embeddings.

For the cross-lingual STS setting, the words of the two languages have to be projected to the same embedding space. To achieve this, we employ the translation matrix model of Mikolov et al. (2013), who have shown that the linear mapping can be established between independently trained embedding spaces. Given a set of translation pairs $\{s_i, t_i\}_{i=1}^{n}$, $s_i \in \mathbb{R}^{d_s}$, $t_i \in \mathbb{R}^{d_t}$ (with d_s and d_t being the sizes of source and target embeddings, respectively), we obtain the translation matrix $M \in \mathbb{R}^{d_s \times d_t}$ by minimizing the sum:

$$\sum_{i=1}^{n} \|s_i M - t_i\|_2$$

Once learned, the matrix M is used to project the embeddings of the whole source language vocabulary to the embedding space of the target language.

3.2 Visual Representations

Image embeddings. We use a standard procedure to obtain visual representations for words, e.g., (Kiela et al., 2015, 2016): we first retrieve n images for the word via Bing image search ($n = 20$ in all experiments).[2] Example images for the four languages we consider in our experiments (cf. Section 5) are shown in Figure 1.

We next run a deep convolutional neural network (CNN) pre-trained on the ImageNet classification task (Russakovsky et al., 2015) and extract the 4096-dimensional vector from the pre-softmax layer to represent each image. We opt for the VGG network (Simonyan and Zisserman, 2014) which, according to Kiela et al. (2016), has a slight edge on the two other alternatives – AlexNet (Krizhevsky et al., 2012) and GoogLeNet (Szegedy et al., 2015). We used the MMFeat toolkit (Kiela, 2016) to facilitate the process of image retrieval and CNN-based feature extraction.

Visual similarity. Because we retrieve more than one image per word, our visual representation of the word is a set of image embedding vectors. This allows for different visual similarity measures taking as input two sets of image embeddings (Kiela et al., 2015), given in Table 1.

3.3 Multi-Modal Representations

In order to compute multi-modal STS scores, one can combine linguistic and visual embeddings of words and sentences in a number of ways. Here, we explore three different levels of combining visual and linguistic representations, to which we refer as *early fusion*, *middle fusion*, and *late fusion*. We also experiment with selective

[2] Our choices were based on the findings from a recent systematic study on visual representation for multi-modal semantics (Kiela et al., 2016): (1) Visual representations from images obtained via Google and Bing image search are of similar quality. We opted for Bing for logistic reasons; (2) The performance of multi-modal models in semantic tasks typically saturates for n in the interval $[10, 20]$.

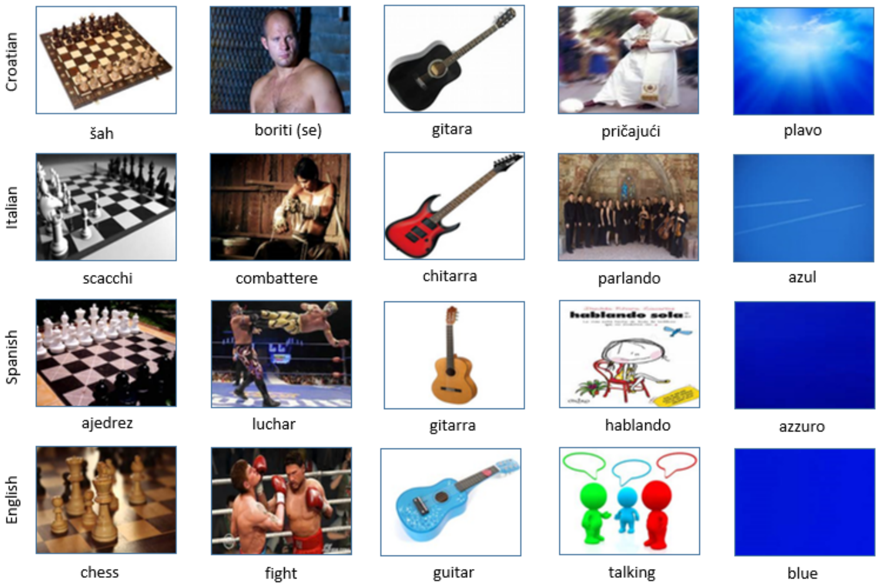

Figure 1: Example images (Bing image search)

Measure	Computation
Avg-Max	$\frac{1}{n}\sum_{e_i\in\mathcal{I}(w_i)}\max_{e_j\in\mathcal{I}(w_j)}\cos(e_i,e_j)$
Max-Max	$\max_{e_i\in\mathcal{I}(w_i)}\max_{e_j\in\mathcal{I}(w_j)}\cos(e_i,e_j)$
Sim-Avg	$\cos\left(\frac{1}{n}\sum_{e_i\in\mathcal{I}(w_i)} e_i, \frac{1}{n}\sum_{e_j\in\mathcal{I}(w_j)} e_j\right)$
Sim-Max	$\cos\left(\max_{el}\mathcal{I}(w_i), \max_{el}\mathcal{I}(w_j)\right)$

Table 1: Visual similarity measures for two sets of n images. $\mathcal{I}(w)$ is the set of image embeddings of word w. Function $\max_{el}$ computes the single element-wise maximum of the input vectors.

inclusion of visual information into the multi-modal representations, based on the measure of image dispersion.

Early fusion. This type of fusion concatenates (‖) the visual and linguistic embeddings of words:

$$e_{ef}(w) = e_v(w) \,\|\, e_t(w) \tag{1}$$

where $e_v(w)$ is the visual embedding of the word w, computed either by averaging or element-wise maxing of word's image embeddings, and $e_t(w)$ is the linguistic embedding of the word w. The similarity of two words is then simply computed as the cosine similarity of their fused multi-modal vectors.

Middle fusion. This type of fusion is performed at the sentence level. We first independently compute the aggregate linguistic representation and the aggregate visual representation for the whole sentence by averaging linguistic embeddings and visual embeddings of its words, respectively. The middle-fusion sentence representation is then the concatenation of the aggregated linguistic and visual representations. Let S be the set of words of a sentence. The middle-fusion sentence representation is then given as follows:

$$e_{mf}(S) = \left(\frac{1}{|S|}\sum_{w \in S} e_v(w)\right) \mid\mid \left(\frac{1}{|S|}\sum_{w \in S} e_t(w)\right) \tag{2}$$

Note that the middle-fusion representation of each sentence equals to averaging the early-fusion representations of its constituent words.

Late fusion. The late fusion combines the visual and linguistic signal at the level of similarity scores rather than at the embedding level. Thus, it may be applied both for computing word and sentence similarities. Let sim_v be the similarity measure (cf. Section 4) for two words or sentences computed using their visual representations, and let sim_t be their similarity computed using their linguistic representations. The late-fusion similarity is then computed as the linear combination of the uni-modal similarities, i.e., as $a \cdot sim_v + b \cdot sim_t$. The default late-fusion model uses $a = b = 0.5$.

Selective inclusion of visual information. Previous studies (Hill et al., 2013; Kiela et al., 2014) show that visual signal does not improve the semantic representation equally for all concepts. In fact, the inclusion of visual information deteriorates semantic representations for abstract concepts (e.g., *honesty, love, freedom*). In order to selectively include the visual information, we need a measure reflecting the quality of the visual signal. To this end, we use the *image dispersion* score (Kiela et al., 2014). A concept's image dispersion is the cosine distance between image embeddings, averaged over all pairs of images obtained for the concept w:

$$id(w) = \frac{1}{\binom{\mathcal{I}(w)}{2}} \sum_{\substack{e_i, e_j \in \mathcal{I}(w) \\ i \neq j}} 1 - \cos(e_i, e_j) \tag{3}$$

High image dispersion indicates that the images obtained for the concept are diverse. This means that the concept does not have a standard visual representation due to, e.g., its abstractness or its inherent polysemous nature.

We extend our middle-fusion and late-fusion models with selective inclusion of visual information. For the middle fusion, we measure the average image dispersion of all the words in a sentence. If the larger of the image dispersions of sentences in comparison scores above treshold τ,[3] we compare only the linguistic sentence embeddings. Otherwise, we compare the "middle-fused" multi-modal embeddings of the two sentences. For the late fusion, we compute the linear combination coefficients a and b as functions of image dispersions (the formula is given for words, but we also apply it to sentences in an analogous manner):

$$\begin{aligned} sim_{lf}(w_i, w_j) = (1 - id(w_i, w_j)) \cdot sim_v(w_i, w_j) \\ + id(w_i, w_j) \cdot sim_t(w_i, w_j) \end{aligned} \tag{4}$$

where $id(w_i, w_j)$ represents the larger of the image dispersions of the words w_i and w_j.[4]

4 Unsupervised STS Measures

In the previous section we explained the different levels at which we may combine visual and linguistic representations. However, we still have to define the actual STS measures that compute similarity scores for given pairs of sentences. We propose two simple unsupervised scores for measuring textual similarity. Both scores are agnostic of the actual modality used: this means that we can swap linguistic, visual, and multi-modal vectors as desired without altering the actual STS measure.

Optimal aligment similarity. Following the ideas from successful unsupervised STS models (Han et al., 2013; Sultan et al., 2014), we aim to align words between the two sentences at hand. Aiming to devise language-independent STS models (i.e., language-specific tools that could help better align the words are off-limits), we can resort to word similarity measures as the sole information source guiding the alignment process. This STS measure is based on the optimal alignment between the words of the two input sentences. Given the similarity scores for all pairs of words between the sentences S_1 and S_2, we are looking for an alignment $\{(w^i_{S_1}, w^i_{S_2})\}^N_{i=1}$ (N is the number of aligned pairs, equal to the number of words in the shorter of the sentences) that maximizes the sum of the pairwise similarities, i.e.:

$$\max_{\{w^i_{S_1}, w^i_{S_2}\}^N_{i=1}} \sum_{i=1}^{N} sim(w^i_{S_1}, w^i_{S_2}) \tag{5}$$

As this is a prototypical assignment problem, we find the optimal word alignment using the Hungarian algorithm (Kuhn, 1955), which provides the solution in polynomial time.[5] Because pairs of longer sentences will be assigned larger similarity scores on the account of more aligned word pairs simply due to length, we normalize the above sum of similarities with the length of each of the two input sentences,

[3] In all experiments, we set the treshold to the middle of the image dispersion range, i.e., $\tau = 0.5$.

[4] We also experimented with $b = 1$, but it yielded inferior performance. This implies that the contribution of the "useful" (i.e., non-dispersed) visual signal should outweigh the linguistic signal.

[5] The time complexity of the algorithm, also known as the Kuhn-Munkres algorithm, is $\mathcal{O}(n^3)$.

respectively. Finally, the optimal alignment similarity is the average of these two length-normalized similarity scores.

Aggregation similarity. In order to compute the aggregation similarity, we first compute the aggregate vector representation for each of the two sentences and then compare these aggregate sentence vectors. The aggregate vector representation of the sentence S is computed simply as the mean of the vectors of its words, i.e., $e(S) = \frac{1}{|S|}\sum_{w \in S} e(w)$. The aggregation similarity score is then computed as the cosine similarity between the aggregate vector representations of the two sentences, i.e., $sim_{agg} = \cos(e(S_1), e(S_2))$. Note that, for multi-modal STS models, the aggregation similarity based on early-fusion word vectors is equivalent to the similarity of middle-fused sentence representations.

5 Evaluation

In this section, we provide all details relevant to the evaluation of our unsupervised multi-modal STS models, from the description of datasets to the discussion of the experimental results.

Datasets. We use two different STS datasets to evaluate all models: we opt for very different STS datasets to gain more insight about the effectiveness of visually-informed STS models in different settings. The first dataset is the evaluation portion of the Microsoft Research video captions dataset (MSRVID) from the SemEval 2012 STS challenge Agirre et al. (2012). MSRVID consists of 750 pairs of short English sentences containing rather concrete concepts (people and animals performing simple actions, e.g., *"A woman is slicing onions"*). We couple the MSRVID dataset with the cross-lingual English-Spanish STS dataset (NEWS-16) from the SemEval 2016 STS shared task Agirre et al. (2016). NEWS-16 comprises 301 pairs of long sentences taken from news stories.

Considering (1) that MSRVID is a monolingual English dataset and NEWS-16 considers only one language pair and (2) that we aim to evaluate cross-lingual STS models on several language pairs, we derived other cross-lingual versions of these datasets. In addition to the standard monolingual English evaluation on the MSRVID dataset, we perform cross-lingual evaluations for three different language pairs: English-Spanish (EN-ES), English-Italian (EN-IT), and English-Croatian (EN-HR). We selected Spanish because of a readily available ES-EN NEWS-16 dataset and Italian because we had access to native speakers.

To test the claim that the proposed approach is language-independent, we also include an under-resourced language in the evaluation. As for Italian, we chose Croatian because we had access to a native speaker of that language. We created the additional cross-lingual datasets (all three language pairs for MSRVID; EN-IT and EN-HR for NEWS-16) by: (1) translating one of the sentences from each pair

	MSRVID		NEWS-16	
Language	ASL	AID	ASL	AID
English	3.42	0.51	17.0	0.68
Spanish	3.59	0.58	17.2	0.65
Italian	4.66	0.66	20.9	0.70
Croatian	3.54	0.70	17.7	0.71

Table 2: Statistics of the STS evaluation datasets.

to another language via Google translate and (2) having native speakers fix the machine translation errors.[6]

We depict the differences between the datasets in terms of the average sentence length in number of words (ASL) and the average image dispersion of words (AID) in Table 2, for all four languages. The average image dispersion is much lower on the MSRVID dataset (especially for English), which implies that the NEWS-16 dataset has larger portion of concepts that simply do not have a standardized visual representation. A closer manual inspection of the two datasets revealed that NEWS-16, besides having more abstract concepts than MSRVID, contains also a much larger number of polysemous words. This is not surprising considering the news-story origin of the sentences in NEWS-16. The differences in average image dispersion between the languages on the MSRVID dataset imply that Bing image search does not perform equally well for all languages.

Linguistic embeddings. We used the readily available word vectors for English (200-dimensional GloVe vectors trained on 6B tokens corpus), Spanish (300-dimensional Skip-Gram vectors trained on a 1.5B tokens corpus), and Italian (300-dimensional Skip-Gram vectors trained on a 2B tokens corpus). For Croatian, we trained 200-dimensional Skip-Gram embedding vectors on the 1.2B token version of the hrWaC corpus (Ljubešić and Erjavec, 2011).

To train the translation matrices, we selected the 4200 most frequent English words and translated them to the other three languages via Google translate, as done in prior work (Mikolov et al., 2013; Vulić et al., 2016). Native speakers of target languages fixed incorrect machine translations. We learned the optimal values of the translation matrices stochastically with the Adam algorithm (Kingma and Ba, 2014) on the 4000 word translation pairs. The obtained results of the evaluation of the translation quality on the remaining 200 test pairs – 58.8% P@5 for EN-ES, 56.3% for EN-IT, and 56.2% for EN-HR – are comparable to those reported in the original study from Mikolov et al. (2013).

STS Models in Evaluation. Our general approach includes several methods for measuring visual similarity (Table 1), different types of multi-modal information

[6] We make the STS datasets and the multi-modal STS code freely available at `http://tinyurl.com/jc8rd57`.

	MSRVid				News-16		
Model	EN-EN	EN-ES	EN-IT	EN-HR	EN-ES	EN-IT	EN-HR
Linguistic-only							
Txt-OA	74.9	57.3	50.6	55.3	82.7	79.2	78.8
Txt-Agg	74.7	54.9	42.9	51.1	57.3	48.1	54.5
Visual-only							
Vis-OA-Avg-Max	76.5	70.4	63.1	45.0	56.9	57.5	47.7
Vis-Agg-Sim-Avg	77.6	71.8	63.1	38.2	18.0	12.4	4.4
Multi-modal							
EF-OA-Avg	77.0	71.5	59.7	33.3	52.8	48.9	41.7
MF-Avg	77.8	72.0	63.8	38.9	19.1	14.8	1.3
LF-Word-OA	76.6	67.9	60.9	58.3	78.1	74.9	71.0
LF-Sent	80.8	**73.1**	**65.4**	59.2	78.0	74.0	71.3
Multi-modal with selective inclusion of visual information							
MF-Avg-ID	78.1	70.6	50.0	53.9	57.4	50.2	54.5
LF-Word-OA-ID	77.3	64.3	56.9	58.8	82.7	79.3	78.6
LF-Sent-ID	**81.0**	71.8	63.4	**61.0**	**83.1**	**79.6**	**79.5**

Table 3: STS performance on the MSRVid and News-16 datasets (Pearson ρ).

fusion (Section 2) and two STS measures (Section 4). For brevity, we present results only for the following models:

i) Linguistic-only models employ linguistic embeddings with optimal alignment or aggregation similarity (Txt-OA and Txt-Agg);

ii) Visual-only models use optimal alignment or aggregation similarity with the visual similarities from Table 1 that yield the best performance (Vis-OA-Avg-Max and Vis-Agg-Sim-Avg);

iii) Multi-modal models exploit both the linguistic and visual signal by combining early or middle fusion with the averaged image embedding (EF-OA-Avg and MF-Avg). Additionally, LF-Word-OA performs the late fusion at the word level with optimal alignment similarity, whereas LF-Sent model computes the average of the similarity scores produced by the best-performing linguistic-only model and the best-performing visual-only model on the respective dataset. For the last three models we also evaluate variants with image dispersion-based weighting (MF-Avg-ID, LF-Word-OA-ID and LF-Sent-ID).

Results and Discussion. Results using Pearson correlation between human and automatic similarity scores are shown in Table 3.

i) Multi-modal vs. uni-modal. The visual-only models tend to outperform the linguistic-only models on MSRVID. The multi-modal models further improve the performance of the visual-only models. We believe that this is the result of good visual representations we are able to obtain for concrete concepts, which are abundant in MSRVID. In the multi-modal landscape, the late fusion at the level of similarity scores (i.e., the LF-SENT model) seems to be the best way to combine visual and linguistic information.

The performance of the visual-only models on NEWS-16 is, however, much lower than the performance of the linguistic-only models. We believe that this is the direct consequence of obtaining rather noisy visual signal for the majority of concepts in this dataset. We observe that only 17.9% of English words from NEWS-16 have the image dispersion score below 0.5 (the statistics is 38.7% on MSRVID). The number is even lower for the other three languages. Therefore, the direct multi-modal models (i.e., without the selective inclusion of visual information) also perform worse than the linguistic-only models.

Aggregation-based models (TXT-OA, VIS-AGG-AVG, and MF-AVG) perform comparably to their respective optimal alignment counterparts (TXT-OA, VIS-OA-AVGMAX, and EF-OA-AVG) on MSRVID, but display drastically lower performances on NEWS-16. The explanation for this is rather intuitive – it is harder to aggregate the meaning of a sentence from the meaning of its words for long than for short sentences. On the other hand, by aligning pairs of words and accounting for the number of alignments, the optimal alignment similarity is not affected by the sentence length.

ii) Monolingual vs. cross-lingual. The performance gap on MSRVID in favor of visual-only and multi-modal models is significantly larger in the cross-lingual settings than in the monolingual English setting. On one hand, the cross-lingual linguistic-only models suffer from the imperfect mappings between monolingual embedding spaces. On the other hand, the visual signal seems not to deteriorate as much in quality for other languages. The performance of visual-only and multi-modal models is naturally lower for language pairs with languages for which more dispersed visual signals are used (IT and HR, see the scores in Table 2).

iii) Selective inclusion of visual information. The models that selectively include visual information do not consistently improve the results of the direct multi-modal models on MSRVID. Since the impact of the visual signal is scaled according to the the larger of the image dispersions, the selection model might discard useful visual information for a word/sentence on one side, because of the poor visual information on the other side. On the other hand, we have less informative visual representations across the board on NEWS-16: here, a selective inclusion of visual information in the similarity-level late fusion model (LF-SENT-ID) has a slight edge on the linguistic-only model (TXTOA). This improvement is small due to a shortage of concepts with sufficiently coherent visual representations in NEWS-16. This suggests that more sophisticated image extraction and content selection methods are required in future work.

iv) Comparison with state-of-the-art. For the monolingual English MSRVID dataset and the cross-lingual EN-ES NEWS-16 dataset, we also compare our results with the best-performing systems from the corresponding SemEval shared tasks. Šarić et al. (2012) reach 88% correlation on MSRVID, which is 7% better than our LF-SENT-ID model. The system of Brychcín and Svoboda (2016) achieves the correlation score of 91% on the EN-ES NEWS-16, 8% above the performance of LF-SENT-ID. We find these gaps in performance to be reasonably low, given that both state-of-the-art systems use a set of expensive language-specific tools (e.g., dependency parsers, NER). Moreover, the system of Šarić et al. (2012) is supervised, whereas the Brychcín and Svoboda (2016) require a full-blown MT system.

6 Conclusion

Semantic representations of meaning that combine signals from visual and linguistic input tend to outperform uni-modal models exploiting only linguistic information across a variety of semantic tasks. In this work, we have investigated the effects of leveraging visual information in measuring semantic textual similarity (STS) of short texts. We have retrieved images for single-word concepts and extracted visual embeddings via a transferred deep CNN (VGG). We fused visual and linguistic signals at three different levels of granularity and plugged the variety of representations (linguistic, visual, multi-modal) into two simple unsupervised STS measures. In addition, we investigated the selective inclusion of visual information in multi-modal STS models based on image dispersion.

Experimental results suggest that the visual-only models outperform the linguistic-only models by a wide margin on datasets containing a large number of concrete concepts, especially in the cross-lingual setting. Moreover, the multi-modal STS models with selective inclusion of visual information seem to provide a performance boost even for the dataset for which the visual signal is dispersed.

The experiments show that the performance of visual-only and multi-modal models highly depends on the quality (dispersion) of images obtained for the concepts. Our future efforts will thus aim to devise methods for extracting and selecting better visual representations for visually dispersed concepts (e.g., by clustering the retrieved images by similarity and considering only images from the largest cluster).

References

Agirre, E., C. Banea, C. Cardiec, D. Cerd, M. Diabe, A. Gonzalez-Agirrea, W. Guof, I. Lopez-Gazpioa, M. Maritxalara, R. Mihalcea, et al. (2015). Semeval-2015 Task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *SemEval*, pp. 252–263.

Agirre, E., C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe (2016). Semeval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval*, pp. 497–511.

Agirre, E., M. Diab, D. Cer, and A. Gonzalez-Agirre (2012). Semeval-2012 Task 6: A pilot on semantic textual similarity. In *SemEval*, pp. 385–393.

Aziz, W. and L. Specia (2011). Fully automatic compilation of Portuguese-English and Portuguese-Spanish parallel corpora. In *STIL*, pp. 234–238.

Bär, D., C. Biemann, I. Gurevych, and T. Zesch (2012). UKP: Computing semantic textual similarity by combining multiple content similarity measures. In **SEM*, pp. 435–440.

Bergsma, S. and B. V. Durme (2011). Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI*, pp. 1764–1769.

Bergsma, S. and R. Goebel (2011). Using visual information to predict lexical preference. In *RANLP*, pp. 399–405.

Bruni, E., N. Tran, and M. Baroni (2014). Multimodal distributional semantics. *Journal of Artiifical Intelligence Research 49*, 1–47.

Brychcín, T. and L. Svoboda (2016). UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In *SemEval*, pp. 588–594.

Franco-Salvador, M., P. Gupta, and P. Rosso (2013). Cross-language plagiarism detection using a multilingual semantic network. In *ECIR*, pp. 710–713.

Han, L., A. Kashyap, T. Finin, J. Mayfield, and J. Weese (2013). UMBC EBIQUITY-CORE: Semantic textual similarity systems. In *SemEval*, pp. 44–52.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena 42*(1-3), 335–346.

Hill, F., D. Kiela, and A. Korhonen (2013). Concreteness and corpora: A theoretical and practical analysis. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 75–83.

Islam, A. and D. Inkpen (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD) 2*(2), 10.

Jimenez, S. (2016). SERGIOJIMENEZ at SemEval-2016 Task 1: Effectively combining paraphrase database, string matching, WordNet, and word embedding for semantic textual similarity. In *SemEval*, pp. 749–757.

Kiela, D. (2016). MMFeat: A toolkit for extracting multi-modal features. In *ACL (Demos)*, pp. 55–60.

Kiela, D. and L. Bottou (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, pp. 36–45.

Kiela, D., F. Hill, A. Korhonen, and S. Clark (2014). Improving multi-modal representations using image dispersion: Why less is sometimes mor. In *ACL*, pp. 835–841.

Kiela, D., L. Rimell, I. Vulić, and S. Clark (2015). Exploiting image generality for lexical entailment detection. In *ACL*, pp. 119–124.

Kiela, D., A. L. Verő, and S. Clark (2016). Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. In *EMNLP (to appear)*.

Kiela, D., I. Vulić, and S. Clark (2015). Visual bilingual lexicon induction with transferred ConvNet features. In *EMNLP*, pp. 148–158.

Kingma, D. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). ImageNet classification with deep convolutional neural networks. In *NIPS*, pp. 1097–1105.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly 2*(1-2), 83–97.

Lakoff, G. and M. Johnson (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*.

Ljubešić, N. and T. Erjavec (2011). hrWaC and siWaC: Compiling Web corpora for Croatian and Slovene. In *TSD*, pp. 395–402.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science 59*(1), 617–645.

Mikolov, T., Q. V. Le, and I. Sutskever (2013). Exploiting similarities among languages for machine translation. *CoRR abs/1309.4168*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119.

Oliva, J., J. I. Serrano, M. D. del Castillo, and Á. Iglesias (2011). Symss: A syntax-based measure for short-text semantic similarity. *Data & Knowledge Engineering 70*(4), 390–405.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *EMNLP*, pp. 1532–1543.

Potthast, M., A. Barrón-Cedeño, B. Stein, and P. Rosso (2011). Cross-language plagiarism detection. *Language Resources and Evaluation 45*(1), 45–62.

Resnik, P. and N. A. Smith (2003). The Web as a parallel corpus. *Computational Linguistics 29*(3), 349–380.

Roller, S. and S. Schulte Im Walde (2013). A multimodal LDA model integrating textual, cognitive and visual modalities. In *EMNLP*, pp. 1146–1157.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision 115*(3), 211–252.

Šarić, F., G. Glavaš, M. Karan, J. Šnajder, and B. D. Bašić (2012). Takelab: Systems for measuring semantic text similarity. In *SemEval*, pp. 441–448.

Shutova, E., D. Kiela, and J. Maillard (2016). Black holes and white rabbits: Metaphor identification with visual features. In *NAACL-HLT*, pp. 160–170.

Silberer, C. and M. Lapata (2012). Grounded models of semantic representation. In *EMNLP*, pp. 1423–1433.

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

Specia, L., S. Frank, K. Sima'an, and D. Elliott (2016). A shared task on multimodal machine translation and crosslingual image description. In *WMT*, pp. 543–553.

Sultan, M. A., S. Bethard, and T. Sumner (2014). DLS@CU: Sentence similarity from word alignment. In *SemEval*, pp. 241–246.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2015). Going deeper with convolutions. In *CVPR*, pp. 1–9.

Vulić, I., D. Kiela, S. Clark, and M.-F. Moens (2016). Multi-modal representations for improved bilingual lexicon learning. In *ACL*, pp. 188–194.