Weak Supervision helps Emergence of Word-Object Alignment and improves Vision-Language Tasks

Corentin Kervadec, Grigory Antipov, Moez Baccouche, Christian Wolf

Weak Supervision helps Emergence of Word-Object Alignment and improves Vision-Language Tasks

Corentin Kervadec [1] and **Grigory Antipov** [2] and **Moez Baccouche** [3] and **Christian Wolf** [4]

Abstract. The large adoption of the self-attention (i.e. transformer model) and BERT-like training principles has recently resulted in a number of high performing models on a large panoply of vision-and-language problems (such as Visual Question Answering (VQA), image retrieval, etc.). In this paper we claim that these State-Of-The-Art (SOTA) approaches perform reasonably well in structuring information inside a single modality but, despite their impressive performances, they tend to struggle to identify fine-grained inter-modality relationships. Indeed, such relations are frequently assumed to be implicitly learned during training from application-specific losses, mostly cross-entropy for classification. While most recent works provide inductive bias for inter-modality relationships via cross attention modules, in this work, we demonstrate (1) that the latter assumption does not hold, i.e. modality alignment does not necessarily emerge automatically, and (2) that adding weak supervision for alignment between visual objects and words improves the quality of the learned models on tasks requiring reasoning. In particular, we integrate an object-word alignment loss into SOTA vision-language reasoning models and evaluate it on two tasks VQA and Language-driven Comparison of Images. We show that the proposed fine-grained inter-modality supervision significantly improves performance on both tasks. In particular, this new learning signal allows obtaining SOTA-level performances on GQA dataset (VQA task) with pre-trained models without finetuning on the task, and a new SOTA on NLVR2 dataset (Language-driven Comparison of Images). Finally, we also illustrate the impact of the contribution on the models reasoning by visualizing attention distributions.

1 Introduction

High-capacity deep neural networks trained on large amount of data currently dominate methods addressing problems involving either vision *or* language, or both of these modalities jointly. Examples for vision-language tasks are image retrieval task [15] (retrieve an image given a query sentence); image captioning [22] (describe the content of an input image in one or more sentences), and Visual Question Answering [2] (VQA: textually answer a question on an input image) *etc*. These tasks require different forms of reasoning, among which we find the capacity to analyze instructions – *e.g.* the question in VQA –, or the ability to fuse modalities or to translate one modality into another one – *e.g.* in image captioning. Additionally, they often require different levels of understanding, from a global image-text comparison to fine-grained object-word matchings.

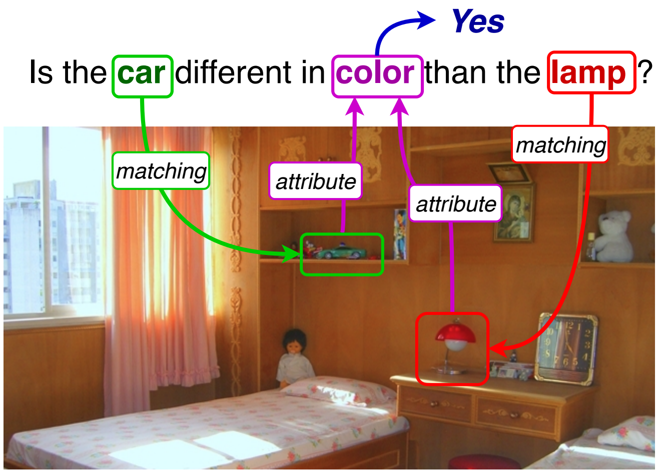

Figure 1. In the context of vision+language problems, we show that the alignment of visual objects to words in an input sentence does not naturally emerge from task oriented losses and propose a new auxiliary training objective addressing this problem. The presented image and question are taken from GQA [14] dataset.

In this context, a large panoply of high-performing models adopt self-attention architectures [35] and BERT-like [8] training objectives, which complement the main task-related loss with other auxiliairy losses correlated to the task. The common point of this large body of work is the large-scale training of unified vision-language encoders on image-sentence pairs. However, despite their ability to model interactions unique to one modality (i.e. *intra*-relationships), we observe that these State-Of-The-Art (SOTA) approaches tend to struggle to identify fine-grained object-word relationships (*inter*-relationships, or cross-modality relationships). These relationships are important, which can be illustrated in the example of VQA: answering a question given an input image requires the detection of certain objects in the image, which correspond to words in the question, and eventually the detection of more fine-grained relationships between visual objects, which are related to entities in the sentence.

In the literature, the alignment or matching of words to visual objects is generally assumed to be implicitly learned from application-specific losses — mostly cross-entropy for classification — thanks to the inductive biases provided by the encoder's architecture, i.e. the possibility of the model to *represent* this kind of matching. In this

[1] Université de Lyon, INSA-Lyon, LIRIS UMR CNRS 5205, Villeurbanne, France; Orange Labs, Cesson-Sévigné, France; corentin.kervadec@orange.com
[2] Orange Labs, Cesson-Sévigne, France; grigory.antipov@orange.com
[3] Orange Labs, Cesson-Sévigne, France; moez.baccouche@orange.com
[4] Université de Lyon, INSA-Lyon, LIRIS UMR CNRS 5205, Villeurbanne, France; christian.wolf@insa-lyon.fr

work we show that (1) modality alignment (*cf.* Figure 1) does not necessarily emerge automatically and (2) that adding weak supervision for alignment between visual objects and words improves the quality of the learned models on tasks requiring visual reasoning.

Our contributions are as follows:

- We enhance vision-language encoder approaches by adding explicit weak supervision of object-word alignment, taking into account the uncertainty present in the detection result of the vision module.
- We improve the accuracy of SOTA vision-language models on the VQA task GQA [14] dataset) *without the need of finetuning*, to achieve SOTA-level results. In other words, with our new objective, pre-training is sufficient for SOTA results.
- On the task of Language-driven Comparison of Images, requiring to reason over two images and one sentence, our proposed model outperforms the current SOTA model on the challenging NLVR2 dataset.
- We show visualizations of attention maps, which corroborate the claim that word-object alignment does not naturally emerge from task losses, while it is discovered by our weak supervision signal.

2 Related Works

Vision-language tasks – *Vision and language understanding* is a broad area and can take several forms at many different levels of granularity. Some tasks focus on matching problems, as for instance *Image Retrieval*, which requires finding the most relevant image given a query sentence [15], [20]. The inverse problem — namely *Sentence Retrieval* — has also been explored [15]. A similar task with finer granularity is *Visual Grounding*, where the model must associate image regions to words or sentences [16], [28].

Other tasks require more high-level reasoning over images and sentences, which, in general, requires multi-modal interactions but also the ability to compare, count or find relations between objects in the image. In *Visual Question Answering* (VQA) [2] [14] we ask questions (given as input) about an input image and the model must predict the answer. Answering the questions requires a variety of skills: finding relations, counting, comparing colors or other visual features, materials, sizes, shapes, *etc*. The binary task of *Language-driven Comparison of Images* takes as input triplets $(img_1, img_2, sentence)$ and requires predicting whether the sentence truly describes the image pair [32].

Finally, some tasks involve the generation of one modality from the other. *Image captioning* consists in translating an image into text [22]. Similarly, some tasks aim to generate questions about an image [21]. Inversely, it is also possible to generate an image from a caption [24]. However, such multimodal generation is out of the scope of our work.

Vision-language multi-modal fusion – Early work in vision and language understanding focused on separate models for each modality followed by multi-modal fusion [29]. In this context, bi-linear fusion is an expressive family of models, which, however, suffers from overparametrization and therefore overfitting. Subsequent work addressed this by creating low-rank decompositions of the fusion tensors, either through Tucker tensor compositions as in MUTAN [5], or block tensor decompositions like in BLOCK [6].

However the general tendency is to move towards holistic architectures, modeling all the interactions between modalities, and also between different objects in the visual modality. Object level reasoning, i.e. the analysis of visual data in the form of a collection of previously detected local entities/objects, has become a general tendency in computer vision beyond VQA, also seen in video analysis [4] etc. In this context, the Relation Network [31] considers all the pairwise interactions between visual objects. [34], [25] and [23] apply variants of Graph Convolutional Network [18] on visual objects and question words for VQA. [38] and [9] go a step further by modeling multimodal interactions via adapting transformer [35] principles to vision and language. We call them holistic because they consider both intra-modality (inside a modality) and inter-modality (fusion between modalities) relationships.

Multi-task pretraining – A second tendency is the evolution of training from task-specific supervision signals to a set of different losses, which are related to general vision-language understanding, and whose supervision signal can successfully be transferred to different downstream tasks.

This use of auxiliary tasks and knowledge transfer can be performed on both modalities: on language, the use of word embeddings such as GloVe [26] or BERT [8] is frequent. On vision, usual objectives include the use of pre-trained object detectors such as Faster RCNN [30], BUTD[1] for VQA and image captioning, and SCAN [20] for image retrieval.

More recent work shows that a joint pre-training over both modalities can benefit downstream vision-language tasks. This is achieved by setting up strategies to learn a vision-language representation in a multi-task fashion similar to BERT [8] in Natural Language Processing (NLP). Thereby, LXMERT [33] and VilBERT [23] use holistic architectures to learn a vision-language encoder trained on a large-scale amount of images-sentences pairs. Their encoder is then transferred to specific vision-language tasks, where they generally achieve SOTA results.

Symbolic representation for visual reasoning – Aside from these approaches, others address the visual reasoning problem by constructing a symbolic view of vision, language and of the reasoning process. Thus, [37] uses reinforcement learning to learn a program generator predicting a functional program from a given question in order to model the reasoning. NSM [12] predicts a probabilistic graph from the image to obtain an abstract latent space which is then processed as a state machine.

Our work follows tendencies in SOTA and is based on a holistic joint vision-language encoder with multiple training objectives. We improve on the SOTA by adding an additional objective, which (weakly) supervises the model to align objects and words referring to the same entity. This new objective complements the inductive bias inherent in current models, which allows learning of cross-modality relationships.

3 Vision-Language Encoder

In this Section, we present our vision-language encoder which is used for learning multimodal embeddings. The encoder is built upon the recent work, and in particular [38]. The overall architecure of our model is presented in Figure 2. Below, we firstly present the embeddings extraction part of our encoder (Section 3.1), and then focus on its central part (Section 3.2).

3.1 Input Embeddings

Vision Input – On the vision side, we use an object detector – Faster-RCNN [30] – to extract object level-visual features from the input image as in [1]. Similar to hard attention mechanisms, this enforces the system to reason on object level rather than on the pixel

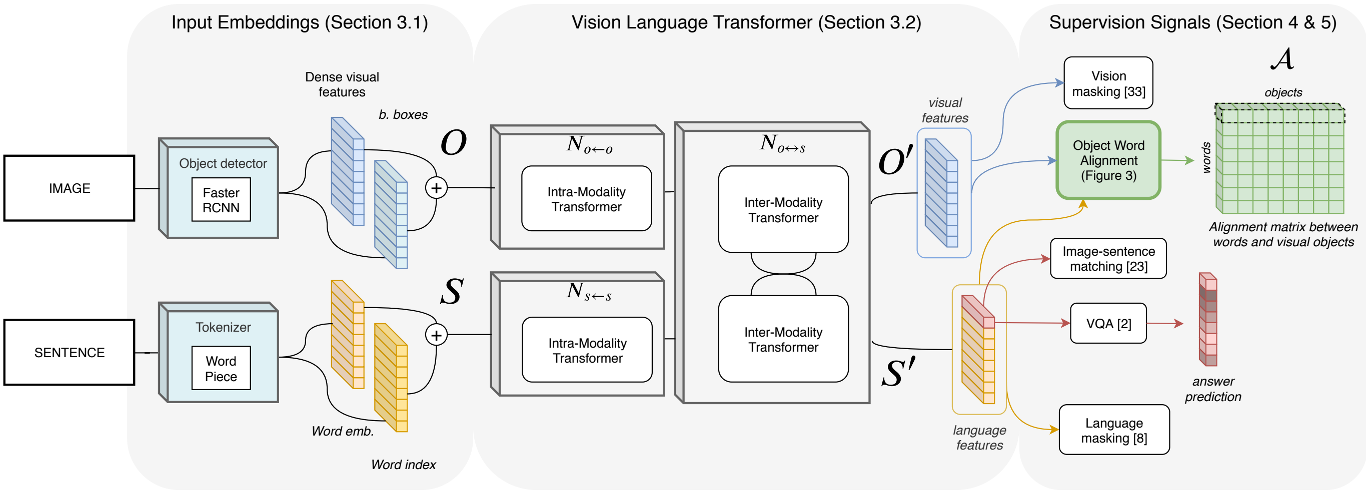

Figure 2. Architecture of our vision-language encoder (Section 3) and the respective supervision tasks (Sections 4 and 5). From left to right: (left), image and sentence are embedded into *input embeddings* (O, S) (Section 3.1); (middle), they are then processed by the *vision-language transformer* (Section 3.2) modeling both intra- and inter-modality interactions in order to obtain multimodal embeddings (O', S'); (right), the encoder is supervised using objectives introduced in Section 4 and our soft object-word alignment proposed in Section 5.

level or global level. Thus, for each image we extract N_o=36 bounding boxes and associated 2048-dimensional visual features:

$$(f, b) = RCNN(I), \tag{1}$$

where I is the input image and f=$\{f_0, ..., f_{N_o-1}\}$ and b=$\{b_0, ..., b_{N_o-1}\}$ are, respectively, the dense visual features and the bounding boxes detected for objects. Box and dense vectors are fused to obtain position-aware object level embeddings $O = [o_0, \ldots, o_i, \ldots, o_{N_o-1}]$.

Language Input – On the language side, sentences are tokenized using the WordPiece tokenizer [36]. As common in language processing, a special token `[CLS]` is added at the beginning of the tokenized sentence, which encodes the multimodal information of the image and sentence. The transformation of this token, performed during the forward pass through the network, corresponds to the prediction of the answer to the task. Tokens are embedded into d-dimensional vectors using a look-up table learned during a pre-training phase which concentrates on language only. The index position of the word is added to the dense vector as a positional encoding in order to obtain index-aware word level embeddings $S = [s_0, \ldots, s_i, \ldots, s_{N_s-1}]$.

3.2 Vision-Language Transformer

The neural model encodes the independent vision and language embeddings (O, S) described above and transforms them as they pass through the network:

$$O', S' = Encoder(O, S), \tag{2}$$

where (O', S') are the updated output embeddings. We resort to the widely-used transformer architecture [35] adapted to vision-language problems as in [9] and [38].

The vision-language transformer is composed of two self-attention modules of type *intra-modality transformer* and *inter-modality transformer* as defined in [9]. They take as input one input sequence (in case of intra-modality) or two input sequences (in case of inter-modality) and calculate an output sequence:

$$x_i = T(q_i, k_j, v_j) = \sum_j \alpha_{ij} v_j, \tag{3}$$

where q_i, k_j and v_j are, respectively the *query*, *key* and *value* vectors in R^d [35], which are calculated as linear mappings from the input sequences. Their exact definition depends on the type (inter vs. intra) and is given further below.

The α_{ij} represent an attention map, which predicts how the different elements of the input sequences attend to each other:

$$\alpha_{ij} = softmax_i(\frac{q_i^T k_j}{\sqrt{d}}), \tag{4}$$

where d is the number of dimensions of the embedding space. As in [35], we use multi-head attention, where the embeddings are split into H parts, transformations are calculated in parallel, and predictions concatenated. Each transformer layer is followed by a residual connection, a layer normalization [3], and a feed-forward layer.

Intra-Modality Transformer blocks – allow to model the interactions inside one modality. Thus, their *query*, *key* and *value* vectors come from the same modality. They are defined as follows:

$$s' = T_s(s^q, s^k, s^v) \tag{5}$$

$$o' = T_o(o^q, o^k, o^v), \tag{6}$$

where $x^q = W^q x$, $x^k = W^k x$ and $x^v = W^v x$.

Inter-Modality Transformer blocks – model information flowing between both modalities vision and language. They are defined similarly to the intra-modality transformer but the *key* and *value* vectors are crossed between the modalities:

$$s' = T_{s \leftarrow o}(s^q, o^k, o^v) \tag{7}$$

$$s'' = T_s(s'^q, s'^k, s'^v) \tag{8}$$

$$o' = T_{o \leftarrow s}(o^q, s^k, s^v) \tag{9}$$

$$o'' = T_o(o'^q, o'^k, o'^v) \tag{10}$$

Stacked Architecture – The vision-language transformer is built by stacking the previously defined modules as shown in Figure 2: the image/sentence input data is passed through detectors/tokenizers, embedding extractors, several intra-modality transformer blocks and finally several cross-modality transformer blocks. Summarizing, the model contains $N_{o \leftarrow o}$ and $N_{s \leftarrow s}$ stacked intra-modality transformer followed by $N_{o \leftrightarrow s}$ inter-modality transformers.

The vision-language transformer provides inductive biases for modeling intra-modality relationships (*e.g.* sentence dependency graph for language, scene graphs for vision) and inter-modality relationships (*e.g.* vision-language alignment). However, as shown below, inductive biases are not sufficient for learning inter-modal interactions, it is therefore necessary to define adequate supervision signals.

4 Supervised Objectives

We train the vision-language encoder defined in Section 3 following the recently widely-adapted strategy of combining BERT-like [8] self-supervised signals with task-specific supervision signals, which has been applied to various problems in vision and language — *e.g.* [33] [23]. We select four supervision signals: vision masking [33], language masking [8], image-sentence matching [23] and visual question answering [2], which are briefly described below.

Vision/Language Masking – This signal aims to supervise the encoder's ability to reconstruct missing information in language and vision. More precisely, we randomly mask each language token (resp. visual object) with a probability of 0.15 and ask the model to predict the missing words (resp. objects). Therefore we add two classifiers – for *vision masking*[5] and *language masking* – on top of the vision language encoder and supervise via a cross-entropy loss. [33] proposes to take the object detector prediction as ground truth in order to get over the disparity of visual annotation. Additionally, we also supervise the model to regress the masked objects' Faster-RCNN features via L2 loss.

Image-Sentence Matching – BERT [8] proposes *next sentence prediction* supervision by asking to predict if two sentences are consecutive in a given text, or randomly sampled from a corpus. Its vision-language equivalent is *image-sentence matching* [33] [23], where the model has to predict whether a given sentence matches a given image or not. Thus, we randomly replace the image in each sentence-image pair with a probability of 0.5. We add a feed-forward layer on top of the `[CLS]` output embedding to predict whether the pair matches or not. This global matching is supervised using a binary cross-entropy loss.

Visual Question Answering – Our model is applicable to a wide range of vision-language problems (in Section 6 we evaluate it to two different tasks, namely VQA and Language-driven Comparison of Images). At the same time, independently of the target vision-language task, pretraining on VQA helps reasoning as shown in [33]. The VQA task is defined as a classification problem over a set of most frequent answers. In this work, we perform this classification from a prediction head attached to the `[CLS]` token and supervise it using a cross-entropy loss.

[5] It is worth noticing that the vision masking requires to predict both the object classes and its attributes (*e.g.* color, materials, *etc.*)

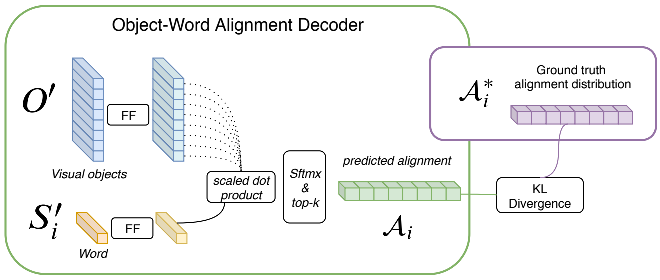

Figure 3. The proposed vision-language alignment decoder and the respective weakly-supervised loss. In this illustration, we present the alignment prediction $\mathcal{A}_i$ between one word S'_i and the visual objects O'. FF stands for feed-forward layers.

5 Weak Supervision of Object-Word Alignment

The presented SOTA vision-language supervision signals – *i.e.*, *vision/language masking*, *image-sentence matching* and *VQA* – have proved their efficiency at encoding rich vision-language embeddings [33] [23]. However, none of them explicitly supervises the object-word alignment. At the same time, matching words and visual objects referring to a same high-level entity is a natural prerequisite for visual reasoning.

The reason why such supervision has not been proposed before is probably that this fine-grained matching property is frequently assumed to be implicitly learned via inter-modality attention modules training. In this work, we claim that the word-object alignment does not necessarily emerge automatically but rather requires explicit supervision.

Vision-Language Alignment Decoder – We propose to add a vision-language alignment decoder on top of the encoder. The whole model is then supervised to predict the object-word alignment matrix $\mathcal{A}$ from the encoder's outputs (O', S'). First, (O', S') are projected into a joint space using a feed-forward layer with layer normalization [3] and residual connection. We obtain $(\hat{O}, \hat{S})$ from which we compute $\mathcal{A}$:

$$\mathcal{A} = \frac{\hat{S} \otimes \hat{O}}{\sqrt{d}}, \tag{11}$$

where $\otimes$ is the outer product. In other words, the alignment scalar $\mathcal{A}_{ij}$ is computed as the scaled-dot-product between each object-word pair (o_i, s_j), as shown in Figure 3:

$$\mathcal{A}_{ij} = \frac{\hat{s}_i . \hat{o}_j^T}{\sqrt{d}} \tag{12}$$

For each word s_i we only keep the top-k highest predictions and apply a softmax:

$$\mathcal{A}_i = softmax_j(top_k(\mathcal{A}_{ij})) \tag{13}$$

In this work, we empirically set $k = 3$. This way, we compute from each word a probability distribution $\mathcal{A}_i$ over the set of visual objects detected by Faster-RCNN. A high probability $\mathcal{A}_{ij}$ means word s_i and object o_j refer to the same high-level entity. The dedicated loss L_{align} is defined using Kullback-Leibler (KL) divergence:

$$L_{align} = KL(\mathcal{A}^*, \mathcal{A}), \tag{14}$$

Soft Alignment Score: approximating $\mathcal{A}^*$ – Let's suppose we have the ground truth object-word pair $(s_i, b^*_{s_i})$ (*cf.* Section 6.1). This pair is composed of a word or group of words s_i taken from the input sentence and a bounding box $b^*_{s_i}$ indicating the position of the respective object in the image. However we cannot directly use this supervision because both ground truth object-word annotations and the object detector are imperfect. More precisely, (1) the ground truth visual-object annotation is often misaligned with the object detection's bounding box prediction, or (2) the annotated object can simply be not detected at all.

To address this issue we set up a soft-alignment score taking into account both the detection-annotation misalignment and the object detector imperfection. To this end, we consider two criteria: the position one and the semantic one.

Position Criterion - For each ground truth object-word pair $(s_i, b^*_{s_i})$, we compute Intersection over Union (IoU) between object detector's predicted bounding box b_{o_j} and the ground truth object's bounding box $b^*_{s_i}$:

$$P\mathcal{A}^*_{ij} = IoU(b^*_{s_i}, b_{o_j}), \tag{15}$$

A high IoU leads to a high criterion value. Therefore, this criterion permits to give more importance to objects detected in the same image region as the ground-truth object.

Semantic Criterion - At the same time, we cannot only rely on positional information. Indeed, we also have to take into account the semantics of the object detector's prediction. This would avoid to align a word with a well-localized but a semantically-different object (according to the detector). Therefore we define the semantic criterion which computes the semantic similarity between a word s_i and the object's class c_{o_j} – and attribute a_{o_j} – predicted by the detector:

$$S\mathcal{A}^*_{ij} = \frac{3}{4}S(s_i, c_{o_j}) + \frac{1}{4}S(s_i, a_{o_j}), \tag{16}$$

where $S(\cdot, \cdot)$ compute the cosine similarity between the GloVe embeddings of the class/attribute names. We bias the similarity toward object class as we empirically found it more relevant than the attribute prediction.

Finally, we combine the two criteria in order to obtain a soft alignment score for each object-word pair in the annotation:

$$\mathcal{A}^*_{ij} = \frac{norm_j(P\mathcal{A}^*_{ij}) + norm_j(S\mathcal{A}^*_{ij})}{2} \tag{17}$$

The resulting soft-alignment scores are normalized over the objects such as:

$$\sum_{j}^{n_{objects}} \mathcal{A}^*_{ij} = 1 \tag{18}$$

Hence the ground truth soft alignment score $\mathcal{A}^*_i$ of word s_i is a probability distribution over the set of visual objects detected by the object detector.

The soft alignment score defined in this Section is by construction approximate. Therefore, we refer to the designed supervision signal as weak.

6 Experiments

In this section we evaluate the learned vision-language encoder on two tasks, namely VQA [2] and the Language-driven Comparison of Images [32].

6.1 Training Data

Following [33], we construct our dataset as the concatenation of the two public ones, namely: MSCOCO [22] and Visual Genome [19]. These datasets provide images annotated with captions. The VQA annotations are taken from three datasets (based on images from either MSCOCO or Visual Genome): VQA 2.0 [10], GQA [14] and VG-QA [19]. Consequently, our dataset is composed of 9.18M image-sentence pairs (a sentence can be either a caption or a question).

The object-word alignment scores, which are defined in Section 5, are calculated based on the annotations extracted from GQA and Visual Genome. In GQA dataset, salient question words and answers are annotated with visual pointers. A visual pointer consists in a bounding box corresponding to the visual region described by the words composing the question or the answer. Nevertheless, as GQA represents only 12% of the dataset, the use of the GQA pointers would have been insufficient.

To alleviate this issue, we augment the pointer annotation with visual grounded annotations from Visual Genome. Every Visual Genome image is accompanied with visual region descriptions forming *(description, bounding box)* pairs. Unlike in GQA, descriptions are full descriptive sentences and not small groups of words. Therefore, the so obtained pointer is less discriminative towards the language part. Therefore, we choose to combine these descriptions in order to obtain sentences with one, two or three pointers. For instance the two descriptions *"the cat playing near the tree"* and *"the yellow bird"* become *"the cat playing near the tree and the yellow bird"*, with the associated bounding boxes.

All in all, by combining annotations from GQA and Visual Genome, we gather roughly 6M image-sentence pairs annotated with pointers. In other words, about 70% of the total number of the image-sentence pairs in the dataset have fine-grained object-word alignment annotations.

6.2 Evaluation Tasks

To evaluate the reasoning quality of our vision-language encoder supervised with the object-word alignment, we evaluate it on two tasks requiring to reason over image and text.

The first one is the VQA task. It consists in predicting the answer asked about an image. This task is challenging as it requires a high-level understanding of vision and language. For evaluation, we select the most recent and, arguably, the most challenging VQA-dataset today, namely GQA. As GQA is already used during the vision-language encoder training, we do not find it necessary to finetune our model on the datatset.

The second is the Language-driven Comparison of Images. We choose the Natural Language for Visual Reasoning (NLVR2) dataset [32]. NLVR2 is composed of triplets $(img_1, img_2, sentence)$ where img_1 and img_2 are two images and $sentence$ is a sentence describing one or both images. The goal is to predict if the sentence is true. It is worth noticing that NLVR2 data is not viewed during the encoder training, therefore it truly evaluates the generalization capacity of our method. We use the same finetuning strategy as in [33]. Thus we concatenate the two encoder's output $[CLS]$ embeddings – obtained with $(img_1, sentence)$ and $(img_2, sentence)$ pairs – and pass them through a feed-forward layer. We then use a binary cross-entropy loss.

Table 1. Evaluation of the proposed object-word alignment weak supervision on the GQA [14] dataset. The presented results are calculated on the dataset's test-std split. The GQA's accuracy is presented in the last column. The exact definitions of all other (auxiliary) metrics can be found in [14].

Models	Binary	Open	Validity	Plausibility	Consistency	Distribution	Acc.
Human [14]	91.2	87.4	98.9	97.2	98.4	-	89.3
BUTD [1]	66.6	34.8	96.2	84.6	78.7	6.0	49.7
MAC [13]	71.23	38.9	96.2	84.5	81.6	5.3	54.1
LCGN [11]	73.7	42.3	**96.5**	**84.8**	84.7	4.7	57.0
LXMERT [33]	77.2	45.5	96.4	84.5	89.6	5.7	60.3
NSM [12]	**78.9**	**49.3**	96.4	84.3	**93.3**	**3.7**	**63.2**
ours	**76.9**	**46.1**	**96.3**	**84.7**	**89.7**	**5.3**	**60.5**

6.3 Results

Training Details – We train our vision language encoder using Adam optimizer [17] during 20 epochs. However, the VQA supervision is only added after 10 epochs, following [33]. We set the learning to 10^{-4} with warm starting and learning rate decay. The batch size is 512. On the architecure side, the number of stacked inter- and intra-modality transformers is $N_{o \leftrightarrow s} = 5$, $N_{o \leftarrow o} = 5$ and $N_{s \leftarrow s} = 9$. The encoder hidden's vectors are of dimension $d = 768$ and each attention layer has H=12 heads. Moreover, to reduce computation, we set the maximum sentence length to $N_s = 20$ tokens and the number of visual objects to $N_o = 36$. Training is done on four V100 GPUs.

For NLVR2 [32], we finetune during 4 epochs using Adam optimizer [17]. The learning rate is set to $5 * 10^{-5}$ and the batch size is 32. We only supervise with the task-specific binary objective, *i.e.*, we drop all the supervision signals used for encoder training.

Table 2. Impact of the proposed object-word alignment weak supervision on the VQA task. The presented results are calculated on the GQA [14] test-std split.

Models	Consistency	Accuracy
ours (w/o alignment supervision)	79.5	54.9
ours (with alignment supervision)	**89.7**	**60.5**

Visual Question Answering – Table 1 compares the results of applying our vision-language encoder on the VQA task versus the recent published works. As one may observe, our model obtains the 2nd-best SOTA-result, just after the NSM model [12]. The latter is fundamentally different from our approach (contrary to NSM, our model operates on raw images rather than on the scene graphs). Moreover, it is important to highlight that, unlike previous work, our model has not been finetuned on the target dataset making the obtained result even more significant.

In order to quantify the impact of the our object-word alignment weak supervision on the VQA task, we evaluate the two versions of our model, with and without the proposed loss, on the GQA dataset. The results are reported in Table 2. One may observe that the proposed weak supervision boosts the accuracy with $+5.6$ points. Moreover, when we focus on the metric which explicitly measures the reasoning capacity of the model, namely the consistency, our weakly-supervised alignment allows to gain more than $+10$ points. This demonstrates that, by enforcing the model to explicitly align words with visual objects, we obtained a finer multimodal representation.

Natural Language for Visual Reasoning (NLVR2) – As shown in Table 3, our method outperforms the published[6] SOTA accuracy on NLVR2 with a gain of $+1$ point. Furthermore, we have performed the same ablation analysis as for the VQA task (i.e. with and without the object-word alignment weak supervision), and the obtained results are summarized in Table 4. These results are coherent with those calculated on the VQA task confirming the advantage of the proposed supervision. Note that the scores in Table 4 are reported both for unbalanced and balanced subsets of the NLVR2 dataset. This split takes into account the visual biases present in the dataset. The benefit of our fine-grained alignment supervision method is constant between both subsets, showing that the gain is not bias-related.

Table 3. Evaluation of the proposed object-word alignment weak supervision on the NLVR2 evaluation splits. Models marked with * have been ran by the authors of [32].

Models	Dev.	Test-P
MAC* [13]	50.8	51.4
FiLM* [27]	51.0	52.1
CNN+RNN* [32]	53.4	52.4
MaxEnt [32]	54.1	54.8
LXMERT [33]	74.9	74.5
ours	**75.8**	**75.5**

Table 4. Impact of the proposed object-word alignment weak supervision on the Visual Reasoning grounded by Language task. The presented results are calculated on the Test-P part of the NLVR2 dataset.

Models	Test-P	Unbalanced	Balanced
ours (w/o alignment supervision)	74.5%	76.0%	73.1%
ours (with alignment supervision)	**75.5%**	**77.2%**	**74.5%**

7 Visualizing Reasoning

The reasoning capabilities of high-capacity deep neural networks are notoriously difficult to interpret, as inputs and intermediate activations are embedded in high-dimensional spaces in non trivial applications. Vision-language models are no exceptions, therefore we propose visualizations of some of the key activations of the proposed model. Such visualizations — when wisely chosen — can be a step toward more interpretable models, especially in the field of visual reasoning, where distinguishing real reasoning (i.e. which follows causal chains) from educated guesses (i.e. exploiting subtle statistic

[6] At the time of the writing, an unpublished work, called UNITER [7], reported a better result on NLVR2.

Figure 4. Visualization of the attention maps of the penultimate (=4th) inter-modality transformer. Word-object alignment does not emerge naturally for the baseline (without object-word alignment supervision), whereas our model with the proposed weakly-supervised objective learns to pay strong cross-attention on co-occurring combinations of words and objects in the scene. In the attention maps, rows represent words and columns represent visual objects. For the sake of visibility, we display the bounding box of the detected object with the highest activation regarding to the selected word. The predicted answer (underlined) is written after the question. Its corresponding language token is $[CLS]$, *i.e.*, the first row in attention maps.

biases in the data) can be difficult. The visualizations in this Section are obtained from the dev-test split of the GQA [14] dataset.

We inspect the attention maps inside the inter-modality transformers, which illustrates the information flow between the two modalities (vision and language). Generally, attention maps convey information on the importance that a neural map poses on local areas in input or activations. In the particular case of our model, the inter-modalilty attention map visualize how modalities are fused by the model, as they give weight to outputs for a given word as a function of a given object (or vice-versa).

Following equation (4), we visualize the values α_{ij}, showing the attention given to the pair (s_i, o_j). We visualize the map of the 4^{th} inter-modality transformer and sum the maps over the 12 parallel attention heads, comparing the maps of our proposed model with and without the proposed object-word alignment supervision in Figure 4.

The effectiveness of the new object-word alignment objective is corroborated by attention units which are higher for object-word pairs referring to the same entity in our model. We observe a radically different behavior in the baseline's attention maps, where attention is less-fine grained: roughly uniform attention distributions indicate that the layer outputs of all words attend to roughly the same objects.

Caveat: we do not want to imply, that the exact word-object alignment in the inter-modality layer is indispensable for a given model to solve a reasoning task, as a complex neural network can model relationships in the data in various different layers. However, we do argue, that some form of word-object alignment is essential for solving vision-language tasks, as the model is required to query whether concepts from the question are present in the image, and eventually query their relationships to other concepts. Inductive bias has been added to model for this type of reasoning in the form of inter-modality layers, and it is therefore natural to inspect whether this cross-attention emerges at this exact place. We would also like to point out that we do not force or favor word-object alignment at a specific layer, as our proposed objective is injected through a new module attached to the inter-modality layer (see Figure 2). The attention maps show that the objective is successfully propagated from the new alignment head to the inter-modality layer.

Conclusion

In this work, we design a vision-language encoder and train it with a novel object-word alignment weak supervision. To this end, we carefully design a soft alignment signal taking into account both spatial and semantic alignment between the words and the detected visual objects. We empirically show the benefit of this new supervision on two tasks requiring to reason over images, namely the VQA and the Language-driven Comparison of Images on which we obtain the SOTA-level accuracies. We also provide a qualitative visualization of the attention distributions of our model, showing that attention units are higher for object-word pairs referring to the same entity, and that the proposed object-word alignment does not emerge naturally without supervision. Future work will explore the application of this weak supervision signal to other vision-language tasks, including image retrieval.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, 'Bottom-up and top-down attention for image captioning and visual question answering', in *CVPR*, pp. 6077–6086, (2018).

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, 'Vqa: Visual question answering', in *ICCV*, pp. 2425–2433, (2015).

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, 'Layer normalization', *NeurIPS*, (2016).

[4] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori, 'Object level visual reasoning in videos', in *ECCV*, (2018).

[5] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome, 'Mutan: Multimodal tucker fusion for visual question answering', in *ICCV*, pp. 2612–2620, (2017).

[6] Hedi Ben-Younes, Rémi Cadene, Nicolas Thome, and Matthieu Cord, 'Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection', *AAAI*, (2019).

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu, 'Uniter: Learning universal image-text representations', *arXiv 1909.11740*, (2019).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', in *NAACL-HLT*, pp. 4171–4186, (2019).

[9] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li, 'Dynamic fusion with intra-and inter-modality attention flow for visual question answering', in *CVPR*, pp. 6639–6648, (2019).

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, 'Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering', in *CVPR*, (2017).

[11] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko, 'Language-conditioned graph networks for relational reasoning', *ICCV*, (2019).

[12] Drew Hudson and Christopher D Manning, 'Learning by abstraction: The neural state machine', in *NeurIPS*, pp. 5901–5914, (2019).

[13] Drew A Hudson and Christopher D Manning, 'Compositional attention networks for machine reasoning', *ICLR*, (2018).

[14] Drew A. Hudson and Christopher D. Manning, 'Gqa: A new dataset for real-world visual reasoning and compositional question answering', in *CVPR*, (June 2019).

[15] Andrej Karpathy and Li Fei-Fei, 'Deep visual-semantic alignments for generating image descriptions', in *CVPR*, pp. 3128–3137, (2015).

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg, 'Referitgame: Referring to objects in photographs of natural scenes', in *EMNLP*, pp. 787–798, (2014).

[17] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *ICLR*, (2015).

[18] Thomas N Kipf and Max Welling, 'Semi-supervised classification with graph convolutional networks', *ICLR*, (2016).

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., 'Visual genome: Connecting language and vision using crowdsourced dense image annotations', *IJCV*, **123**(1), 32–73, (2017).

[20] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, 'Stacked cross attention for image-text matching', in *ECCV*, pp. 201–216, (2018).

[21] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou, 'Visual question generation as dual task of visual question answering', in *CVPR*, pp. 6116–6124, (2018).

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, 'Microsoft coco: Common objects in context', in *ECCV*, pp. 740–755, (2014).

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, 'Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks', in *NeurIPS*, pp. 13–23, (2019).

[24] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov, 'Generating images from captions with attention', *arXiv 1511.02793*, (2015).

[25] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot, 'Learning conditioned graph structures for interpretable visual question answering', in *NeurIPS*, pp. 8334–8343, (2018).

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning, 'Glove: Global vectors for word representation', in *EMNLP*, pp. 1532–1543, (2014).

[27] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, 'Film: Visual reasoning with a general conditioning layer', in *AAAI*, (2018).

[28] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, 'Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models', in *ICCV*, pp. 2641–2649, (2015).

[29] Mengye Ren, Ryan Kiros, and Richard Zemel, 'Exploring models and data for image question answering', in *NeurIPS*, pp. 2953–2961, (2015).

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, 'Faster r-cnn: Towards real-time object detection with region proposal networks', in *NeurIPS*, pp. 91–99, (2015).

[31] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap, 'A simple neural network module for relational reasoning', in *NeurIPS*, pp. 4967–4976, (2017).

[32] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi, 'A corpus for reasoning about natural language grounded in photographs', in *ACL*, (2019).

[33] Hao Tan and Mohit Bansal, 'Lxmert: Learning cross-modality encoder representations from transformers', in *EMNLP-IJCNLP*, pp. 5103–5114, (2019).

[34] Damien Teney, Lingqiao Liu, and Anton van den Hengel, 'Graph-structured representations for visual question answering', in *CVPR*, pp. 1–9, (2017).

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *NeurIPS*, pp. 5998–6008, (2017).

[36] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al., 'Google's neural machine translation system: Bridging the gap between human and machine translation', *CoRR*, (2016).

[37] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum, 'Neural-symbolic vqa: Disentangling reasoning from vision and language understanding', in *NeurIPS*, pp. 1031–1042, (2018).

[38] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, 'Deep modular co-attention networks for visual question answering', in *CVPR*, pp. 6281–6290, (2019).