LEARNING GROUNDED SENTENCE REPRESENTATIONS BY JOINTLY USING VIDEO AND TEXT INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual grounding of language is an active research field aiming at enriching text-based representations with visual information. In this paper, we propose a new way to leverage visual knowledge for sentence representations. Our approach transfers the structure of a visual representation space to the textual space by using two complementary sources of information: (1) the *cluster information*: the implicit knowledge that two sentences associated with the same visual content describe the same underlying reality and (2) the *perceptual information* contained within the structure of the visual space. We use a joint approach to encourage beneficial interactions during training between textual, perceptual, and cluster information. We demonstrate the quality of the learned representations on semantic relatedness, classification, and cross-modal retrieval tasks.

1 INTRODUCTION

Building linguistic vectors that represent semantics is a long-standing issue in Artificial Intelligence. Distributional Semantic Models (Mikolov et al., 2013; Peters et al., 2018) are well-known recent efforts in this direction, making use of the *distributional hypothesis* (Harris, 1954) on text corpora to learn word embeddings. At another granularity level, having high-quality general-purpose sentence representations is crucial for all models that encode sentences into semantic vectors, such as the ones used in machine translation (Bahdanau et al., 2014) or question answering (Sagara & Hagiwara, 2014). Moreover, encoding semantics of sentences is paramount because sentences describe relationships between objects and thus convey complex and high-level knowledge better than individual words, which mostly refer to a single concept (Norman, 1972).

Relying only on text can lead to biased representations and unrealistic predictions (e.g., text-based models could predict that "the sky is green" (Baroni, 2016)). Besides, it has been shown that human understanding of language is *grounded* in physical reality and perceptual experience (Fincher-Kiefer, 2001). To overcome this limitation, one emerging approach is the *visual grounding of language*, which consists of leveraging visual information, usually from images, to enhance word representations. Two methods showing substantial improvements have emerged: (1) the *sequential* technique combines textual and visual representations that were separately learned (Bruni et al., 2014; Silberer & Lapata, 2014), and (2) the *joint* method learns a common multimodal representation from multiple sources simultaneously (Lazaridou et al., 2015). In the case of words, the latter has proven to produce representations that perform better on intrinsic and downstream tasks.

While there exist numerous approaches to learning sentence representations from text corpora only, and to learning multimodal word embeddings, the problem of the *visual grounding of sentences* is quite new to the research community. To the best of our knowledge, the only work in the field is Kiela et al. (2018). The authors propose a sequential model: linguistic vectors, learned from a purely textual corpus, are concatenated with grounded vectors, which were independently learned from a captioning dataset. However, the two sources are considered separately, which might prevent beneficial interactions between textual and visual modalities during training.

We propose a joint model to learn multimodal sentence representations, based on the assumption that the meaning of a sentence is simultaneously grounded in its textual and visual contexts. In our case, the textual context of a sentence consists of adjacent sentences in a text corpus. Within a distinct dataset, the visual context is learned from a paired video and its associated captions. Indeed, we propose to use videos instead of images because of their temporal aspect, since sentences often describe actions grounded in time. The key challenge is to capture visual information. Usually, to transfer information from the visual space to the textual one, one space is *projected* onto the other (Kiela et al., 2018; Lazaridou et al., 2015). However, as pointed out by Collell & Moens (2018), projections are not sufficient to transfer neighborhood structure between modalities. In our work, we rather propose to exploit the visual space by *preserving* the overall structure, i.e. conserving the similarities between related elements across spaces. More precisely, we take visual context into account by distinguishing two types of complementary information sources. First, the *cluster information*, which consists in the implicit knowledge that sentences associated with the same video refer to the same underlying reality. Second, the *perceptual information*, which is the high-level information extracted from a video using a pre-trained CNN.

Regarding these considerations, we formulate three Research Questions (RQ):

• **RQ1**: Is perceptual information useful to improve sentence representations?

• **RQ2**: Are cluster and perceptual information complementary, and does their combination compete with previous models based on projections between visual and textual spaces?

• **RQ3**: Is a joint approach better suited than a sequential one regarding the multimodal acquisition of textual and visual knowledge?

Our contribution is threefold: (1) We propose a joint multimodal framework for learning grounded sentence representations; (2) We show that cluster and perceptual information are complementary sources of information; (3) To the best of our knowledge, obtained results achieve state-of-the-art performances on multimodal sentence representations.

2 Joint Multimodal framework for sentence representation

2.1 Model overview

Our framework learns multimodal representations for sentences by jointly leveraging the textual and visual contexts of a sentence. The textual resource is a large text corpus C_T of ordered sentences. The visual resource is a distinct video corpus C_V, whose videos are associated with one or more descriptive captions.

A sentence S is represented by $s = F_\theta(S)$ and its corresponding video V_S by $v_s = G_{\theta'}(V_S)$, where F (resp. G) is a sentence (resp. video) encoder parameterized by θ (resp. θ'). We propose to use a *joint* approach where the sentence encoder F_θ is learned by jointly optimizing a textual objective $\mathcal{L}_\mathcal{T}(\theta)$ on C_T and a visual objective $\mathcal{L}_\mathcal{V}(\theta, \theta')$ on C_V. So far, this method has only been applied to words, with good results (Lazaridou et al., 2015; Zablocki et al., 2018). Note that C_T and C_V are not parallel corpora but that θ is shared between both objectives; in other terms, sentence representations are influenced by their distinct textual and visual contexts. Any sentence encoder F_θ and textual objective $\mathcal{L}_\mathcal{T}$ can be used such as SkipTought (Kiros et al., 2015), FastSent (Hill et al., 2016) or QuickThought (Logeswaran & Lee, 2018). In this paper, we focus on SkipThought, and present evidences that our approach also improves over FastSent (section 4.3). In the following, we introduce hypotheses and their derived objectives to tackle the modeling of $\mathcal{L}_\mathcal{V}$.

2.2 Leveraging the visual context

Most visual grounding works use projections between the textual space and the visual space (Kiela et al., 2018; Lazaridou et al., 2015) to integrate visual information. However, when a cross-modal mapping is learned, the projection of the source modality does not resemble the target modality, in the sense of neighborhood topology (Collell & Moens, 2018). This suggests that projections between spaces is not an appropriate approach to incorporate visual semantics. Instead, we propose a new way to structure the textual space with the help of the visual modality.

Without even considering the content of videos, the fact that sentences describe or not a same underlying reality is an implicit source of information that we name the *cluster information*. For convenience, two sentences are said to be *visually equivalent* (resp. *visually different*) if they are associated with the same video (resp. different videos), i.e. if they describe the same (resp. different) underlying reality. We call *cluster* a set of visually equivalent sentences. Leveraging the cluster information may be useful to improve the structure of the textual space: intuitively, representations of visually equivalent sentences should be close, and representations of visually different sentences should be separated. We thus formulate the following hypothesis (see red elements in Figure 1):

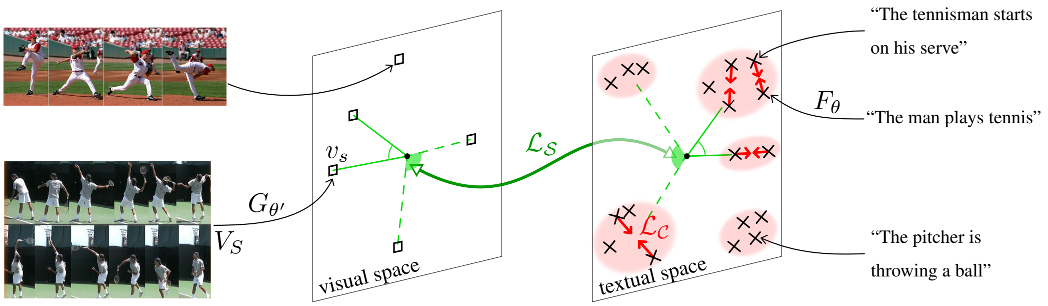
"The tennisman starts on his serve"
F_θ
"The man plays tennis"
"The pitcher is throwing a ball"
$\mathcal{L}_\mathcal{S}$
$G_{\theta'}$
V_S
v_s
visual space
textual space
$\mathcal{L}_\mathcal{C}$

Figure 1: Illustration of the cluster and perceptual hypotheses. Red circles indicate visual clusters. Red arrows represent the gradient of the loss derived from the cluster hypothesis (**C**), which gathers visually equivalent sentences. For clarity's sake, the term in equation 1 that separates negative pairs is not represented. The green arrow and angles illustrate the loss derived from the perceptual hypothesis (**S**), which requires cosine similarities to correlate across the two spaces. The point at the center of each space is the origin.

Cluster Hypothesis (**C**): A sentence should be closer to a visually equivalent sentence than to a visually different sentence.

We translate this hypothesis into the constraint $\cos(s, s^+) \leq \cos(s, s^-)$, where s^+ (resp. s^-) is a visually equivalent (resp. different) sentence to s. Following Karpathy & Li (2015); Carvalho et al. (2018), we use a max-margin ranking loss to ensure the gap between both terms is higher than a fixed margin m:

$$\mathcal{L}_\mathcal{C} = \sum_{(s,s^+)} \sum_{s^-} \max(0, m - \cos(s, s^+) + \cos(s, s^-)) \tag{1}$$

where (s, s^+) cover visually equivalent pairs; visually different sentences s^- are randomly sampled.

The cluster hypothesis ignores the structure of the visual space and only uses the visual modality as a proxy to assess if two sentences are visually equivalent or different. Moreover, a ranking loss simply drives visually different sentences apart in the representation space, even if their corresponding videos are closely related.

To cope with this limitation, we suggest to take into account the structure of the visual space and use the content of videos, and then propose a novel approach which does not require cross-modal projections. The intuition is that the structure of the textual space should be modeled on the structure of the visual one to extract visual semantics. We choose to preserve *similarities* between related elements across spaces. We thus formulate the following hypothesis, illustrated with green elements in Figure 1:

Perceptual hypothesis (**P**): The similarity between two sentences in the textual space should be correlated with the similarity between their corresponding videos in the visual space.

We translate this hypothesis into the loss $\mathcal{L}_\mathcal{P} = -\rho_{vis}$, where $\rho_{vis} = \rho(\cos(s, s'), \cos(v_s, v_{s'}))$ and ρ is the Pearson correlation.

The final multimodal loss is a linear combination of the aforementioned objectives, weighted by hyperparameters α_T, α_P and α_C:

$$\mathcal{L}(\theta, \theta') = \underbrace{\alpha_T.\mathcal{L}_\mathcal{T}(\theta)}_{\text{textual context}} + \underbrace{\alpha_P.\mathcal{L}_\mathcal{P}(\theta, \theta') + \alpha_C.\mathcal{L}_\mathcal{C}(\theta)}_{\text{visual context } \mathcal{L}_\mathcal{V}} \tag{2}$$

2.3 Video modeling

To evaluate the impact of visual semantics on sentence grounding, we examine several types of visual context. As done in Yao et al. (2016); Guo et al. (2016), visual features are extracted using the penultimate layer of a pretrained CNN. A video is represented as a set of n images $(I_k)_{k\in[1,n]}$. Let $(i_k)_{k\in[1,n]}$ be the representations of these images obtained with the pre-trained CNN. We present below three simple ways to represent a video V. Note that our model can be generalized to more complex video representations (Ji et al., 2010; Simonyan & Zisserman, 2014b).

One Frame (F): this simple setting amounts at keeping the first frame and ignoring the rest of the sequence (any other frame might be used). The visual context vector is $v = i_1$.

Average (A): the temporal aspect is ignored, and the scene is represented by the average of the individual frame features: $v = \frac{1}{n}\sum_{k=1}^{n} i_k$ (Zha et al., 2015).

Temporal Grounding (T): the intuition is that, in a video, not all frames are relevant to sentence understanding. An attention mechanism allows us to focus on important frames. We set: $v = \sum_{k=1}^{n} \beta_k i_k$, where $\beta_k = \text{softmax}(<\sum_w u_w, N.i_k>)$. The sum ranges over the words w of the sentence s, u_w is the fixed pretrained word embedding of w, and N is a learned projection.

3 Evaluation protocol

3.1 Datasets

Textual dataset. Following Kiros et al. (2015); Hill et al. (2016), we use the Toronto BookCorpus dataset as the textual corpus C_T. This corpus consists of 11K books: this makes a total of 74M ordered sentences, with an average of 13 words per sentence.

Visual dataset. We use the MSVD dataset (Chen & Dolan, 2011) as the visual corpus C_V. This video captioning dataset consists of 1970 videos and 80K English descriptions. On average, a video lasts 10 seconds and has about 41 associated sentences.

3.2 Baselines and Scenarios

Model Scenarios. We test different variants of our multimodal model presented in section 2. We note these variants $\mathbf{M}_V^I(\alpha_T, \alpha_P, \alpha_C)$, which depend on:
• the *initialization* $I \in \{p, \varnothing\}$: the sentence encoder F_θ is either pretrained using the textual objective $\mathcal{L}_\mathcal{T}$ ($I = p$), or initialized randomly ($I = \varnothing$).
• the *visual representation* $V \in \{F, A, T, R\}$: where F, A or T are the video modelings described in Section 2.3. We introduce a baseline R, where visual vectors are randomly sampled from a normal distribution to measure the information brought by the video content.

Baselines. We propose two extensions of multimodal word embedding models to sentences:
• *Projection* (**P**): Inspired by Lazaridou et al. (2015), this baseline is projecting videos in the textual space, while our model keeps both spaces separated. The visual loss is a ranking objective:

$$\mathcal{L}_\mathcal{V} = \sum_s \sum_{v_-} \max(0, m' - \cos(s, W.v_-) + \cos(s, W.v_s)) \tag{3}$$

where W is a trainable projection matrix and m' a fixed margin. We note $\mathbf{P}_V^I(\alpha_T)$ the variants of this baseline using the global loss $\mathcal{L} = \alpha_T.\mathcal{L}_\mathcal{T} + \mathcal{L}_\mathcal{V}$.

• *Sequential* (**SEQ**): Inspired by Collell Talleda et al. (2017), we learn a linear regression model (W, b) to predict the visual representation from the SkipThought representations. The multimodal sentence embedding is the concatenation of the original SkipThought vector and its predicted representation: $\mathbf{ST} \oplus W\mathbf{ST} + b$, projected into a lower-dimensional space using PCA. This baseline can also be seen as a simpler variant of the model in Kiela et al. (2018).

3.3 Tasks and measures

In line with previous works on sentence embeddings (Kiros et al., 2015; Hill et al., 2016), we consider several benchmarks to evaluate the quality of our learned multimodal representations:

Semantic relatedness: We use two well-known semantic similarity benchmarks: STS (Cer et al., 2017) and SICK (Marelli et al., 2014), which consist of pairs of sentences that are associated with human-labeled similarity scores. STS is subdivided in three textual sources: *Captions* contains sentences with a strong visual content, describing everyday-life actions, whereas the others contain more abstract sentences: news headlines in *News* and posts from users forum in *Forum*. Correlations (Spearman/Pearson) are measured between the cosine similarity of our learned sentence embeddings and human-labeled scores. Hyperparameters are tuned on SICK/trial (results on SICK/train+test are reported in tables).

Classification benchmarks: We use six sentence classification benchmarks: paraphrase identification (MSRP) (Dolan et al., 2004), opinion polarity (MPQA) (Wiebe & Cardie, 2005), movie review sentiment (MR) (Pang & Lee, 2005), subjectivity/objectivity classification (SUBJ) (Scott et al., 2004), question-type classification (TREC) (Voorhees, 2001) and customer product reviews

Table 1: **RQ1**: Validation of the semantic hypothesis by comparing video modelings on semantic relatedness. We use model $\mathbf{M}^p_\bullet(0, 1, 0)$. Correlation results are given in the form $\rho_{\text{Spearman}}/\rho_{\text{Pearson}}$.

	Model	STS/All	STS/Cap.	STS/News	STS/For.	SICK
Text-only	**ST**	40/41	44/42	38/42	21/22	52/55
Visual info.	R	51/53	62/62	39/43	23/24	56/57
	F	57/59	75/75	**41/46**	24/26	**60**/61
	A	**58/60**	75/75	**41**/45	23/26	59/**63**
	T	57/**60**	**76/76**	**41/46**	**25/27**	**60/63**

(CR) (Hu & Liu, 2004). For each dataset, a logistic regression classifier is learned from the extracted sentence embeddings; we report the classification accuracy.

Cross-modal retrieval on COCO: We consider the image search/annotation tasks on the MS COCO dataset (Lin et al., 2014). A pairwise triplet-loss is optimized in order to bring corresponding sentences and images closer in a multimodal latent space. Evaluation is performed using Recall@K.

Structural measures: To analyze the quality of the textual space, we report some measures (computed in $\%$) defined on the MSVD test set:
• ρ_{vis} measures if the similarities between sentences correlate with the similarities between videos.
• $E_{intra} = \mathbb{E}_{v_s = v_{s'}}[cos(s, s')]$ measures the homogeneity of each cluster, by measuring the average similarity of sentences within a cluster.
• $E_{inter} = \mathbb{E}_{v_s \neq v_{s'}}[cos(s, s')]$ measures how well clusters are separated from each other (i.e. average similarity between sentences of two different clusters).

3.4 Implementation details

Videos are sampled at a 3 frames per second rate; afterwards, frames are processed using a pretrained VGG network (Simonyan & Zisserman, 2014a). The multimodal loss $\mathcal{L}$ is optimized with Adam optimizer (Kingma & Ba, 2014) and a learning rate $\lambda = 8.10^{-4}$. Hyperparameters are tuned using the Pearson correlation measure on SICK trial: $m = m' = 0.5$, $\mu = 2.5.10^{-4}$, and mini-batch size of 32 for $\mathcal{L}_\mathcal{V}$. We perform extensive experiments with $\mathcal{L}_\mathcal{T}$ based on the SkipThought model, using an embedding size of 2400 and the same network hyperparameters as in Kiros et al. (2015).

4 Experiments and Results

4.1 Validation of the perceptual hypothesis (RQ1)

The perceptual hypothesis holds that the information within videos is useful to ground sentence representations. In our model, this hypothesis translates into the perceptual loss $\mathcal{L} = \mathcal{L}_\mathcal{P}$ (i.e. model $\mathbf{M}^p_\cdot(0, 1, 0)$). Since the perceptual loss is the only component exploiting video content, we compare, in Table 1, the different video encoders on intrinsic evaluation benchmarks, namely semantic relatedness. The first observation is that our model **M** outperforms the purely textual baseline **ST** for all video encoders, which shows that perceptual information from videos is useful to improve representations. We also observe that using random visual anchors (R) improves over **ST**. This validates our cluster hypothesis, since grouping visually equivalent sentences improves representation – even when anchors bear no perceptual semantics. We further observe that $F, A, T > R$, which shows that the perceptual information from videos brings a more semantically meaningful structure to the representation space. Finally, regarding the different ways to encode a video, we observe that leveraging more than one frame can be slightly beneficial to learn grounded sentence representations, e.g. A obtains $+3.3\%$ average relative improvement over F on $\rho^{\text{SICK}}_{Pearson}$. Selecting relevant frames (T) in the video rather than considering all frames with equal importance (A) improves the quality of the embeddings.

It is worth noting that discrepancies between the modeling choices F, A, T are relatively low. This could be explained by the fact that videos from the MSVD dataset are short (10 seconds on average) and contain very few shot transitions. Thus, nearly all frames can provide a relevant visual context for associated sentences. We believe that higher differences would be exhibited for a dataset containing longer videos. In the remaining experiments, we therefore select A as the video model, since it offers a good balance between effectiveness (T) and efficiency (F).

Table 2: **RQ2**: Influence of visual hypotheses on the structure of the representation space.

		ρ_{vis}^{A}	E_{intra}	E_{inter}	STS/All	STS/Cap.	STS/News	STS/For.	SICK
	Model	Structural measures			Semantic relatedness				
Text-only	**ST**	16	43	25	40/41	44/42	38/42	21/22	52/55
Projection **P**	$\mathbf{P}_A^p(0)$	37	63	06	62/67	82/84	43/48	**29/31**	61/75
Cluster $\mathbf{M}_c$	$\mathbf{M}_A^p(0, \quad 0, 1)$	38	**66**	**01**	62/66	83/84	41/46	22/24	**62/76**
Perceptual $\mathbf{M}_p$	$\mathbf{M}_A^p(0, \quad 1, 0)$	**53**	44	18	58/60	75/75	41/45	23/26	59/63
Both combined $\mathbf{M}_b$	$\mathbf{M}_A^p(0, 0.1, 1)$	46	63	02	**64/68**	**84/85**	**44/49**	27/29	**62/76**

4.2 Complementarity of cluster and perceptual information (RQ2)

We study here the influence of perceptual and cluster information on the embedding space structure. To do so, we report, in Table 2, the structural measures on three versions of our model – $\mathbf{M}_c$ (cluster information), $\mathbf{M}_p$ (perceptual information) and $\mathbf{M}_b$ (combination of both), as well as on baselines **ST** and **P**. For **M** and **P**, we discard the textual loss to isolate the effect of the different hypotheses. As expected, solely using cluster information leads to the highest E_{intra} and lowest E_{inter}, which suggests that $\mathbf{M}_c$ is the most efficient model at separating visually different sentences. Using only perceptual information in $\mathbf{M}_p$ logically leads to highly correlated textual and visual spaces (highest ρ_{vis}^{A}), but the local neighborhood structure is not well preserved (lowest E_{intra} and highest E_{inter}). $\mathbf{M}_b$ and **P** are optimized for both forming well-separated clusters and capturing the perceptual information within the representation space. This translates into a high E_{intra} and low E_{inter}. However, the main difference lies in the fact that $\mathbf{M}_b$ is better at preserving the geometry of the visual space (higher ρ_{vis}^{A}). This difference results in better performances for our model $\mathbf{M}_b$ in terms of semantic relatedness compared to **P**. It reinforces our claim that both visual and perceptual information complement each other for sentence representation. Therefore, in the remaining experiments, we use the combined model $\mathbf{M}_A^{\bullet}(\alpha_T, .1, 1)$, that we note $\mathbf{M}_A^{\bullet}(\alpha_T)$ for clarity's sake.

4.3 Performance of scenarios and baselines (RQ2 and RQ3)

Table 4.3 reports the effectiveness of the sentence embeddings obtained from our scenarios and baselines on semantic relatedness and classification tasks. We first observe that multimodal models generally outperform the text-only baseline **ST** on both semantic relatedness and classification benchmarks. Interestingly, we notice that the STS/Captions benchmark gives the highest discrepancies compared to the text-only baseline, probably because these sentences have a highly visual content. Second, we notice that a high α_T leads to high classification scores, whereas a low α_T leads to high semantic relatedness scores. There is a trade-off between semantic relatedness and classification scores, that we can set properly by tuning α_T. Indeed, properly weighting the textual contribution in the global loss $\mathcal{L}$ is task-dependent, for every grounding model. This echoes the problem reported in Faruqui et al. (2016) in the context of word embeddings: there is no strong correlation between the semantic relatedness scores and extrinsic evaluation (e.g. classification) scores.

As a qualitative analysis, we illustrate in Table 3 that, due to our multimodal model, concrete knowledge acquired via visual grounding can be transferred to abstract sentences. To do so, we manually build abstract sentence queries using words with low concreteness (between 2.5 and 3.5) from the USF dataset (Nelson et al., 2004). Then, nearest neighbors are retrieved from all sentences of the MS COCO training set. We see that our multimodal model is more accurate than the purely textual model to capture visual meaning, even for sentences that are not inherently visual. For example, on the first line of Table 3, **ST**'s sentence contradicts the query by depicting the man as "smiling", whereas **M**'s sentence gives a concrete vision of horror: "grabs his head while screaming". The observation that perceptual information propagates from concrete sentences to abstract ones is analogous to findings made in previous research on word embeddings (Hill & Korhonen, 2014).

Table 3: Qualitative analysis: finding the nearest neighbor of a given query in the textual space.

Query	**ST**	$\mathbf{M}_A^p(0, 0.1, 1)$
A man is **horrified**	An older man in a suit is smiling	The man is holding his face and screaming
This is a **tragedy**	I think this is a huge food court	View from the survivor of a motorcycle accident
Two people are in **love**	Two people are out in the ocean kitesurfing	A couple of people that are next to each other

Table 4: **RQ2,3**: Semantic relatedness and classification performances. $\mathbf{M}(\alpha_T)$ stands for $\mathbf{M}(\alpha_T, 0.1, 1)$. Note that, in all models, sentence vectors have the same dimension (2400).

			STS/All	STS/Cap.	STS/News	STS/For.	SICK	MSRP	MPQA	MR	SUBJ	TREC	CR
		Model	Semantic relatedness					Classification					
Baselines	Text-only	**ST**	40/41	44/42	38/42	21/22	52/55	71.6	86.2	75.9	92.1	89.4	**82.5**
	Sequential	**SEQ**	47/44	70/59	37/44	29/24	58/69	70.0	86.1	75.8	92.2	89.2	81.9
	Projection	$\mathbf{P}^p_A(500)$	41/39	57/47	38/43	19/19	54/59	73.1	86.2	76.7	92.5	88.8	81.3
		$\mathbf{P}^p_A(100)$	46/46	64/60	38/43	18/18	54/61	72.9	86.4	77.0	92.6	89.8	81.0
		$\mathbf{P}^p_A(\ 10\ )$	54/57	76/76	38/44	20/20	59/69	72.2	86.3	77.0	92.9	88.8	81.4
		$\mathbf{P}^p_A(\ 1\ )$	58/62	81/82	40/46	22/22	62/74	71.3	86.3	76.0	92.2	89.6	81.2
		$\mathbf{P}^p_A(\ 0\ )$	62/67	82/84	43/48	29/31	61/75	70.8	86.0	71.6	88.2	87.4	77.8
Our models	From scratch	$\mathbf{M}^{\varnothing}_A(500)$	50/51	71/69	37/43	20/19	60/70	72.2	86.3	77.2	**93.0**	**90.4**	81.0
		$\mathbf{M}^{\varnothing}_A(100)$	54/56	78/76	38/44	23/22	**64**/75	72.9	86.6	**77.3**	92.8	89.0	81.7
		$\mathbf{M}^{\varnothing}_A(\ 10\ )$	54/56	77/76	39/44	24/23	**64**/73	71.2	86.4	76.0	92.6	87.2	81.3
		$\mathbf{M}^{\varnothing}_A(\ 1\ )$	55/59	77/79	40/46	26/25	**64**/71	71.1	86.3	76.0	92.2	89.6	81.1
		$\mathbf{M}^{\varnothing}_A(\ 0\ )$	58/60	83/83	**45/50**	**37/34**	60/72	65.3	76.2	60.4	71.2	69.0	65.9
	Pretrained	$\mathbf{M}^p_A(500)$	44/43	61/54	39/44	19/19	54/60	**74.4**	86.1	76.3	92.6	88.8	81.3
		$\mathbf{M}^p_A(100)$	49/50	70/66	38/43	17/17	56/65	73.2	86.2	76.8	92.5	89.2	81.5
		$\mathbf{M}^p_A(\ 10\ )$	56/59	79/78	38/44	19/19	61/73	72.6	**86.7**	76.2	92.4	88.6	81.5
		$\mathbf{M}^p_A(\ 1\ )$	60/64	82/83	40/46	22/23	63/**76**	71.6	86.2	76.0	92.0	88.8	81.1
		$\mathbf{M}^p_A(\ 0\ )$	**64/68**	**84/85**	44/49	27/29	62/**76**	70.1	85.9	72.7	89.6	86.8	78.0

To further answer **RQ2**, we compare our model **M** with the projection baseline **P**. Our model obtains higher results than **P** on semantic relatedness tasks and comparable ones on classification tasks. For example, $\mathbf{M}^p$ has $5\%/3\%$ average relative improvement over **P** on semantic relatedness tasks. This suggests that preserving the structure of the visual space is more effective than learning cross-modal projections, as outlined in section 4.2. Indeed, this statement is strengthened by the fact that our model also improves over a sequential state-of-the-art model (Kiela et al., 2018). Since their textual baseline is weaker than ours (due to differences in the encoder and the dimensionality), we do not report their results in Table 4.3. However, we compare, between both approaches, the discrepancy Δ between the best multimodal model and the respective text-only baseline, while keeping dimensionality constant. On the benchmarks MPQA, MR, SUBJ and MSRP, our Δ is higher than theirs. For example, $\Delta^{\text{Kiela et al.}}_{\text{MSRP}} = 0.7$ and $\Delta^{\mathbf{M}^p_A(500)}_{\text{MSRP}} = 74.4 - 71.6 = 2.8$.

To answer **RQ3**, we compare joint and sequential approaches. We notice that joint models **M** and **P** globally perform better than the sequential baseline **SEQ** on classification and semantic relatedness tasks. For instance, $\mathbf{M}^{\varnothing}_A(500)$ has $5\%/9\%$ average relative improvement (resp. 1%) over **SEQ** on semantic relatedness (resp. classification benchmarks). Therefore, the joint approach shows superior performances to the sequential one, confirming results reported for grounded word embeddings (Zablocki et al., 2018). Finally, our models trained from scratch perform slightly better than pretrained ones. This might be due to the fact that visual and textual information are integrated in a joint manner from the beginning of training, which leads to better interactions between visual and textual modalities.

To further evaluate the quality of the embeddings, we perform cross-modal retrieval experiments on the COCO dataset (Lin et al., 2014). In Table 5, we report the results of our best performing models, which corroborates our previous statements on semantic relatedness and classification.

Finally, we probe that our model is independent from the choice of the textual encoder and objective $\mathcal{L}_T$, we use the FastSent model (Hill et al., 2016) instead of the SkipThought model. We observe similar improvements in performances (e.g. $\Delta_{\text{STS}} = 4/4$ and $\Delta_{\text{SICK}} = 7/7$ for the best performing model $\mathbf{M}^p_{\mathbf{A}}(0)$), confirming that our visual grounding strategy applies to any textual model.

Table 5: Model performances on COCO cross-modal retrieval task. R@k is the recall at k metric.

	R@1	R@5	R@10	R@1	R@5	R@10
Model	Image Search			Image Annotation		
ST	31.8	66.5	79.9	**25.0**	58.0	73.4
$\mathbf{P}_A(0)$	30.9	65.4	79.0	24.5	57.3	73.3
$\mathbf{M}^p_A(0)$	32.1	**67.0**	**80.1**	24.5	58.1	**74.3**
$\mathbf{M}^p_A(1)$	**33.2**	**67.0**	79.9	**25.0**	**58.7**	73.9

5 Related work

Sentence representations: Several approaches have been proposed over the last years to build semantic representations for sentences. On the one hand, supervised techniques produce task-specific sentence embeddings. For example, in a classification context, they are built using recurrent networks with LSTM (Hochreiter & Schmidhuber, 1997), recursive networks (Socher et al., 2013), convolutional networks (Kalchbrenner et al., 2014), or self-attentive networks (Lin et al., 2017). On the other hand, unsupervised methods aim at producing more general and task-independent sentence representations. Closer to our contribution, SkipThought (Kiros et al., 2015) and FastSent (Hill et al., 2016) are based on the distributional hypothesis (Harris, 1954) applied to sentences, i.e. *sentences that appear in similar contexts should have similar meanings*. In the SkipThought model, a sentence is encoded with a GRU network, and two GRU decoders are trained to reconstruct the adjacent sentences. In FastSent, the embedding of a sentence is the sum of its word embeddings; the learning objective is to predict all words in the adjacent sentences using a negative sampling loss. The present paper extends these works by integrating visual information.

Language grounding: To understand the way language conveys meaning, the traditional approach consists of considering language as a purely symbolic system based on words and syntactic rules (Chomsky, 1980; Burgess & Lund, 1997). However, Fincher-Kiefer (2001); W. Barsalou (1999) insist on the intuition that language has to be grounded in the real world and perceptual experience. The importance of real-world grounding is stressed in Gordon & Van Durme (2013), where an important bias is reported: the frequency at which objects, relations, or events occur in natural language are significantly different from their real-world frequency. Thus, leveraging visual resources, in addition to textual resources, is a promising way to acquire common-sense knowledge (Lin & Parikh, 2015; Yatskar et al., 2016) and cope with the bias between text and reality.

Following this intuition, Multimodal Distributional Semantic Models have been developed to cope with the lack of perceptual grounding in Distributional Semantic Models (Mikolov et al., 2013; Pennington et al., 2014). Two lines of work can be distinguished. First, the *sequential* approach separately builds textual and visual representations and combines them, via concatenation (Kiela & Bottou, 2014; Collell Talleda et al., 2017), linear weighted combination (Bruni et al., 2011), and Canonical Correlation Analysis (Loeub & Reichart, 2016). Second, the *joint* approach is intuitively closer to the way humans learn language semantics by hearing words and sentences in perceptual contexts. The advantage is that the visual information of *concrete words* is transferred to more abstract words that do not necessarily have associated visual data (Hill et al., 2014). Closer to our contribution, Lazaridou et al. (2015) presents the Multimodal Skip-Gram model, where the Word2vec objective (Mikolov et al., 2013) is optimized jointly with a max-margin ranking objective aiming at bringing concrete word vectors closer to their corresponding visual features. Similarly, Zablocki et al. (2018) show that not only the visual appearance of objects is important to word understanding, but also their context in the image, i.e. surroundings and neighboring objects. However, these models learn word representations while our model is intended to learn sentence representations.

Very recently, Kiela et al. (2018) have set ground for multimodal sentence representations. The authors propose a sequential method: language-only representations obtained from a text corpus (Toronto BookCorpus) are concatenated to grounded sentence vectors obtained from a caption dataset (MS COCO). A LSTM sentence encoder is trained to predict the representation of the corresponding image using a ranking loss and/or to predict other captions depicting the same image. Our work is different in several ways from theirs: we use a *joint* approach instead of a sequential one, and we distinguish and exploit cluster and perceptual information; moreover, we use videos instead of sentences and our framework is applicable to any textual sentence representation model.

6 Conclusion

In this paper, we proposed a joint multimodal model to learn sentence representations and our learned grounded sentence embeddings show state-of-the-art performances. Besides, our main findings are the following: (1) Both perceptual and cluster information are useful to learn sentence representations, in a complementary way. (2) Preserving the structure of the visual space, by modeling textual similarities on visual ones, outperforms a strategy based on projecting one space into the other. (3) A *joint* approach is more appropriate than a *sequential* method to learn multimodal representation for sentences. As future work, we would investigate the contribution of the temporal knowledge contained in videos for sentence grounding.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

Marco Baroni. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13, 2016.

Elia Bruni, Giang Binh Tran, and Marco Baroni. Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pp. 22–32, 2011. ISBN 978-1-937284-16-9.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, January 2014. ISSN 1076-9757.

Curt Burgess and Kevin Lund. Modelling parsing constraints with high-dimensional context space. 12, 03 1997.

Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pp. 35–44, 2018.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055, 2017.

David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 190–200, 2011. ISBN 978-1-932432-87-9.

Noam Chomsky. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15, 1980.

Guillem Collell and Marie-Francine Moens. Do neural network cross-modal mappings really bridge modalities? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 462–468, 2018.

Guillem Collell Talleda, Teddy Zhang, and Marie-Francine Moens. Imagined visual representations as multimodal embeddings. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. AAAI, 2017.

Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*, 2004.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, RepEval@ACL 2016, Berlin, Germany, August 2016*, pp. 30–35, 2016.

Rebecca Fincher-Kiefer. Perceptual components of situation models. *Memory & Cognition*, 29(2): 336–343, Mar 2001. ISSN 1532-5946. doi: 10.3758/BF03194928.

Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pp. 25–30, 2013. ISBN 978-1-4503-2411-3.

Zhao Guo, Lianli Gao, Jingkuan Song, Xing Xu, Jie Shao, and Heng Tao Shen. Attention-based lstm with semantic consistency for videos captioning. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pp. 357–361, 2016. ISBN 978-1-4503-3603-1.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Felix Hill and Anna Korhonen. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 255–265, 2014.

Felix Hill, Roi Reichart, and Anna Korhonen. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296, 2014.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 1367–1377, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 168–177, 2004. ISBN 1-58113-888-1.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 495–502, 2010. ISBN 978-1-60558-907-7.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 655–665, 2014.

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.

Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 36–45, 2014.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 408–418, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3294–3302, 2015.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pp. 153–163, 2015.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pp. 740–755, 2014.

Xiao Lin and Devi Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 2984–2993, 2015.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

Hagar Loeub and Roi Reichart. Effective combination of language and vision through model composition and the R-CCA method. *CoRR*, abs/1609.08810, 2016.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

Marco Marelli, Luisa Bentivogli, Marco G Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *SemEval@COLING*, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pp. 3111–3119, 2013.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407, 2004.

Donald A Norman. Memory, knowledge, and the answering of questions. 1972.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pp. 115–124, 2005.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–1543, 2014.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 2227–2237, 2018.

Tsukasa Sagara and Masafumi Hagiwara. Natural language neural network and its application to question-answering system. *Neurocomputing*, 142:201–208, 2014.

Donia Scott, Walter Daelemans, and Marilyn A. Walker (eds.). *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, 2004. ACL.

Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 721–732, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014a.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pp. 568–576, 2014b.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1631–1642, 2013.

Ellen M. Voorhees. Overview of the trec 2001 question answering track. In *In Proceedings of the Tenth Text REtrieval Conference (TREC*, pp. 42–51, 2001.

Lawrence W. Barsalou. Perceptual symbol systems. 22:577–609; discussion 610, 09 1999.

Janyce Wiebe and Claire Cardie. Annotating expressions of opinions and emotions in language. language resources and evaluation. In *Language Resources and Evaluation (formerly Computers and the Humanities*, pp. 2005, 2005.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. *Describing videos by exploiting temporal structure*, volume 11-18-December-2015, pp. 4507–4515. Institute of Electrical and Electronics Engineers Inc., United States, 2 2016. doi: 10.1109/ICCV.2015.512.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 193–198, 2016.

Éloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. Learning multi-modal word representation grounded in visual context. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

Shengxin Zha, Florian Luisier, Walter Andrews, Nitish Srivastava, and Ruslan Salakhutdinov. Exploiting image-trained CNN architectures for unconstrained video classification. In *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pp. 60.1–60.13, 2015.