24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Evaluating Multimodal Representations on Visual Semantic Textual Similarity

Oier Lopez de Lacalle[1] and **Ander Salaberria**[1] and **Aitor Soroa**[1] and **Gorka Azkune**[1] and **Eneko Agirre**[1]

Abstract.

The combination of visual and textual representations has produced excellent results in tasks such as image captioning and visual question answering, but the inference capabilities of multimodal representations are largely untested. In the case of textual representations, inference tasks such as Textual Entailment and Semantic Textual Similarity have been often used to benchmark the quality of textual representations. The long term goal of our research is to devise multimodal representation techniques that improve current inference capabilities. We thus present a novel task, Visual Semantic Textual Similarity (vSTS), where such inference ability can be tested directly. Given two items comprised each by an image and its accompanying caption, vSTS systems need to assess the degree to which the captions in context are semantically equivalent to each other. Our experiments using simple multimodal representations show that the addition of image representations produces better inference, compared to text-only representations. The improvement is observed both when directly computing the similarity between the representations of the two items, and when learning a siamese network based on vSTS training data. Our work shows, for the first time, the successful contribution of visual information to textual inference, with ample room for benchmarking more complex multimodal representation options.

1 Introduction

Language understanding is a task proving difficult to automatize, because, among other factors, much of the information that is needed for the correct interpretation of an utterance is not explicit in text [5]. This contrasts with how natural is language understanding for humans, who can cope easily with information absent in text, using common sense and background knowledge like, for instance, typical spatial relations between objects. From another perspective, it is well-known that the visual modality provides complementary information to that in the text. In fact, recent advances in deep learning research have led the field of computer vision and natural language processing to significant progress in tasks that involve visual and textual understanding. Tasks that include visual and textual content include Image Captioning [11], Visual Question Answering [1], and Visual Machine Translation [13], among others.

On the other hand, progress in language understanding has been driven by datasets which measure the quality of sentence representations, specially those where inference tasks are performed on top of sentence representations, including textual entailment [8, 4] and semantic textual similarity (STS). In STS [6], for instance, pairs of sentences have been annotated with similarity scores, with top scores for semantically equivalent sentences and bottom scores for completely unrelated sentences. STS provides a unified framework for extrinsic evaluation of multiple semantic aspects such as compositionality and phrase similarity. Contrary to related tasks, such as textual entailment and paraphrase detection, STS incorporates the notion of graded semantic similarity between the pair of textual sentences and is symmetric.

Figure 1. A sample with two items, showing the influence of images when judging the similarity between two captions. While the similarity for the captions alone was annotated as low (1.8), when having access to the images, the annotators assigned a much higher similarity (4). The similarity score ranges between 0 and 5.

In this paper we extend STS to the visual modality, and present Visual Semantic Textual Similarity (vSTS), a task and dataset which allows to study whether better sentence representations can be built when having access to the corresponding images, in contrast with having access to the text alone. Similar to STS, annotators were asked to score the similarity between two items, but in this case each item comprises an image and a textual caption. Systems need to predict the human score. Figure 1 shows an instance in the dataset, with similarity scores in the captions. The example illustrates the need to re-score the similarity values, as the text-only similarity is not applicable to the multimodal version of the dataset: the annotators return a low similarity when using only text, while, when having access to the corresponding image, they return a high similarity. Although a dataset for multimodal inference exists (visual textual entailment [39]) that dataset reused the text-only inference labels.

The vSTS dataset aims to become a standard benchmark to test the contribution of visual information when evaluating the similarity of sentences and the quality of multimodal representations, allowing to test the complementarity of visual and textual information for improved language understanding. Although multimodal tasks such as image captioning, visual question answering and visual machine translation already show that the combination of both modalities can be effectively used, those tasks do not separately benchmark the inference capabilities of multimodal visual and textual representations.

[1] University of the Basque Country, Spain, emails: {oier.lopezdelacalle, ander.salaberria, a.soroa, gorka.azcune, e.agirre}@ehu.eus

We evaluate a variety of well-known textual, visual and multimodal representations in supervised and unsupervised scenarios, and systematically explore if visual content is useful for sentence similarity. For text, we studied pre-trained word embeddings such as GloVe [27], pre-trained language models like GPT-2 and BERT [12, 30], sentence representations fine-tuned on an entailment task like USE [7], and textual representations pre-trained on a multimodal caption retrieval task like VSE++ [14]. For image representation we use a model pre-trained on Imagenet (ResNet [17]). In order to combine visual and textual representations we used concatenation and learn simple projections. Our experiments show that the text-only models are outperformed by their multimodal counterparts when adding visual representations, with up to 24% error reduction.

Our contributions are the following: (1) We present a dataset which allows to evaluate visual/textual representations on an inference task. The dataset is publicly available under a free license[2]. (2) Our results show, for the first time, that the addition of image representations allows better inference. (3) The best text-only representation is the one fine-tuned on a multimodal task, VSE++, which is noteworthy, as it is better than a textual representation fine-tuned in a text-only inference task like USE. (4) The improvement when using image representations is observed both when computing the similarity directly from multimodal representations, and also when training siamese networks. At the same time the improvement holds for all textual representations, even those fine-tuned on a similarity task.

2 Related Work

The task of Visual Semantic Textual Similarity stems from previous work on textual inference tasks. In textual entailment, given a textual premise and a textual hypothesis, systems need to decide whether the first entails the second, they are in contradiction, or none of the previous [8]. Popular datasets include the Stanford Natural Language Inference dataset [4]. As an alternative to entailment, STS datasets comprise pairs of sentences which have been annotated with similarity scores. STS systems are usually evaluated on the STS benchmark dataset [6][3]. In this paper we present an extension of STS, so we present the task in more detail in the next section.

Textual entailment has been recently extended with visual information. A dataset for **visual textual entailment** was presented in [39]. Even if the task is different from the text-only counterpart, they reused the text-only inference ground-truth labels without re-annotating them. In fact, they annotate a small sample to show that the labels change. In addition, their dataset tested pairs of text snippets referring to a single image, and it was only useful for testing grounding techniques, but not to measure the complementarity of visual and textual representations. The reported results did not show that grounding improves results, while our study shows that the inference capabilities of multimodal visual and textual representations improve over text-only representations. In related work, [40] propose visual entailment, where the premise is an image and the hypothesis is textual. The chosen setting does not allow to test the contribution of multimodal representationn with respect to unimodal ones.

The complementarity of visual and text representations for improved language understanding was first proven on word representations, where word embeddings were combined with visual or perceptual input to produce multimodal representations [15]. The task of Visual Semantic Textual Similarity is also related to other multimodal tasks such as Image Captioning [3, 16], Text-Image Retrieval [2, 28] and Visual Question Answering [1].

Image Captioning is a task that aims to generate a description of a given image. The task is related to ours in that it is required an understanding of the scene depicted in the image, so the system can generate an accurate description of it. Unlike vSTS, image captioning is a generation task in which evaluation is challenging and unclear, as the defined automatic metrics are somewhat problematic [36]. On the other hand, **Text-Image Retrieval** task requires to find similarities and differences of the items in two modalities, so we can distinguish relevant and irrelevant texts and images regarding the query. Apart from not checking inference explicitly, the other main difference with regards to vSTS is that, in retrieval, items are ranked from most to least similar, whereas the vSTS task consists on scoring an accurate real valued similarity. A comprehensive overview is out of the scope, and thus we focus on the most related vision and language tasks. We refer the reader to [26] for a survey on vision and language research.

Many of these tasks can be considered as extensions of previously existing NLP taks. For instance, Image Captioning can be seen as an extension of conditional language modeling [10] or natural language generation [32], whereas Visual Question Answering is a natural counterpart of the traditional Question Answering in NLP.

Regarding multimodal and unimodal **representation learning**, convolutional neural networks (CNN) have become the standard architecture for generating representations for images [23]. Most of these models learn transferable general image features in tasks such as image classification, and detection, semantic segmentation, and action recognition. Most used transferable global image representations are learned with deep CNN architectures such as AlexNet [22], VGG [33], Inception-v3 [34], and ResNet [17] using large datasets such as ImageNet [11], MSCOCO [24] and Visual Genome [21]. Recently, Graph Convolution Networks (GCN) showed to be promising way to distill multiple input types multimodal representations [41].

Language representation is mostly done with pretrained word embeddings like Glove [27] and sequence learning techniques such as Recurrent Neural Networks (RNN) [18]. Recently, self-attention approaches like Transformers [37] provided transferable models (BERT, GPT-2, among others [12, 30]) that significantly improve many state-of-the-art tasks in NLP. Alternatively, sentence representations have been fine-tuned on an entailment task [7]. We will present those used in our work in more detail below.

3 The Visual STS Dataset

STS assesses the degree to which two sentences are semantically equivalent to each other. The annotators measure the similarity among sentences, with higher scores for more similar sentences. The annotations of similarity were guided by the scale in Table 1, ranging from 0 for no meaning overlap to 5 for meaning equivalence. Intermediate values reflect interpretable levels of partial overlap in meaning.

Similarity definitions:
5: Completely equivalent: They mean the same thing.
4: Mostly equivalent: Some unimportant details differ.
3: Roughly equivalent: Some important information differs/missing.
2: Not equivalent but share some details.
1: Not equivalent but on the same topic.
0: Completely dissimilar.

Table 1. Similarity scores with the definition of each ordinal value. Definitions are the same as used in STS datasets [6]

[2] https://oierldl.github.io/vsts/

[3] See for instance recent models evaluated on STS benchmark http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

In this work, we extend the STS task with images, providing visual information that models use, and assess how much visual content can contribute in a language understanding task. The input of the task now consists of two items, each comprising an image and its corresponding caption. In the same way as in STS, systems need to score the similarity of the sentences with the help of the images. Figure 1 shows an example of an instance in the dataset.

In previous work reported in a non-archival workshop paper [9], we presented a preliminary dataset which used the text-only ground-truth similarity scores. The 819 pairs were extracted from a subset of the STS benchmark, more specifically, the so called STS-images subset, which contains pairs of captions with access to images from PASCAL VOC-2008 [31] and Flickr-8K [19]. Our manual analysis, including examples like Figure 1, showed that in many cases the text-only ground truth was not valid, so we decided to re-annotated the dataset but showing the images in addition to the captions (the methodology is identical to the AMT annotation method mentioned below). The correlation of the new annotations with regard to the old ones was high (0.9ρ) showing that the change in scores was not drastic, but that annotations did differ. The annotators tended to return higher similarity scores, as the mean similarity score across the dataset increased from 1.7 to 2.1. The inter-tagger correlation was comparable to the text-only task, showing that the new annotation task was well-defined.

From another perspective, the fact that we could only extract 819 pairs from existing STS datasets showed the need to sample new pairs from other image-caption datasets. In order to be effective in measuring the quality of multimodal representations, we defined the following desiderata for the new dataset: (1) Following STS datasets, the similarity values need to be balanced, showing a uniform distribution; (2) Paired images have to be different to avoid making the task trivial, as hand analysis of image-caption datasets showed that two captions of the same image tended to be paraphrases of each other; (3) The images should not be present in more than one instance, to avoid biases in the visual side; (4) It has to contain a wide variety of images so we can draw stronger conclusions. The preliminary dataset fulfilled 2 and 3, but the dataset was skewed towards low similarity values and the variety was limited.

3.1 Data Collection

The data collection of sentence-image pairs comprised several steps, including the selection of pairs to be annotated, the annotation methodology, and a final filtering stage.

1. Sampling data for manual annotation. We make use of two well-known image-caption datasets. On one hand, Flickr30K dataset [29] that has about 30K images with 5 manually generated captions per image. On the other hand, we use the Microsoft COCO dataset [24], which contains more than 120K images and 5 captions per image. Using both sources we hope to cover a wide variety of images.

In order to select pairs of instances, we did two sampling rounds. The goal of the first run is to gather a large number of varied image pairs with their captions which contain interesting pairs. We started by sampling images. We then combined two ways of sampling pairs of images. In the first, we generated pairs by sampling the images randomly. This way, we ensure higher variety of paired scenes, but presumably two captions paired at random will tend to have very low similarity. In the second, we paired images taking into account their visual similarity, ensuring the selection of related scenes with a higher similarity rate. We used the cosine distance of the top-layer of a pretrained ResNet-50 [17] to compute the similarity of images. We collected an equal number of pairs for the random and visual similarity strategy, gathering, in total, $155,068$ pairs. As each image has 5 captions, we had to select one caption for each image, and we decided to select the two captions with highest word overlap. This way, we get more balanced samples in terms of caption similarity[4].

The initial sampling created thousands of pairs that were skewed towards very low similarity values. Given that manual annotation is a costly process, and with the goal of having a balanced dataset, we used an automatic similarity system to score all the pairs. This text-only similarity system is an ensemble of feature-based machine learning systems that uses a large variety of distance and machine-translation based features. The model was evaluated on a subset of STS benchmark dataset [6] and compared favorably to other baseline models. As this model is very different from current deep learning techniques, it should not bias the dataset sampling in a way which influences current similarity systems.

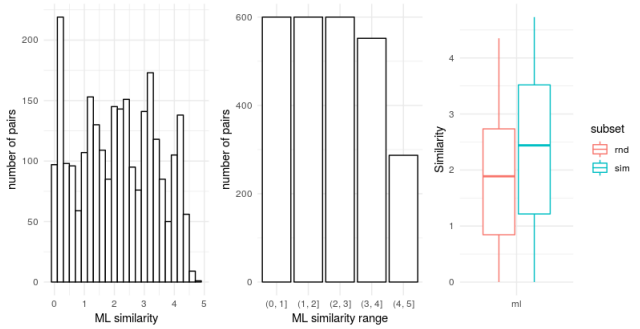

Figure 2. Histograms of the similarity distribution in the 2639 sample, according to the automatic text-only system (left and middle plots), and the distribution of the similarity of each sampling strategy (**rnd** stands for random image sampling and **sim** stands for image similarity driven sampling).

The automatic scores were used to sample the final set of pairs as follows. We defined five similarity ranges ($(0,1], \ldots, (4,5]$) and randomly selected the same amount of pairs from the initial paired sample. We set a sampling of maximum 3000 instances (i.e 600 instances per range). Given the fact that the high similarity range had less than 600 instances, we collected a total of 2639 potential text-image candidate pairs for manual annotation. Figure 2 shows the proposed methodology can sample approximately a uniform distribution with the exception of the higher similarity values (left and middle plots). In addition, we show that the lower predicted similarities are mainly coming from random sampling, whereas, as expected, the higher ones come from similar images.

2. Manual annotations. In order to annotate the sample of 2639 pairs, we used Amazon Mechanical Turk (AMT). Crowdworkers followed the same instructions of previous STS annotation campaigns [6], very similar to those in Table 1. Annotators needed to focus on textual similarity with the aid of aligned images. We got up to 5 scores per item, and we discarded annotators that showed low correlation with the rest of the annotators ($\rho < 0.75$). In total 56 annotators took part. On average each crowdworker annotated 220 pairs, where the amounts ranged from 19 to 940 annotations.

[4] We tried random sampling over captions too, but we ended up with a more unbalanced selection.

Regardless the annotation amounts, most of the annotators showed high correlations with the rest of the participants. We computed the annotation correlation by aggregating the individual Pearson correlation with averaged similarity of the other annotators. The annotation shows high correlation among the crowdworkers ($\rho = 0.89 \pm 0.01$) comparable to that of text-only STS datasets.

	#Pairs	Mean	Median	STD	#Zeros
Item similarity	2639	1.96	1.80	1.65	549
Item disagreement	2639	0.60	0.55	0.45	724

Table 2. Overall item similarity and disagreement of the AMT annotations.

Table 2 shows the average **item similarity** and **item disagreement** in the annotation. We defined item disagreement as the standard deviation of the annotated similarity value. The low average similarity can be explained by the high number of zero-similarity pairs. Item disagreement is moderately low (about 0.6 points out of 5) which is in accordance with the high correlation between the annotators.

3. Selection of difficult examples. In preliminary experiments, the evaluation of two baseline models, word overlap and the ensemble system mentioned before, showed that the sampling strategy introduced a large number of trivial examples. For example, the word overlap system attained 0.83 ρ. This high correlation could be the result of using word-overlap in the first sampling round. In order to create a more challenging dataset where to measure the effectiveness of multimodal representations, we defined the **easiness** metric to filter out some of the easy examples from the annotated dataset.

We defined easiness as an amount of discrepancy provided by an example regarding the whole dataset. Taking the inner product of the Pearson correlation formula as basis, we measure the easiness of an annotated example i as follows:

$$e_i = \left(\frac{o_i - \overline{o}}{s_o}\right)\left(\frac{gs_i - \overline{gs}}{s_{gs}}\right) \quad (1)$$

where o_i is the word-overlap similarity of the i-th pair, $\overline{o}$ is the mean overlap similarity in the dataset, and s_o is the standard deviation. Similarly, variable gs_i is the gold-standard value of the i-th pair, and $\overline{gs}$ and s_{gs} are the mean and standard deviation of gold values in the dataset, respectively. We removed 30% of the *easiest* examples and create a more challenging dataset of 1858 pairs, reducing ρ to 0.57 for the word-overlap model, and to 0.66 ρ (from 0.85) for the ML based approach.

3.2 Dataset Description

The full dataset comprises both the sample mentioned above and the 819 pairs from our preliminary work, totalling 2677 pairs. Figure 3 shows the final item similarity distribution. Although the distribution is skewed towards lower similarity values, we consider that all the similarity ranges are sufficiently well covered.

Average similarity of the dataset is 1.9 with a standard deviation of 1.36 points. The dataset contains 335 zero-valued pairs out of the 2677 instances, which somehow explains the lower average similarity.

4 Evaluation of Representation Models

The goal of the evaluation is to explore whether representation models can have access to images, instead of text alone, have better inference abilities. We consider the following models.

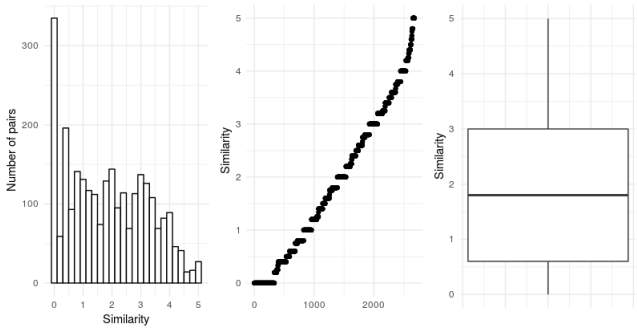

Figure 3. Similarity distribution of the visual STS dataset. Plots show three views of the data. Histogram of the similarity distribution of ground-truth values (left plot), sorted pairs according to their similarity (middle) and boxplot of the similarity values (right).

ResNet [17] is a deep network of 152 layers in which the residual representation functions are learned instead of learning the signal representation directly. The model is trained over 1.2 million images of ImageNet, the ILSRVC subset of 1000 image categories. We use the top layer of a pretrained ResNet-152 model to represent the images associated to text. Each image is represented with a vector of 2048 dimensions.

GloVe. The Global Vector model [27] is a log-linear model trained to encode semantic relationships between words as vector offsets in the learned vector space, combining global matrix factorization and local context window methods. Since GloVe is a word-level vector model, we build sentence representations with the mean of the vectors of the words composing the sentence. The pre-trained model from GloVe considered in this paper is the 6B-300d, with a vocabulary of 400k words, 300 dimension vectors and trained on a dataset of 6 billion tokens.

BERT. The Bidirectional Encoder Representations from Transformer [12] implements a novel methodology based on the so-called *masked language model*, which randomly masks some of the tokens from the input, and predicts the original vocabulary id of the masked word based only on its context. The BERT model used in our experiments is the BERT-Large Uncased (24-layer, 1024-hidden, 16-heads, 340M parameters). In order to obtain the sentence-level representation we extract the token embeddings of the last layer and compute the mean vector, yielding a vector of 1024 dimensions.

GPT-2. The Generative Pre-Training-2 model[30] is a language model based on the transformer architecture, which is trained on the task of predicting the next word, given all the previous words occurring in some text. In the same manner to BERT and GloVe, we extract the token embeddings of the last layer and compute the mean vector to obtain the sentence-level representation of 768 dimensions. The GPT-2 model used in our experiments was trained on a very large corpus of about 40 GB of text data with 1.5 billion parameters.

USE. The Universal Sentence Encoder [7] is a model for encoding sentences into embedding vectors, specifically designed for transfer learning in NLP. Based on a deep averaging network encoder, the model is trained for varying text lengths, such as sentences, phrases or short textbfs, and in a variety of semantic tasks including STS. The encoder returns the vector of the sentence with 512 dimensions.

VSE++. The Visual-Semantic Embedding [14] is a model trained for image-caption retrieval. The model learns a joint space of aligned images and captions. The model is an improvement of the original introduced by [20], and combines a ResNet-152 over images with a bidirectional Recurrent Neural Network (GRU) over the sentences.

Texts and images are projected onto the joint space, obtaining representations of 1024 dimension both for images and texts. We used projections of images and texts in our experiments. The VSE++ model used in our experiments was pre-trained on the Microsoft COCO dataset [24] and the Flickr30K dataset [29]. Table 3 summarizes the sentence and image representations used in the evaluation.

Model	Modality	dimensions
RESNET	Image	2048
VSE++(IMG)	Image	1024
GLOVE	Text	300
BERT	Text	1024
GPT-2	Text	768
USE	Text	512
VSE++(TEXT)	Text	1024
CONCAT	multimodal	-
PROJECT	multimodal	-

Table 3. Summary of the text and image representation models used.

4.1 Experiments

Experimental Setting. We split the vSTS dataset into training, validation and test partitions sampling at random and preserving the overall score distributions. In total, we use 1338 pairs for training, 669 for validation, and the rest of the 670 pairs were used for the final testing. Similar to the STS task, we use the Pearson correlation coefficient (ρ) as the evaluation metric of the task.

STS models. Our goal is to keep similarity models as simple as possible in order to directly evaluate textual and visual representations and avoid as much as possible the influence of the parameters that intertwine when learning a particular task. We defined two scenarios: the supervised and the unsupervised scenarios.

In the **supervised scenario** we train a Siamese Regression model in a similar way presented in [35]. Given a sentence/image pair, we wish to predict a real-valued similarity in some range $[1, K]$, being $K = 5$ in our experiments. We first produce sentence/image representations h_L and h_R for each sentence in the pair using any of the unimodal models described above, or using a multimodal representations as explained below. Given these representations, we predict the similarity score o using a regression model that takes both the distance and angle between the pair (h_L, h_R):

$$h_x = h_L \odot h_R, \quad (2)$$
$$h_+ = |h_L - h_R|, \quad (3)$$
$$h_s = \sigma(W^{(h)}[h_x, h_+] + b^{(h)}), \quad (4)$$
$$o = W^{(o)}h_s + b^{(o)} \quad (5)$$

Note that the distance and angle concatenation ($[h_x, h_+]$) yields a $2 * d$-dimensional vector. The resulting vector is used as input for the non-linear hidden layer (h_s) of the model. Contrary to [35], we empirically found that the estimation of a continuous value worked better than learning a softmax distribution over $[1, K]$ integer values. The loss function of our model is the Mean Square Error (MSE), which is the most commonly used regression loss function.

In the **unsupervised scenario** similarity is computed as the cosine of the produced h_L and h_R sentence/image representations.

Multimodal representation. We combined textual and image representations in two simple ways. The first method is concatenation of the text and image representation (CONCAT). Before concatenation we applied the L2 normalization to each of the modalities. The second method it to learn a common space for the two modalities before concatenation (PROJECT).

$$h_1 = \sigma(W^{(1)}m_1, +b^{(1)}), \quad (6)$$
$$h_2 = \sigma(W^{(2)}m_2, +b^{(2)}), \quad (7)$$
$$h_m = [h_1, h_2] \quad (8)$$

The projection of each modality learns a space of d-dimensions, so that $h_1, h_2 \in \mathbb{R}^d$. Once the multimodal representation is produced (h_m) for the left and right pairs, vectors are directly plugged into the regression layers. Projections are learned *end-to-end* with the regression layers and the MSE as loss function.

Hyperparameters and training details. We use the validation set to learn parameters of the supervised models, and to carry an exploration of the hyperparameters. We train each model a maximum of 300 epochs and apply *early-stopping* strategy with a *patience* of 25 epochs. For early stopping we monitor MSE loss value on validation. For the rest, we run a grid search for selecting the rest of the hyperparameter values. We explore *learning rate* values (0.0001, 0.001, 0.01, 0.05), *L2 regularization* weights (0.0, 0.0001, 0.001, 0.01), and different hidden layer (h_s) dimensions (50,100, 200, 300). In addition, we activate and deactivate batch normalization in each layer for each of the hyperparameter selection.

4.2 Results

The unsupervised scenario. Table 4 reports the results using the item representations directly. We report results over train and dev partitions for completeness, but note that none of them was used to tune the models. As it can be seen, multimodal representations consistently outperform their text-only counterparts. This confirms that, overall, visual information is helpful in the semantic textual similarity task and that image and sentence representation are complementary. For example, the BERT model improves more than 13 points when visual information provided by the RESNET is concatenated. GLOVE shows a similar or even larger improvement, with similar trends for USE and VSE++(TEXT)[5].

Although VSE++(IMG) shows better performance than RESNET when applying them alone, further experimentation showed lower complementarity when combining with textual representation (e.g. 0.807ρ in test combining textual and visual modalities of VSE++). This is something expected as VSE++(IMG) is pre-trained along with the textual part of the VSE++ model on the same task. We do not show the combinations with VSE++(IMG) due to the lack of space.

Interestingly, results show that images alone are valid to predict caption similarity (0.627ρ in test). Actually, in this experimental setting RESNET is on par with BERT, which is the best purely unsupervised text-only model. Surprisingly, GPT-2 representations are not useful for text similarity tasks. This might be because language models tend to forget past context as they focus on predicting the next token [38]. Due to the low results of GPT-2 we decided not to combine it with RESNET.

[5] VSE++ +RESNET in the table.

Model	Modality	train ρ	dev ρ	test ρ
GLOVE	text	0.576	0.580	0.587
BERT	text	0.641	0.593	0.612
GPT-2	text	0.198	0.241	0.210
USE	text	0.732	0.747	0.720
VSE++(TEXT)	text	0.822	**0.812**	**0.803**
RESNET	image	0.638	0.635	0.627
VSE++(IMG)	image	0.677	**0.666**	**0.662**
GLOVE+RESNET	mmodal	0.736	0.732	0.730
BERT+RESNET	mmodal	0.768	0.747	0.745
USE+RESNET	mmodal	0.799	0.806	0.787
VSE++ +RESNET	mmodal	0.846	**0.837**	**0.826**

Table 4. The unsupervised scenario: train, validation and test results of the unsupervised models.

The supervised scenario. Table 5 show a similar pattern to that in the the unsupervised setting. Overall, models that use a conjunction of multimodal features significantly outperform unimodal models, and this confirms, in a more competitive scenario, that adding visual information helps learning easier the STS task. The gain of multimodal models is considerable compared to the text-only models. The most significant gain is obtained when GLOVE features are combined with RESNET. The model improves more than 15.0 points. In this case, the improvement over BERT is lower, but still considerable with more than 4.0 points.

In the same vein as in the unsupervised scenario, features obtained with a RESNET can be as competitive as some text based models (e.g. BERT). GPT-2, as in the unsupervised scenario, does not produce useful representations for semantic similarity tasks. Surprisingly, the regression model with GPT-2 features is not able to learn anything in the training set. As we did in the previous scenario, we do not keep combining GPT-2 with visual features.

Multimodal version of VSE++ and USE[6] are the best model among the supervised approaches. Textual version of USE and VSE++ alone obtain very competitive results and outperforms some of the multimodal models (the concatenate version of GLOVE and BERT with RESNET). Results might indicate that text-only with sufficient training data can be on par with multimodal models, but, still, when there is data scarcity, multimodal models can perform better as they have more information over the same data point.

Comparison between projected and concatenated models show that projected models attain slightly better results in two cases, but the best overall results are obtained when concatenating VSE++(TEXT) with RESNET. Although concatenation proofs to be a hard baseline, we expect that more sophisticated combination methods like grounding [25] will obtain larger gains in the future.

5 Discussion

5.1 Contribution of the Visual Content

Table 6 summarizes the contribution of the images on text representations in test partition. The contribution is consistent through all text-based representations. We measure the absolute difference (**Diff**) and the error reduction (**E.R**) of each textual representation with the multimodal counterpart. For the comparison we chose the best text model for each representation. As expected we obtain the largest improvement ($22-26\%$ E.R) when text-based unsupervised models are combined with image representations. Note that unsupervised models are not learning anything about the specific task, so the more information in the representation, the better. In the case of USE and VSE++ the improvement is significant but not as large as the purely unsupervised models. The best text-only representation is the one fine-tuned on a multimodal task, VSE++, which is noteworthy, as it is better than a textual representation fine-tuned in a text-only inference task like USE.

[6] VSE++ +RESNET and USE+RESNET models.

Model	Modality	train ρ	dev ρ	test ρ
GLOVE	text	0.819	0.744	0.702
BERT	text	0.888	0.775	0.781
GPT-2	text	0.265	0.285	0.246
USE	text	0.861	0.824	0.810
VSE++(TEXT)	text	0.883	**0.831**	**0.825**
RESNET	image	0.788	**0.721**	**0.706**
VSE++(IMG)	image	0.775	0.703	0.701
CONCAT: GLOVE+RESNET	mmodal	0.899	0.830	0.794
CONCAT: BERT+RESNET	mmodal	0.889	0.805	0.797
CONCAT: USE+RESNET	mmodal	0.892	0.859	0.841
CONCAT: VSE++ +RESNET	mmodal	0.915	**0.864**	**0.852**
PROJECT: GLOVE+RESNET	mmodal	0.997	0.821	0.826
PROJECT: BERT+RESNET	mmodal	0.996	0.825	0.827
PROJECT: USE+RESNET	mmodal	0.998	0.850	0.837
PROJECT: VSE++ +RESNET	mmodal	0.998	**0.853**	**0.847**

Table 5. Supervised scenario: Train, validation and test results of the unsupervised models

Improvement is consistent for the supervised models. Contrary to the unsupervised setting, these models are designed to learn about the task, so there is usually less room for the improvement. Still, GLOVE+RESNET shows an error reduction of 12.9 in the test set. Finally, USE and VSE++ show smaller improvements when we add visual information into the model.

Scenario	Repr	text	mmodal	Diff	E.R
Unsup	GLOVE	0.587	0.730	0.143	24.4
Unsup	BERT	0.612	0.745	0.133	21.7
Unsup	USE	0.720	0.787	0.067	9.3
Unsup	VSE++	0.803	0.826	0.023	2.9
Sup	GLOVE	0.702	0.793	0.091	12.9
Sup	BERT	0.781	0.827	0.046	5.8
Sup	USE	0.810	0.841	0.031	3.8
Sup	VSE++	0.825	0.852	0.027	3.3

Table 6. Contribution of images over text representations on test.

Figure 4 displays some examples where visual information positively contributes predicting accurately similarity values. Examples show the case where related descriptions are lexicalized in a different way so a text-only model (GLOVE) predicts low similarity between captions (top two examples). Instead, the multimodal representation GLOVE+RESNET does have access to the image and can predict more accurately the similarity value of the two captions. The examples in the bottom show the opposite case, where similar set of words are used to describe very different situations. The text based model overestimates the similarity of captions, while the multimodal model corrects the output by looking at the differences of the images.

On the contrary, Figure 5 shows that images can also be misleading, and that the task is not as trivial as combining global representations of the image. In this case, related but different captions are supported by very similar images, and as a consequence, the multimodal model overestimates their similarity, while the text-only model focuses on the most discriminating piece of information in the text.

Figure 4. Examples of the contribution of the visual information in the task. **gs** for gold standard similarity value, **text** and **mm** for text-only and multimodal models, respectively. On top examples where related descriptions are lexicalized differently and images help. On the bottom cases where similar words are used to describe different situations.

5.2 The effect of hyperparameters

Neural models are sensitive to hyperparameters, and we might think that results on the supervised scenario are due to hyperparameter optimization. Figure 6 displays the variability of ρ in development across all hyperparameters. Due to space constraints we show text-only and multimodal concatenated models. Models are ordered by mean performance. As we can see, combined models show better mean performance, and all models except Glove exhibit tight variability.

6 Conclusions and Future Work

The long term goal of our research is to devise multimodal representation techniques that improve current inference capabilities.

Figure 5. Example of misleading images. The high similarity of images makes the prediction of the multimodal model inaccurate, while the text only model focuses on the most discriminating piece of information. Note that **gs** refers to the gold standard similarity value, and **text** and **mm** refer to text-only and multimodal models, respectively.

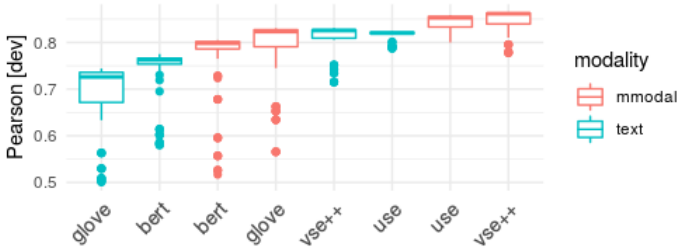

Figure 6. Variability of the supervised models regarding hyperparameter selection on development. The multimodal models use concatenation. Best viewed in colour.

We have presented a novel task, Visual Semantic Textual Similarity (vSTS), where the inference capabilities of visual, textual, and multimodal representations can be tested directly. The dataset has been manually annotated by crowdsourcers with high inter-annotator correlation ($\rho = 0.89$). We tested several well-known textual and visual representations, which we combined using concatenation and projection. Our results show, for the first time, that the addition of image representations allows better inference. The best text-only representation is the one fine-tuned on a multimodal task, VSE++, which is noteworthy, as it is better than a textual representation fine- tuned in a text-only inference task like USE. The improvement when using image representations is observed both when computing the similarity directly from multimodal representations, and also when training siamese networks.

In the future, we would like to ground the text representations to image regions [25], which could avoid misleading predictions due to the global representation of the image. Finally, we would like to extend the dataset with more examples, as we acknowledge that training set is limited to train larger models.

ACKNOWLEDGEMENTS

This research was partially funded by the Basque Government excellence research group (IT1343-19), the NVIDIA GPU grant program, the Spanish MINECO (DeepReading RTI2018-096846-B-C21 (MCIU/AEI/FEDER, UE)) and project BigKnowledge (Ayudas Fundación BBVA a equipos de investigación científica 2018). Ander enjoys a PhD grant from the Basque Government.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, 'VQA: Visual question answering', in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, (2015).

[2] Kobus Barnard and David Forsyth, 'Learning the semantics of words and pictures', in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 408–415. IEEE, (2001).

[3] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank, 'Automatic description generation from images: A survey of models, datasets, and evaluation measures', *Journal of Artificial Intelligence Research*, **55**, 409–442, (2016).

[4] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning, 'A large annotated corpus for learning natural language inference', *arXiv preprint arXiv:1508.05326*, (2015).

[5] Elia Bruni, Nam Khanh Tran, and Marco Baroni, 'Multimodal distributional semantics', *J. Artif. Int. Res.*, **49**(1), 1–47, (January 2014).

[6] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia, 'SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation', in *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 1–14, (2017).

[7] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al., 'Universal sentence encoder for english', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169–174, (2018).

[8] Ido Dagan, Oren Glickman, and Bernardo Magnini, 'The pascal recognising textual entailment challenge', in *Machine Learning Challenges Workshop*, pp. 177–190. Springer, (2005).

[9] Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre, 'Evaluating Multimodal Representations on Sentence Similarity: vSTS, Visual Semantic Textual Similarity Dataset', *Workshop on Closing the Loop Between Vision and Language at ICCV*, (2017).

[10] Wim De Mulder, Steven Bethard, and Marie-Francine Moens, 'A survey on the application of recurrent neural networks to statistical language modeling', *Computer Speech & Language*, **30**(1), 61–98, (2015).

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, 'ImageNet: A Large-Scale Hierarchical Image Database', in *CVPR09*, (2009).

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT - Pre-training of Deep Bidirectional Transformers for Language Understanding.', *CoRR*, **1810**, arXiv:1810.04805, (2018).

[13] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia, 'Multi30k: Multilingual english-german image descriptions', in *5th Workshop on Vision and Language*, (2016).

[14] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler, 'Vse++: Improving visual-semantic embeddings with hard negatives', (2018).

[15] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, 'Devise: A deep visual-semantic embedding model', in *Neural Information Processing Systems (NIPS)*, (2013).

[16] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu, 'Mscap: Multi-style image captioning with unpaired stylized text', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4204–4213, (2019).

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, (2015).

[18] Sepp Hochreiter and Jürgen Schmidhuber, 'Long short-term memory', *Neural Comput.*, **9**(8), 1735–1780, (November 1997).

[19] Micah Hodosh, Peter Young, and Julia Hockenmaier, 'Framing image description as a ranking task: Data, models and evaluation metrics', *J. Artif. Int. Res.*, **47**(1), 853–899, (May 2013).

[20] Andrej Karpathy and Fei-Fei Li, 'Deep visual-semantic alignments for generating image descriptions', in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3128–3137, (2015).

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei, 'Visual genome: Connecting language and vision using crowdsourced dense image annotations', *Int. J. Comput. Vision*, **123**(1), 32–73, (May 2017).

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, 'Imagenet classification with deep convolutional neural networks', in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, USA, (2012).

[23] Yann LeCun, Yoshua Bengio, et al., 'Convolutional networks for images, speech, and time series', *The handbook of brain theory and neural networks*, **3361**(10), 1995, (1995).

[24] 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', *CoRR*, **abs/1405.0312**, (2014).

[25] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, 'Generation and comprehension of unambiguous object descriptions', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, (June 2016).

[26] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow, 'Trends in integration of vision and language research: A survey of tasks, datasets, and methods', *ArXiv*, **abs/1907.09358**, (2019).

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning, 'GloVe: Global vectors for word representation', in *Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. Stanford University, Palo Alto, United States, (January 2014).

[28] 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*, pp. 2641–2649, (2015).

[29] 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', *Int. J. Comput. Vision*, **123**(1), 74–93, (May 2017).

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, 'Language models are unsupervised multitask learners', (2019).

[31] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier, 'Collecting image annotations using Amazon's mechanical turk', in *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 139–147, (2010).

[32] Ehud Reiter and Robert Dale, *Building natural language generation systems*, Cambridge university press, 2000.

[33] Karen Simonyan and Andrew Zisserman, 'Very deep convolutional networks for large-scale image recognition', in *International Conference on Learning Representations*, (2015).

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and ZB Wojna, 'Rethinking the inception architecture for computer vision', (06 2016).

[35] Kai Sheng Tai, Richard Socher, and Christopher D. Manning, 'Improved semantic representations from tree-structured long short-term memory networks', in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 1556–1566, (2015).

[36] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen, 'Measuring the diversity of automatic image descriptions', in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1730–1741, Santa Fe, New Mexico, USA, (August 2018). Association for Computational Linguistics.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in Neural Information Processing Systems 30*, eds., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5998–6008, (2017).

[38] Elena Voita, Rico Sennrich, and I. M. Titov, 'The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives', *ArXiv*, **abs/1909.01380**, (2019).

[39] Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjan, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt, 'Grounded textual entailment', in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2354–2368, Santa Fe, New Mexico, USA, (August 2018).

[40] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav, 'Visual entailment: A novel task for fine-grained image understanding', *arXiv preprint arXiv:1901.06706*, (2019).

[41] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski, 'Graph convolutional networks: a comprehensive review', *Computational Social Networks*, **6**(1), 11, (2019).