Finding beans in burgers:
Deep semantic-visual embedding with localization

Martin Engilberge[1,2], Louis Chevallier[2], Patrick Pérez[2], Matthieu Cord[1]
[1]Sorbonne université, Paris, France [2]Technicolor, Cesson Sévigné, France
{martin.engilberge, matthieu.cord}@lip6.fr
patrick.perez@valeo.com louis.chevallier@technicolor.com

Abstract

Several works have proposed to learn a two-path neural network that maps images and texts, respectively, to a same shared Euclidean space where geometry captures useful semantic relationships. Such a multi-modal embedding can be trained and used for various tasks, notably image captioning. In the present work, we introduce a new architecture of this type, with a visual path that leverages recent space-aware pooling mechanisms. Combined with a textual path which is jointly trained from scratch, our semantic-visual embedding offers a versatile model. Once trained under the supervision of captioned images, it yields new state-of-the-art performance on cross-modal retrieval. It also allows the localization of new concepts from the embedding space into any input image, delivering state-of-the-art result on the visual grounding of phrases.

1. Introduction

Text and image understanding is progressing fast thanks to the ability of artificial neural nets to learn, with or without supervision, powerful distributed representations of input data. At runtime, such nets *embed* data into high-dimensional feature spaces where semantic relationships are geometrically captured and can be exploited to accomplish various tasks. Off-the-shelf already trained nets are now routinely used to extract versatile deep features from images which can be used for recognition or editing tasks, or to turn words and sentences into vectorial representations that can be mathematically analysed and manipulated.

Recent works have demonstrated how such deep representations of images and texts can be jointly leveraged to build *visual-semantic embeddings* [11, 17, 20, 33]. The ability to map natural images and texts in a shared representation space where geometry (distances and directions) might be interpreted is a powerful unifying paradigm. Not only does it permit to revisit visual recognition and captioning tasks, but it also opens up new usages, such as cross-modal content search or generation.

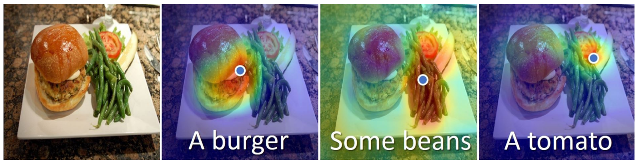

Figure 1. **Concept localization with proposed semantic-visual embedding**. Not only does our deep embedding allows cross-modal retrieval with state-of-the-art performance, but it can also associate to an image, *e.g.*, the hamburger plate on the left, a localization heatmap for any text query, as shown with overlays for three text examples. The circled blue dot indicates the highest peak in the heatmap.

One popular approach to semantic-visual joint embedding is to connect two mono-modal paths with one or multiple fully connected layers [20, 17, 39, 10, 2]: A visual path based on a pre-trained convolutional neural network (CNN) and a text path based on a pre-trained recurrent neural network (RNN) operating on a given word embedding. Using aligned text-image data, such as images with multiple captions from MS-COCO dataset [26], final mapping layers can be trained, along with the optional fine-tuning of the two branches. Building on this line of research, we investigate new pooling mechanisms in the visual path. Inspired by recent work on weakly supervised object localization [45, 7], we propose in particular to leverage selective spatial pooling with negative evidence proposed in [7] to improve visual feature extraction without resorting, *e.g.*, to expensive region proposal strategies. Another important benefit of the proposed joint architecture is that, once trained, it allows localization of arbitrary concepts within arbitrary images: Given an image and the embedding of a text (or any point of the embedding space), we propose a mechanism to compute a localization map, as demonstrated in Fig. 1.

The proposed modification to current approaches, along with additional design and training specifics, leads to a new system whose performance is assessed on two very different tasks. We first establish new state-of-the-art performance on cross-modal matching, effectively composed of two symmetric sub-tasks: Retrieving captions from query images and vice-versa. Without additional fine-tuning, our model with its built-in concept localization mechanism also outperforms existing work on the "pointing game" sentence-grounding task. With its state-of-the-art performance and its mechanism to localize even unseen concepts, our system opens up new opportunities for multi-modal content search.

The rest of the paper is organized as follows. We discuss in Section 2 the related works, on semantic-visual embedding and on weak supervised localization, and position our work. Section 3 is dedicated to the presentation of our own system, which couples selective spatial pooling with recent architectures and which relies on a triplet ranking loss based on hard negatives. We also show how it can be equipped with a concept localization module by exploiting without pooling the last feature maps in the visual path. More details on the system and its training are reported in Section 4, along with various experiments. On the competitive cross-modal retrieval task, our system is shown to outperform current state-of-the-art by a good margin. On the recently proposed task of pointing game, its localization mechanism offers new state-of-the-art performance with no need for retraining. In Section 5, we finally summarize the achievements of our work and outline perspectives.

2. Related Work and Paper Positioning

Deep learning offers powerful ways to embed raw data into high dimensional continuous representations that capture semantics. Off-the-shelf pretrained nets are now routinely used to extract versatile deep features from images [23, 36, 13] as well as from words and sentences [29, 32, 4, 25]. There are many strategies either to fine-tune these deep embeddings or to adapt them through new learned projections. In the following, we review learning methods to handle such mono/cross-modal representations, and we also highlight approaches dealing with spatial localization in this context.

Metric learning for semantic embedding One way to learn advanced visual representations is to consider the required transformation of the raw data as a metric learning problem. Several methods have been proposed to learn such metrics. In pairwise approaches, [43] minimizes the distance within pairs of similar training examples with a constraint on the distance between dissimilar ones. This learning process has been extended to kernel functions as in [28]. Other methods consider triplets or quadruplets of images, which are easy to generate from classification training datasets, to express richer relative constraints among groups of similar and dissimilar examples [40, 12, 3]. This kind of learning strategies has been also considered for deep (Siamese) architecture embeddings in the pairwise framework [37], and recently extended to triplets [15].

To embed words in a continuous space as vector representations, Mikolov *et al.*'s "word2vec" is definitively the leading technique [29]. In recent years, several approaches have been developed for learning operators that map sequences of word vectors to sentence vectors including recurrent networks [14, 4, 25] and convolutional networks [18]. Using word vector learning as inspiration, [21] proposes an objective function that abstracts the skip-gram word model to the sentence level, by encoding a sentence to predict the sentences around it.

In our work, we adopt most recent and effective deep architectures on both sides, using a deep convolutional network (ResNet) for images [13] and a simple recurrent unit (SRU) network [25] to encode the textual information. Our learning scheme is based on fine-tuning (on the visual side) and triplet-based optimization, in the context of cross-modal alignment that we describe now.

Learning cross-modal embedding The Canonical Correlation Analysis (CCA) method is certainly one of the first techniques to align two views of heterogeneous data in a common space [16]. Linear projections defined on both sides are optimized in order to maximize the cross correlation. Recently, non-linear extensions using kernel (KCCA [24]) or deep net (DCCA [1]) have been proposed. [38] exploit DCCA strategies for image-text embeddings, while [44] points out some limitations of this approach in terms of optimization complexity and overfitting and proposes ways to partially correct them. [9] proposes some CCA-based constraint regularization to jointly train two deep nets passing from one view to the other (text/image).

When considering the specific problem of embedding jointly images and labels (classification context), [41, 11] train models that combine a linear mapping of image features into the joint embedding space with an embedding vector for each possible class label. Approaches for the more advanced task of textual image description (captioning) often rely on an encoder/decoder architecture where the encoder consists of a joint embedding [20, 17]. Other works focus on the sole building of such a joint embedding, to perform image-text matching and cross-modal retrieval [11, 10, 27, 34].

Our work stems from this latter class. We aim at generating a joint embedding space that offers rich descriptors for both images and texts. We adopt the contrastive triplet loss that follows the margin-based principle to separate the positive pairs from the negative ones with at least a fixed margin. The training strategy with stochastic gradient descent has to be carefully adapted to the cross-modality of the triplets. Following [10], we resort to batch-based hard mining, but we depart from this work, and from other related approaches, in the way we handle localization information.

Cross-modal embedding and localization Existing works that combine localization and multimodal embedding rely on a two-step process. First, regions are extracted either by a dedicated model, *e.g.*, EdgeBox in [39], or by a module in the architecture. Then the embedding space is used to measure the similarity between these regions and textual data. [31, 17] use this approach on the dense captioning task to produce region annotations. It is also used for phrase localization by [39] where the region with the highest similarity with the phrase is picked.

To address this specific problem of phrase grounding, Xiao *et al.* [42] recently proposed to learn jointly a similarity score and an attention mask. The model is trained using a structural loss, leveraging the syntactic structure of the textual data to enforce corresponding structure in the attention mask.

In contrast to these works, our approach to spatial localization in semantic-visual embedding is weakly supervised and does not rely on a region extraction model. Instead, we take inspiration from other works on weakly supervised visual localization to design our architecture, with no need for a location-dependent loss.

Weakly supervised localization The task of generating image descriptors that include localization information has also been explored. A number of weakly supervised object localization approaches extrapolate localization features while training an image classifier, *e.g.*, [45, 7, 5]. The main strategy consists in using a fully convolutional deep architecture that postpones the spatial aggregation (pooling) at the very last layer of the net. It can be used both for classification and for object detection.

We follow the same strategy, but in the context of multimodal embedding learning, hence with a different goal. In particular, richer semantics is sought (and used for training) in the form of visual description, whether at the scene or at the object level.

3. Approach

The overall structure of the proposed approach, shown in Fig. 2, follows the dual-path encoding architecture of Kiros *et al.* [20]. We first explain its specifics before turning to its training with a cross-modal triplet ranking loss.

3.1. Semantic-visual embedding architecture

Visual path In order to accommodate variable size images and to benefit from the performance of very deep architectures, we rely on fully convolutional residual ResNet-152 [13] as our base visual network. Its penultimate layer outputs a stack of $D = 2048$ feature maps of size $(w, h) = (\frac{W}{32}, \frac{H}{32})$, where (W, H) is the spatial size of the input image. These feature maps retain coarse spatial information that lends itself to spatial reasoning in subsequent layers. Following the weakly supervised learning framework proposed by Durand *et al.* [7, 6], we first transform this stack through a linear adaptation layer of 1×1 convolutions. While in WELDON [7] and in WILDCAT [6] the resulting maps are class-related (one map per class in the former, a fixed number of maps per class in the latter), we do not address classification or class detection here.

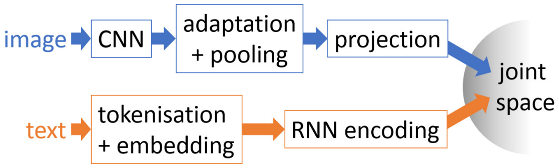

Figure 2. **Two-path multi-modal embedding architecture**. Images of arbitrary size and text of arbitrary length pass through dedicated neural networks to be mapped into a shared representation vector space. The visual path (blue) is composed of a fully convolutional neural network (ResNet in experiments), followed by a convolutional adaptation layer, a pooling layer that aggregates previous feature maps into a vector and a final projection to the final output space; The textual path (orange) is composed of a recurrent net running on sequences of text tokens individually embedded with an off-the-shelf map (word2vec in experiments).

Hence we empirically set the number D' of these new maps to a large value, 2400 in our experiments. A pooling à la WELDON is then used, but again in the absence of classes, to turn these maps into vector representations of dimension D'. A linear projection with bias, followed by ℓ_2 normalization accomplishes the last step to the embedding space of dimension d.

More formally, the visual embedding path is defined as follows:

$$\mathbf{I} \xmapsto{f_{\boldsymbol{\theta}_0}} \mathbf{F} \xmapsto{g_{\boldsymbol{\theta}_1}} \mathbf{G} \xmapsto{\text{sPool}} \mathbf{h} \in \mathbb{R}^{D'} \xmapsto{p_{\boldsymbol{\theta}_2}} \mathbf{x} \in \mathbb{R}^d, \quad (1)$$

where: $\mathbf{I} \in (0, 255)^{W \times H \times 3}$ is the input color image, $f_{\boldsymbol{\theta}_0}(\mathbf{I}) \in \mathbb{R}_+^{w \times h \times D}$ is the output of ResNet's `conv5` parematrized by weights in $\boldsymbol{\theta}_0$, $g_{\boldsymbol{\theta}_1}$ is a convolution layer with $|\boldsymbol{\theta}_1| = D \times D'$ weights and with activation in $\mathbb{R}^{w \times h \times D'}$, sPool is the selective spatial pooling with negative evidence defined in [7]:

$$\mathbf{h}[k] = \max \mathbf{G}[:, :, k] + \min \mathbf{G}[:, :, k], \ k = 1 \cdots D', \quad (2)$$

and $p_{\boldsymbol{\theta}_2}$ is an ℓ_2-normalized affine function

$$p_{\boldsymbol{\theta}_2}(\mathbf{h}) = \frac{A\mathbf{h} + \mathbf{b}}{\|A\mathbf{h} + \mathbf{b}\|_2}, \quad (3)$$

where $\boldsymbol{\theta}_2 = (A, \mathbf{b})$ is of size $(D' + 1) \times d$. We shall denote $\mathbf{x} = F(\mathbf{I}; \boldsymbol{\theta}_{0:2})$ for short this visual embedding.

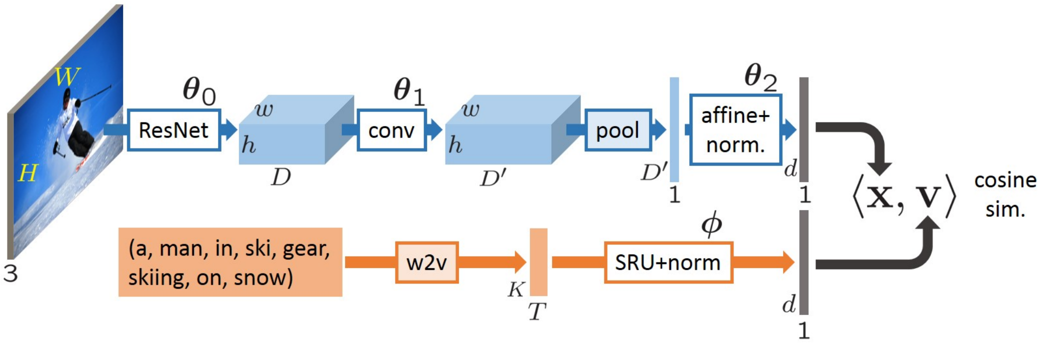

Figure 3. **Details of the proposed semantic-visual embedding architecture**. An image of size $3 \times W \times H$ is transformed into a unit norm representation $\mathbf{x} \in \mathbb{R}^d$; likewise, a sequence of T tokenized words is mapped to a normalized representation $\mathbf{v} \in \mathbb{R}^d$. Training will aim to learn parameters $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\phi})$ such that cross-modal semantic proximity translates into high cosine similarity $\langle \mathbf{x}, \mathbf{v} \rangle$ in the joint embedded space. Boxes with white background correspond to trainable modules, with parameters indicated on top. In our experiments, the dimensions are $K = 620$, $D = 2048$ and $D' = d = 2400$.

Textual path The inputs to this path are tokenized sentences (captions), *i.e.*, variable length sequences of tokens $S = (s_1 \cdots s_T)$. Each token s_t is turned into a vector representation $\mathbf{s}_t \in \mathbb{R}^K$ by the pre-trained word2vec embedding [29] of size $K = 620$ used in [21]. Several RNNs have been proposed in the literature to turn such variable length sequences of (vectorized) words into meaningful, fixed-sized representations. In the specific context of semantic-visual embedding, [20, 10] use for instance gated recurrent unit (GRU) [4] networks as text encoders. Based on experimental comparisons, we chose to encode sentences with the simple recurrent unit (SRU) architecture recently proposed in [25]. Since we train this network from scratch, we take its output, up to ℓ_2 normalization, as the final embedding of the input sentence. There is no need here for an additional trainable projection layer.

Formally, the textual path reads:

$$S \xmapsto{\text{w2v}} \mathbf{S} \xmapsto{\text{normSRU}_{\boldsymbol{\phi}}} \mathbf{v} \in \mathbb{R}^d, \tag{4}$$

where $\mathbf{S} = \text{w2v}(S) = \mathbb{R}^{K \times T}$ is an input sequence of text tokens vectorized with word2vec and $\mathbf{v}$ is the final sentence embedding in the joint semantic-visual space, obtained after ℓ_2-normalizing the output of SRU with parameters $\boldsymbol{\phi}$.

3.2. Training

The full architecture is summarized in Fig. 3. The aim of training it is to learn the parameters $\boldsymbol{\theta}_{0:2}$ of the visual path, as well as all parameters $\boldsymbol{\phi}$ of the SRU text encoder. The goal is to create a joint embedding space for images and sentences such that closeness in this space can be interpreted as semantic similarity. This requires cross-modal supervision such that image-to-text semantic similarities are indeed enforced.[1]

Contrastive triplet ranking loss Following [20], we resort to a contrastive triplet ranking loss. Given a training set $\mathcal{T} = \{(\mathbf{I}_n, S_n)\}_{n=1}^N$ of aligned image-sentence pairs – the sentence describes (part of) the visual scene – the empirical loss to be minimized takes the form:

$$\mathcal{L}(\boldsymbol{\Theta}; \mathcal{T}) = \frac{1}{N} \sum_{n=1}^{N} \Big(\sum_{m \in C_n} \text{loss}(\mathbf{x}_n, \mathbf{v}_n, \mathbf{v}_m) + \sum_{m \in D_n} \text{loss}(\mathbf{v}_n, \mathbf{x}_n, \mathbf{x}_m) \Big), \tag{5}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\phi})$ are the parameters to learn, $\mathbf{x}_n = F(\mathbf{I}_n; \boldsymbol{\theta}_{0:2})$ is the embedding of image n, $\mathbf{v}_n = \text{normSRU}_{\boldsymbol{\phi}}(\text{w2v}(S_n))$ is the embedding of sentence n, $\{S_m\}_{m \in C_n}$ is a set of sentences unrelated to n-th image, $\{\mathbf{I}_m\}_{m \in D_n}$ is a set of images unrelated to n-th sentence. The two latter sets are composed of negative ("constrastive") examples. The triplet loss is defined as:

$$\text{loss}(\mathbf{y}, \mathbf{z}, \mathbf{z}') = \max \big\{0, \alpha - \langle \mathbf{y}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z}' \rangle \big\}, \tag{6}$$

with $\alpha > 0$ a margin. It derives from triplet ranking losses used to learn metrics and to train retrieval/ranking systems. The first argument is a "query", while the second and third ones stand respectively for a relevant (positive) answer and an irrelevant (negative) one. The loss is used here in a similar way, but with a multimodal triplet. In the first sum of Eq. 5, this loss encourages the similarity, in the embedding space, of an image with a related sentence to be larger by a margin to its similarity with irrelevant sentences. The second sum is analogous, but centered on sentences.

[1] Note that mono-modal supervision can also be useful and relatively easier to get in the form, *e.g.*, of categorized images or of categorized sentences. Both are indeed used implicitly when relying on pre-trained CNNs and pre-trained text encoders. It is our case as well as far as the visual path is concerned. However, since our text encoder is trained from scratch, the only pure text (self-)supervision we implicitly use lies in the pre-training of word2vec.

Mining hard negatives In [20, 17], contrastive examples are sampled at random among all images (resp. sentences) in the mini-batch that are unrelated to the query sentence (resp. image). Faghri *et al.* [10] propose instead to focus only on the hardest negatives. We follow the same strategy: For each positive pair in the batch, a single contrastive example is selected in this batch as the one that has the highest similarity with the query image/sentence while not being associated with it. This amounts to considering the following loss for the current batch $\mathcal{B} = \{(\mathbf{I}_n, S_n)\}_{n\in B}$:

$$\mathcal{L}(\mathbf{\Theta};\mathcal{B}) = \frac{1}{|B|}\sum_{n\in B}\Big(\max_{m\in C_n\cap B}\text{loss}(\mathbf{x}_n,\mathbf{v}_n,\mathbf{v}_m) + \max_{m\in D_n\cap B}\text{loss}(\mathbf{v}_n,\mathbf{x}_n,\mathbf{x}_m)\Big). \quad (7)$$

Beyond its practical interest, this mining strategy limits the amount of gradient averaging, making the training more discerning.

3.3. Localization from embedding

As described in Section 2, several works on weak supervised localization [45, 7] combine fully convolutional architectures with specific pooling mechanisms such that the unknown object positions in the training images can be hypothesized. This localization ability derives from the activation maps of the last convolutional layer. Suitable linear combinations of these maps can indeed provide one heatmap per class.

Based on the pooling architecture of [7] which is included in our system and without relying on additional training procedures, we derive the localization mechanism for our semantic-visual embedding. Let's remind that in our case, the number of feature maps is arbitrary since we are not training on a classification task but on a cross-modal matching one. Yet, one can imagine several ways to leverage these maps to try and map an arbitrary vector of the joint embedding space into an arbitrary input image. When this vector is the actual embedding of a word or sentence, this spatial mapping should allow localizing the associated concept(s) in the image, if present. Ideally, a well-trained joint embedding should allow such localization even for concepts that are absent from the training captions.

To this end, we propose the following localization process (Fig. 4). Let $\mathbf{I}$ be an image and $\mathbf{G}$ its associated D' feature maps (Eq. 1). This stack is turned into a stack $\mathbf{G}' \in \mathbb{R}^{w\times h\times d}$ of d heatmaps using the linear part of the projection layer $p_{\mathbf{\theta}_2}$:[2]

$$\mathbf{G}'[i,j,:] = A\mathbf{G}[i,j,:],\ \forall(i,j)\in[\![1,w]\!]\times[\![1,h]\!], \quad (8)$$

which is a 1×1 convolution. Given $\mathbf{v}\in\mathbb{R}^d$ the embedding of a word or sentence (or any unit vector in the embedded space) and $K(\mathbf{v})$ the set of the indices of its k largest entries, the 2D heatmap $\mathbf{H}\in\mathbb{R}^{w\times h}$ associated with the embedded text $\mathbf{v}$ in image $\mathbf{I}$ is defined as:

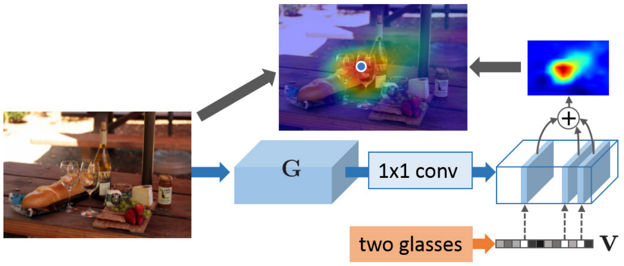

Figure 4. **From text embedding to visual localization.** Given the feature maps **G** associated to an image by our semantic-visual architecture and the embedding of a sentence, a heatmap can be constructed: Learned projection matrix A serves as a 1×1 convolution; Among the d maps thus generated, the k ones associated with the largest among the d entries of **v** are linearly combined. If the sentence relates to a part of the visual scene, like "two glasses" in this example, the constructed heatmap should highlight the corresponding location. Blue dot indicates the heat maximum.

$$\mathbf{H} = \sum\nolimits_{u\in K(\mathbf{v})}\big|\mathbf{v}[u]\big|\times\mathbf{G}'[:,:,u]. \quad (9)$$

In the next section, such heatmaps will be shown in false colors, overlaid on the input image after suitable resizing, as illustrated in Figs. 1 and 4. Note that [35] also proposes to build semantic heatmaps as weighted combinations of feature maps, but with weights obtained by back-propagating the loss in their task-specific network (classification or captionning net). Such heatmaps help visualize which image regions explain the decision of the network for this task.

4. Experiments

Starting from images annotated with text, we aim at producing rich descriptors for both image and text that live in the same embedding space. Our model is trained on the MS-COCO dataset, and benchmarked on two tasks. To evaluate the overall quality of the model we use cross-modal retrieval, and to assess its localization ability we tackle visual grounding of phrases.

4.1. Training

Datasets To train our model, we used the MS-COCO dataset [26][3]. This dataset contains 123,287 images (train+val), each of them annotated with 5 captions. It is originally split into a training set of 82,783 images and a validation set of 40,504 images. The authors of [17] proposed another split (called rVal in the rest of the paper) keeping from the original validation set 5,000 images for

[2] In other words, the pooling is removed. Bias and normalization being of no incidence on the location of the peaks, they are ignored.

[3] http://cocodataset.org

	visual	caption retrieval				image retrieval			
model	backend	R@1	R@5	R@10	Med. r	R@1	R@5	R@10	Med. r
Embedding network [39]	VGG	50.4	79.3	89.4	-	39.8	75.3	86.6	-
2-Way Net [8]	VGG	55.8	75.2	-	-	39.7	63.3	-	-
LayerNorm [2]	VGG	48.5	80.6	89.8	5.1	38.9	74.3	86.3	7.6
VSE++ [10]	R152	64.6	-	95.7	1	52.0	-	92.0	1
Ours	R152	**69.8**	**91.9**	**96.6**	1	**55.9**	**86.9**	**94.0**	1

Table 1. **Cross-modal retrieval results on** MS-COCO**.** On both caption retrieval from images and image retrieval from captions, the proposed architecture outperforms the state-of-the-art systems. It yields an R@1 relative gain of 38% (resp. 40%) with respect to best published results [39] on cross-modal caption retrieval (resp. image retrieval), and 8% (resp 7.5%) with respect to best online results [10].

validation and 5,000 for testing and using the remaining 30,504 as additional training data. To make our results comparable, we trained a model using each split. For evaluation, we also use the MS-COCO dataset, complemented with the annotations from Visual Genome dataset [22][4] to get localization ground-truth when needed.

Image pipeline The image pipeline is pre-trained on its own in two stages. We start from original ResNet-152 [13] pre-trained on ImageNet classification task. Then, to initialize the convolutional adaptation layer $g_{\boldsymbol{\theta}_1}$, we consider temporarily that the post-pooling projection is of size 1000 such that we can train both on ImageNet as well. Once this pre-training is complete, the actual projection layer $p_{\boldsymbol{\theta}_2}$ onto the joint space is put in place with random initialization, and combined with a 0.5-probability dropout layer. As done in [10], random rectangular crops are taken from training images and resized to a fixed-size square (of size 256 × 256).

Text pipeline To represent individual word tokens as vectors, we used pre-trained word2vec with no further fine-tuning. The SRU text encoder [25] is trained from scratch jointly with the image pipeline. It has four stacked hidden layers of dimension 2400. Following [25], 0.25-probability dropout is applied on the linear transformation from input to hidden state and between the layers.

Full model training Both pipelines are trained together with pairs of images and captions, using Adam optimizer [19]. Not every part of the model is updated from the beginning. For the first 8 epochs only the SRU (parameters ϕ) and the last linear layer of the image pipeline ($\boldsymbol{\theta}_2$) are updated. After that, the rest of the image pipeline ($\boldsymbol{\theta}_{0:1}$) is also fine-tuned. The training starts with a learning rate of 0.001 which is then divided by two at every epoch until the seventh and kept fixed after that. Regarding mini-batches, we found in contrast to [10] that their size has an important impact on the performance of our system. After parameter searching, we set this size to 160. Smaller batches result in weaker performance while too large ones prevent the model from converging.

4.2. Comparison to state-of-the-art

MS-COCO retrieval task Our model is quantitatively evaluated on a cross-modal retrieval task. Given a query image (resp. a caption), the aim is to retrieve the corresponding captions (resp. image). Since MS-COCO contains 5 captions per image, recall at r ("R@r") for caption retrieval is computed based on whether at least one of the correct captions is among the first r retrieved ones. The task is performed 5 times on 1000-image subsets of the test set and the results are averaged.

All the results are reported on Tab. 1. We compare our model with recent leading methods. As far as we know, the best published results on this task are obtained by the Embedding Network [39]. For caption retrieval, we surpass it by (19.4%,12.6%,7.2%) on (R@1,R@5,R@10) in absolute, and by (16.1%,11.6%,7.4%) for image retrieval. Three other methods are also available online, 2-Way Net [8], LayerNorm [2] and VSE++ [10]. The first two are on the par with Embedding Network while VSE++ reports much stronger performance. We consistently outperform the latter, especially in terms of R@1. The most significant improvement comes from the use of hard negatives in the loss, without them recall scores are significantly lower (R@1 - caption retrieval: -20,3%, image retrieval: -16.3%).

Note that in [10], the test images are scaled such that the smaller dimension is 256 and centrally cropped to 224 × 224. Our best results are obtained with a different strategy: Images are resized to 400 × 400 irrespective of their size and aspect ratio, which our fully convolutional visual pipeline allows. When using the scale-and-crop protocol instead, the recalls of our system are reduced by approximately 1.4% in average on the two tasks, remaining above VSE++ but less so. For completeness we tried our strategy with VSE++, but it proved counterproductive in this case.

Visual grounding of phrases We evaluate quantitatively our localization module with the pointing game defined by [42]. This task relies on images that are present both in MS-COCO val 2014 dataset and in Visual Genome dataset. The data contains 17,471 images with 86,5582 text region annotations (a bounding box associated

[4] http://visualgenome.org/

	caption retrieval				image retrieval			
model	R@1	R@5	R@10	Med. r	R@1	R@5	R@10	Med. r
Emb. network [39]	40.7	69.7	79.2	-	29.2	59.6	71.7	-
2-Way Net [8]	49.8	67.5	-	-	36.0	55.6	-	-
VSE++ [10]	52.9	-	87.2	1	**39.6**	-	79.5	2
DAN [30]	**55.0**	**81.8**	**89.0**	1	39.4	69.2	79.1	2
Ours (`MS-COCO` only)	46.5	72.0	82.2	2	34.9	62.4	73.5	3

Table 2. **Direct transfer to `Flickr-30K`, with comparison to SoA**. Although cross-validated and trained on `MS-COCO` only, our system delivers good cross-modal retrieval performance on `Flickr-30K`, compared to recent approaches trained on `Flickr-30K`: It is under the two best performing approaches, but above the two others on most performance measures.

Model	Accuracy
"center" baseline	19.5
Linguistic structure [42]	24.4
Ours (`train 2017`)	33.5
Ours (`rVal`)	33.8

Table 3. **Pointing game results**. Our architecture outperforms the state-of-the-art system [42] by more than 9% in accuracy, when trained with either `train` or `rVal` split from `MS-COCO`.

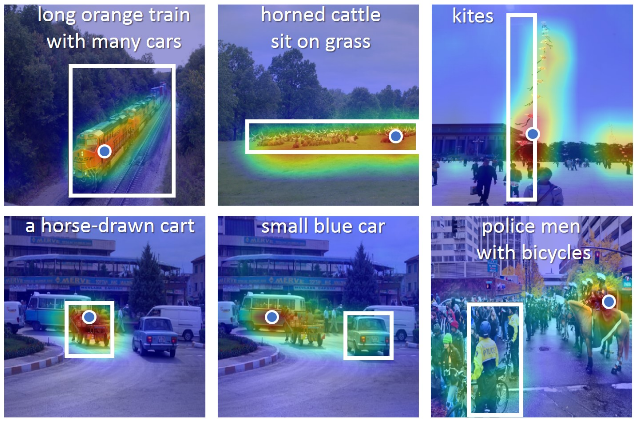

Figure 5. **Pointing game examples**. Images from the `Visual Genome` dataset overlaid with the heatmap localizing the input text according to our system. The white box is the ground-truth localization of the text and the blue dot marks the location predicted by our model for this text. The first four predictions are correct, unlike the last two ones. In the last ones, the heatmap is nonetheless active inside the ground-truth box.

with a caption). The task consists in "pointing" the region annotation in the associated image. If the returned location lies inside the ground-truth bounding box, it is considered as a correct detection, a negative one otherwise. Since our system produces a localization map, the location of its maximum is used as output for the evaluation. For this evaluation, the number of feature maps from $\mathbf{G}'$ that are used to produce the localization map was set through cross-validation to $k = 180$ (out of 2400). We keep this parameter fixed for all presented visualizations.

The quantitative results are reported in Tab. 3 and some visual examples are shown in Fig. 5. We add to the comparison a baseline that always outputs the center of the image as localization, leading to a surprisingly high accuracy of 19.5%. Our model, with an accuracy of 33.8%, offers absolute (resp. relative) gains of 9.4% (resp. 38%) over [42] and of 14% (resp. 73%) over the trivial baseline.

MS-COCO localization and segmentation Following the evaluation scheme for [42], we obtain similar semantic segmentation performance (namely mAP scores of 0.34, 0.24 and 0.15 for IoU@0.3, IoU@0.4 and IoU@0.5 resp.), while our localization module does not benefit from a training to structure the heatmaps. We also performed pointwise object localization on `MS-COCO` using the bounding box annotation, obtaining 57.4 mAP, an improvement of 4% compared to [6].

4.3. Further analysis

Transfer to Flickr30K We propose to investigate how our model trained on `MS-COCO` may be transferred as such to other datasets, namely `Flickr-30K` here. We report the results in Tab. 2. Not surprisingly, our performance is below the best systems [30, 10] trained on `Flickr-30K`. Yet, while not being trained at all on `Flickr-30K`, it outperforms on almost all measures two other recent approaches trained on `Flickr-30K` [8, 39]. Note that *fine-tuning* our system on `Flickr-30K` makes it outperform all, including [30, 10], by a large margin (not reported in Table for the sake of fairness).[5]

Towards zero-shot localization The good performance we obtain in the pointing game highlights the ability of our system to localize visual concepts based on their embedding in the learned joint space. We illustrate further this strength of the system with additional examples, like the one already

[5] We chose to keep the architecture used on `MS-COCO` as it is and to experiment with transfer and fine-tuning. An actual evaluation on `Flickr-30K` would require cross-validation of the various hyper-parameters. This dataset being substantially smaller than `MS-COCO`, such a task is challenging given the size of our architecture with its 2400 new feature maps and its large final embedding dimension of 2400.

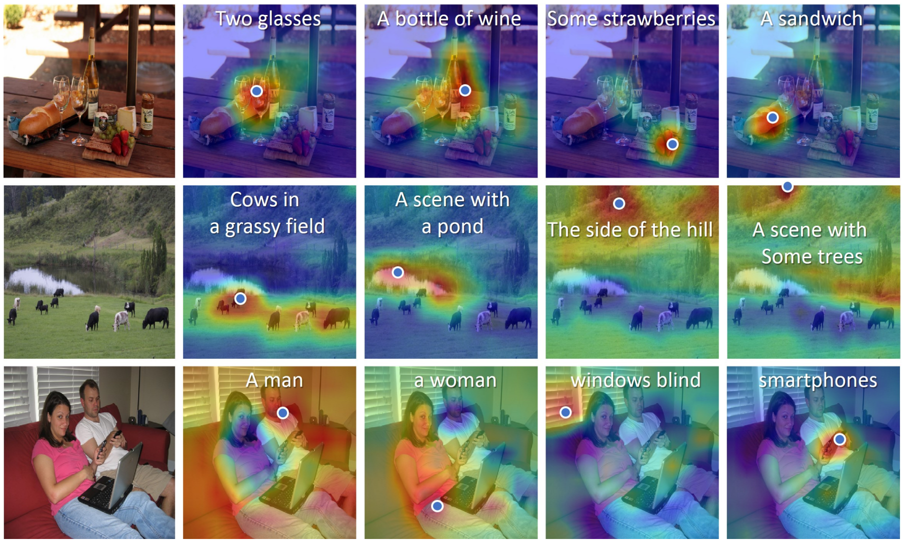

Figure 6. **Localization examples**. The first column contains the original image, the next columns show as overlays the heatmaps provided by the localization module of our system for different captions (superimposed). In each image the circled blue dot marks the maximum value of the heatmap.

presented in Fig. 1. We show in Fig. 6 the heatmaps and associated localizations for home-brewed text "queries" on images from MS-COCO test set. Going one step further, we conducted similar experiments with images from the web and concepts that were checked *not to appear in any of the training captions*, see Fig. 7.

Changing pooling One of the key elements of the proposed architecture is the final pooling layer, adapted from WELDON [7]. To see how much this choice contributes to the performance of the model, we tried instead the Global Average Pooling (GAP) [45] approach. With this single modification, the model is trained following the exact same procedure as the original one. This results in less good results: For caption retrieval (resp. image retrieval), it incurs a loss of 5.3% for R@1 (resp. 4.7%) for instance, and a loss of 1.1% in accuracy in the pointing game.

5. Conclusion

We have presented a novel semantic-visual embedding pipeline that leverages recent architectures to produce rich, comparable descriptors for both images and texts. The use of a selective spatial pooling at the very end of the fully convolutional visual pipeline allows us to equip our system with a powerful mechanism to locate in images the regions corresponding to any text. Extensive experiments show that our model achieves high performance on cross-modal retrieval tasks as well as on phrases localization. We also showed first qualitative results of zero-shot learning, a direction towards which our system could be pushed in the future with, among others, a deeper exploitation of language structure and of its visual grounding.

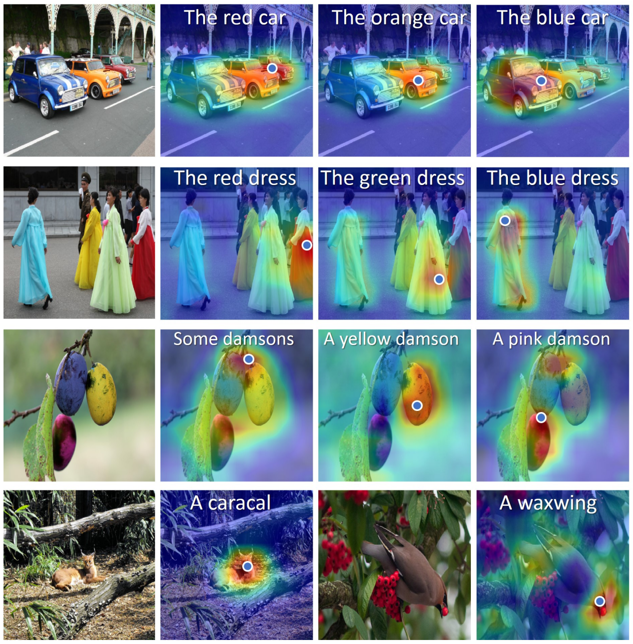

Figure 7. **Toward zero-shot localization**. The first three rows show the ability to differentiate items according to their colors, even if, as in third example, the colors are unnatural and the concept has not been seen at training. This example, and the two last ones could qualify as "zero-shot localization" as damson, caracal, and waxwing are not present in MS-COCO train set.

Acknowledgments This research was supported by the ANR-16-CE23-0006 and IBM Montpellier Cognitive Systems Lab with the loan of a PowerAI server.

References

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, 2013. 2

[2] J. L. Ba, J. R. Kiros, and G. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1, 6

[3] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11:1109–1135, 2010. 2

[4] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS w. on Deep Learning*, 2014. 2, 4

[5] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 3

[6] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *CVPR*, 2017. 3, 7

[7] T. Durand, N. Thome, and M. Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *CVPR*, 2016. 1, 3, 5, 8

[8] A. Eisenschtat and L. Wolf. Linking image and text with 2-way nets. *arXiv preprint arXiv:1608.07973*, 2016. 6, 7

[9] A. Eisenschtat and L. Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017. 2

[10] F. Faghri, D. Fleet, J. R. Kiros, and S. Fidler. VSE++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017. 1, 2, 3, 4, 5, 6, 7

[11] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *NIPS*, 2013. 1, 2

[12] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007. 2

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3, 6

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[15] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *ICLRw*, 2015. 2

[16] H. Hotelling. Relations between two sets of variates. *Biometrika*, 1936. 2

[17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 2, 3, 5

[18] Y. Kim. Convolutional neural networks for sentence classification. In *EMNLP*, 2014. 2

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[20] R. Kiros, R. Salakhutdinov, and R. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 1, 2, 3, 4, 5

[21] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *NIPS*, 2015. 2, 4

[22] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 6

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[24] P. L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *Int. J. Neural Syst.*, 2000. 2

[25] T. Lei and Y. Zhang. Training RNNs as fast as CNNs. *arXiv preprint arXiv:1709.02755*, 2017. 2, 4, 6

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 5

[27] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015. 2

[28] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 2

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2, 4

[30] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2017. 7

[31] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Hierarchical multimodal LSTM for dense visual-semantic embedding. In *CVPR*, 2017. 3

[32] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *EMNLP*, 2014. 2

[33] Z. Ren, H. Jin, Z. Lin, C. Fang, and A. Yuille. Joint image-text representation by gaussian visual-semantic embedding. In *ACMMM*, 2016. 1

[34] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *CVPR*, 2017. 2

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 5

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[37] Y. L. Sumit Chopra, Raia Hadsell. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 2

[38] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 2

[39] L. Wang, Y. Li, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *T-PAMI*, 2017. 1, 3, 6, 7

[40] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009. 2

[41] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011. 2

[42] F. Xiao, L. Sigal, and Y. Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, 2017. 3, 6, 7

[43] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng. Distance metric learning, with application to clustering with side-information. In *NIPS*, 2002. 2

[44] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015. 2

[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 3, 5, 8