Learning Visual Representations with Caption Annotations

Mert Bulent Sariyildiz, Julien Perez, Diane Larlus

NAVER LABS Europe

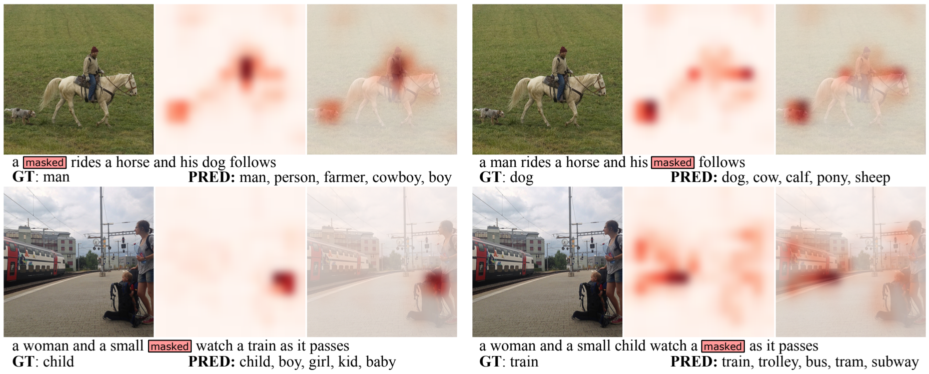

a masked rides a horse and his dog follows
GT: man
PRED: man, person, farmer, cowboy, boy
a man rides a horse and his masked follows
GT: dog
PRED: dog, cow, calf, pony, sheep
a woman and a small masked watch a train as it passes
GT: child
PRED: child, boy, girl, kid, baby
a woman and a small child watch a masked as it passes
GT: train
PRED: train, trolley, bus, tram, subway

Fig. 1: We introduce *image-conditioned masked language modeling* (ICMLM), a proxy task to learn visual representations from scratch given image-caption pairs. This task masks tokens in captions and predicts them by fusing visual and textual cues. The figure shows how the *visual attention* changes as we mask different tokens in a caption (produced by our ICMLM$_{\texttt{tfm}}$ trained on COCO).

Abstract. Pretraining general-purpose visual features has become a crucial part of tackling many computer vision tasks. While one can learn such features on the extensively-annotated ImageNet dataset, recent approaches have looked at ways to allow for noisy, fewer, or even no annotations to perform such pretraining. Starting from the observation that captioned images are easily crawlable, we argue that this overlooked source of information can be exploited to supervise the training of visual representations. To do so, motivated by the recent progresses in language models, we introduce *image-conditioned masked language modeling* (ICMLM) – a proxy task to learn visual representations over image-caption pairs. ICMLM consists in predicting masked words in captions by relying on visual cues. To tackle this task, we propose hybrid models, with dedicated visual and textual encoders, and we show that the visual representations learned as a by-product of solving this task transfer well to a variety of target tasks. Our experiments confirm that image captions can be leveraged to inject global and localized semantic information into visual representations. Project website: `https://europe.naverlabs.com/ICMLM`.

1 Introduction

Large-scale manually annotated datasets [11,62] have been fueling the rapid development of deep learning-based methods in computer vision. Training supervised models over such datasets not only leads to state-of-the-art results, but also enables networks to learn useful image representations that can be exploited on downstream tasks. However, this approach has major limitations. First, the cost and complexity of annotating datasets is considerable, especially when the class taxonomy is fine-grained requiring expert knowledge [11,41,55]. Second, retraining from scratch dedicated models for every new task is inefficient.

Some alternative approaches address these issues and require less curated or fewer annotations [38,59]. At the other extreme of visual representation learning, self-supervised learning methods [6,14,15,17,60,61] do not require annotations and fabricate proxy labels from the data itself. They induce regularities of the data itself, decorrelated from any specific downstream task annotations. Unfortunately, recent findings show that these approaches are not data efficient, *i.e.* they require either extremely large training sets (up to a hundred million images) [6,23] or need to be trained much longer with larger networks to express their full potential [8,25]. Hence they demand huge computational resources.

Interestingly, data often comes with informative metadata for free. For instance, user tags associated with images can be used as image labels [31,38]. Even richer, companion text for images, is sometimes available for free. Using recent sanitation procedures [44], high-quality large-scale captioned datasets can automatically be constructed.

In this paper, we argue that learning visual representations with captions should significantly reduce the scale of the training sets required for pretraining visual representations. If no text is available, in some context it is still easier to acquire short captions than expert-quality-level fine-grained class labels over thousands of categories like in ImageNet [11]. Yet, caption annotations have rarely been used to train visual representations from scratch. Notable exceptions are [19,31,45] which learn image features by training to predict words in a caption or topic probabilities estimated from an associated text. However, none of these approaches use the structure of the entire sentences, *i.e.* they treat words individually. Recent studies [13,43] have shown the superiority of word representations which are conditioned by their surrounding, where the same word has different representations depending on the sentence. We believe such caption representations should also be beneficial for learning image representations.

This paper focuses on the following research questions. *Can we train transferable visual representations from limited sets of image-caption pairs?* If so, *how should we formulate the interaction between images and captions?* To address these questions, we propose several proxy tasks involving images and their captions which we use to train visual representations from scratch. The first one (Sec. 3.1) is intuitive and requires only extracting *image tags* from captions. We propose several ways to do so, and we show that predicting image tags is already competitive compared to other pretraining strategies. Then, to utilize the captions more effectively, and inspired by the recent advances in natural language processing [13], we propose a second proxy task (Sec. 3.2) which employs masked language modeling to learn visual representations. Similar to the first proxy task, it also leverages both images and captions, but it additionally allows visual representations to learn to *localize semantic concepts* in captions. Qualitative results show that the architecture proposed to tackle this second proxy task effectively leverages the text and attends to relevant image regions (see Fig. 1).

Our contributions are threefold. First, we empirically validate that simple tag prediction tasks, where tags are obtained from captions, already learn transferable visual representations. Second, in an attempt to benefit from captions more, we introduce a new task called **i**mage-**c**onditioned **m**asked **l**anguage **m**odeling (ICMLM) and propose two multi-modal architectures to solve this task. Third, we show that solving ICMLM leads to useful visual representations as a by-product. These visual representations, which we obtain using only a hundred thousand captioned images, are competitive with recent self-supervised approaches leveraging a hundred million images, and, in some cases, even fully-supervised approaches showing how powerful a cue text is.

2 Related Work

Pretraining CNNs on an external dataset has become standard practice in computer vision [7,21,46,48], especially for domains or tasks for which data is scarce. The most common strategy is to train a CNN for the ImageNet-1K classification task [47] and then to use it as a feature extractor or to fine-tune it on a target task or domain. Although this scheme has proven to be quite useful, designing fully-annotated datasets represents a significant effort requiring prior knowledge and domain expertise [11]. Thus, alternative research directions have gained interest. We review the ones closest to our work.

Weakly/Webly-supervised learning. Two main research lines have prospered recently. The first line focuses on using *metadata* associated to web data, such as tags or captions for images or videos [53]. Although the signal-to-noise ratio of samples crawled from the web may arguably be lower than carefully-constructed datasets, significant progress has been made leveraging this type of data to pretrain models [9,27]. Among those, to learn visual representations, [31] extracts the most common hashtags and words from the captions and titles of 99 million images in the YFCC100M dataset [53] and train to predict these words using CNNs. Similarly, [38] uses hashtags associated with images from Instagram to construct datasets containing up to 3.5 billion images.

The second line upscales ImageNet. Leveraging ImageNet labels, those approaches produce *pseudo-labels* for additional unlabeled images [58,59]. We note that these methods require initial annotations and extremely large-scale sets of images. In contrary, our models need far less images, 118 thousand images at most, but companion captions to learn visual representations.

Unsupervised representation learning. Self-supervised approaches build a *pretext task* to learn image representations which are decorrelated from any downstream task and they do not require any manual annotations. Often, *proxy* *tasks* consist in predicting missing pieces in manipulated images, for instance context prediction [14], colorization [12,34,60], inpainting of missing portions [42], prediction of image rotations [18], spotting artifacts [30], or cross-channel prediction [61]. Besides, recently, contrastive learning-based unsupervised methods [2,25,40,57] have showed significant improvements. However, computational and data efficiency of these methods are still inferior to supervised models.

It is important to note that most unsupervised approaches are trained on *curated datasets* such as ImageNet for which images were carefully selected to form a well-balanced collection for a diverse set of fine-grained categories. Although these approaches do not directly use ImageNet labels, they implicitly benefit from this careful selection and the resulting underlying structure of the dataset. Indeed, [5,14] show that the feature quality drops when raw data are used instead of ImageNet. Yet, assuming that a curated dataset such as ImageNet is readily available is a strong assumption. Consequently, some works [6,23,39] have evaluated unsupervised methods trained on *uncurated data* [53]. They have concluded that large amounts of raw data (*e.g.* 96 millions images) is required to express the full potential of these approaches. In this work, we focus on learning from a much smaller set of images by leveraging textual information.

Vision and language. Vision and language (VL) have been jointly leveraged to learn cross-modal representations for various VL tasks, such as cross-modal retrieval [20,56], visual question answering [24], captioning [51] or visual grounding [10,29]. Building on the recent advances in natural language processing [13,54], several works have fine-tuned BERT [13] to fuse visual and textual information [37,50,51,52,64] for VL tasks. However, while learning cross-modal representations, such approaches rely on pretrained feature extractors, *i.e.* they use visual features pooled from regions of interest produced by a state-of-the-art detector such as Faster-RCNN [46]. Therefore, their objectives are formulated under the assumption that discriminative visual features are readily available for a list of relevant objects. We note that such feature extractors are already state-of-the-art for most vision tasks, requiring *expensive bounding box annotations* to train. Our approach follows a different path. We focus on learning visual representations *from scratch* for purely visual tasks by leveraging captions.

Learning visual features using text. Only few works have taken advantage of the companion text to learn image representations. [45] creates and solves auxiliary prediction tasks from images with associated captions. [35] constructs label sets out of caption n-grams, and trains CNNs by predicting these labels. [19] extracts topic models for Wikipedia pages using latent Dirichlet allocation and trains a CNN to embed their associated images in this topic space. [22] uses captions to learn image representations for the specific task of semantic retrieval.

We argue that language has a complex structure which cannot be reduced to computing n-grams statistics in a text. Motivated by this, we differ and propose to use a pretrained language model - which can be trained in an unsupervised manner for large text corpora - to represent captions and individual words in them. In our experiments, we show that by doing so it is possible to learn visual representations that are useful for a broad range of tasks.

3 Method

We argue that captions associated with images can provide semantic information about some *observable concepts* that can be captured by image representations. Such concepts can be objects, attributes, or actions that visually appear in images. With this motivation, given a dataset composed of image-caption pairs, we want to formulate non-trivial proxy tasks conditioned on both images and captions such that solving these tasks produce generic visual representations as a by-product. In particular, we want such tasks to properly use the structure of caption sentences, and not only treat them as orderless sets of words.

To this end, we propose two proxy tasks focusing on two distinct objectives to train CNNs to recognize a predefined set of concepts in images. The first proxy task captures *global* semantics in images by predicting image-level tags and is presented in Sec. 3.1. The second proxy task, the image-conditioned masked language modeling task, focuses on *local* semantics in images and is detailed in Sec. 3.2. Experiments show that both proxy tasks are complementary.

Notations. We assume that our dataset $\mathcal{D} = \{(I^i, c^i)\}_1^N$ is composed of N image-caption pairs. We denote by $O = \{o_i\}_1^K$ the set of concepts to be recognized in images. As there can be multiple concepts in an image, we use binary label vectors $\mathbf{y} \in \{0,1\}^K$ to denote the presence of concepts in images, *i.e.* $\mathbf{y}_k = 1$ if concept o_k appears in image I and 0 otherwise. We define two parametric functions ϕ and ψ which respectively embed images and text. More precisely, $\phi : I \rightarrow \mathbf{X} \in \mathbb{R}^{H \times W \times d_\mathrm{x}}$ takes an image I as input and produces $\mathbf{X}$ which is composed of d_x-dimensional visual features over a spatial grid of size $H \cdot W$. Similarly, $\psi : c \rightarrow \mathbf{W} \in \mathbb{R}^{T \times d_\mathrm{w}}$ transforms a caption (a sequence of T tokens) into a set of d_w-dimensional vectors, one for each token. In our models, we train only ϕ, which is a CNN producing visual representations, and we use a pretrained language model as ψ that we freeze during training.

3.1 Capturing image-level semantics

A straightforward way to build a proxy task given image-caption pairs is to formulate a multi-label image classification problem, where, according to its caption, multiple concepts may appear in an image [31,45]. For this setup, we create a label vector $\mathbf{y} \in \{0,1\}^K$ for each image I such that $\mathbf{y}_j = 1$ if concept o_j appears in the image, and 0 otherwise. We denote these labels as *tags*, and name this task as *tag prediction* (TP), illustrated in Fig. 2 (modules **(1)** + **(5)**).

One of the contributions of this work is to consider different ways to define concept sets O from captions. Ground-truth concept vectors can be easily obtained by considering the most frequent bi-grams [31] or even n-grams [35] in captions. More sophisticated ways to obtain artificial labels include using LDA [4] to discover latent topics in captions [19]. In addition to these existing methods, we look for ways to exploit semantics of tokens in captions.

TP$_{\mathbf{Postag}}$. As a first approach, we simply propose to construct label sets by taking into account the *part-of-speech* (POS) tags of tokens in captions. Concretely, we use the off-the-shelf language parser [28] to determine POS tags of tokens

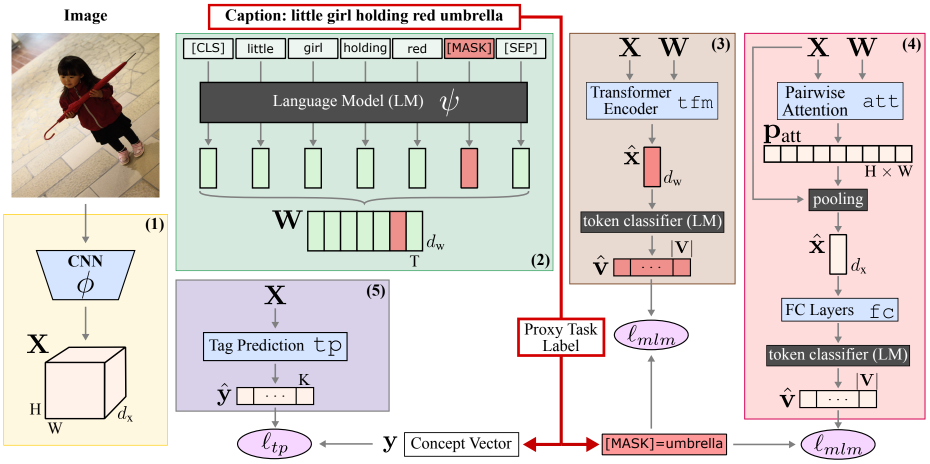


Fig. 2: **Modules used in our models. (1)** a CNN to extract visual features; **(2)** a language model to extract token features; **(3)**, **(4)** and **(5)** respectively correspond to our `tfm`, `att` + `fc` and `tp` modules. Our TP$_\star$, ICMLM$_{\texttt{tfm}}$ and ICMLM$_{\texttt{att-fc}}$ models combine these modules: **(1)** + **(5)**, **(1)** + **(2)** + **(3)** and **(1)** + **(2)** + **(4)**, respectively. Trainable (and frozen) components are colored in blue (and black). Only the CNN is used during target task evaluations.

in captions and gather three label sets of size K, including (i) only nouns, (ii) nouns and adjectives, (iii) nouns, adjectives and verbs. These three label sets are used to train three separate TP$_{\text{Postag}}$ models.

TP$_{\text{Cluster}}$. As mentioned above, we believe it would be beneficial to use the structure of the full caption and not just treat it as an orderless set of tokens as the previously proposed TP$_{\text{Postag}}$. To this end, we use the pretrained BERT$_{\text{base}}$ [13] model to extract sentence-level caption representations. We do this by feeding the caption into BERT$_{\text{base}}$ and taking the representation for the [CLS] token, which is used as a special token to encode sentence-level text representations in BERT$_{\text{base}}$. Then, we cluster the sentence-level representations of all captions using the k-means algorithm and apply hard cluster assignment. This way, the labels are the cluster indices and we train ϕ by learning to predict the cluster assignments of captions from their associated image. K-means learns K cluster centroids $\xi^\star \in \mathbb{R}^{d_w \times K}$ in the caption representation space by minimizing:

$$\xi^\star, \{\mathbf{y}^{i\star}\}_{i=1}^{N} = \operatorname*{arg\,min}_{\substack{\xi \in \mathbb{R}^{d_w \times K}, \\ \{\mathbf{y}^i \in \{0,1\}^K,\ \mathbf{1}_K^\top \mathbf{y}^i = 1\}_{i=1}^{N}}} \sum_{i=1}^{N} \|\psi(c^i)_{[\text{CLS}]} - \xi \mathbf{y}^i\|_2^2, \tag{1}$$

where $\psi(c)_{[\text{CLS}]}$ and $\mathbf{y}^\star$ denote the [CLS] representation of the caption c and of the one-hot cluster assignment vector obtained for c. Note that $\mathbf{y}^\star$ is used as the label for image I. In case there are multiple captions for an image, we simply aggregate the cluster labels of all captions associated to that image.

Training TP$_\star$ models. Once we have crafted image labels over a chosen set of concepts (either using POS tags or cluster assignments), following [38], we normalize the binary label vectors to sum up to one, *i.e.* $\mathbf{y}^\top \mathbf{1}_K = 1$, for all samples. Then we train models by minimizing the categorical cross-entropy:

$$\ell_{\text{tp}} = - \mathop{\mathbb{E}}_{(I,c)\in\mathcal{D}} \left[\sum_{k=1}^{K} \mathbf{y}_k \log(p(\hat{\mathbf{y}}_k \mid I)) \right], \tag{2}$$

where $p(\hat{\mathbf{y}}_k \mid I) = \frac{\exp(\hat{\mathbf{y}}_k)}{\sum_j \exp(\hat{\mathbf{y}}_j)}$, $\hat{\mathbf{y}}_k = \mathtt{tp}(\phi(I))_k$, and $\mathtt{tp} : \mathbb{R}^{H\times W\times d_{\text{x}}} \rightarrow \mathbb{R}^K$ is a parametric function performing tag predictions.

3.2 Capturing localized semantics

The previous section presents a cluster prediction task where the structure of the sentence is leveraged through the use of the [CLS] output of the pretrained BERT$_{\text{base}}$. Yet, this has a major limitation: token-level details may largely be ignored especially when captions are long [3]. Our experiments also support this argument, *i.e.* TP$_{\text{Cluster}}$ performs on par with or worse than TP$_{\text{Postag}}$. To address this issue, we propose a second learning protocol that learns to explicitly *relate* individual concepts appearing in both an image and its caption.

To this end, we extend the natural language processing task known as Masked Language Model (MLM) [13] into an *image-conditioned* version. The MLM task trains a language model by masking a subset of the tokens in an input sentence, and then by predicting these masked tokens. Inspired by this idea, we introduce the Image-Conditioned Masked Language Model (ICMLM) task. Compared to MLM, we propose to predict masked tokens in a caption by using the visual information computed by ϕ. This way, we learn visual representations that should be informative enough to reconstruct the missing information in captions.

For this task, for each image-caption pair (I, c), we assume that there is at least one concept appearing in the caption c. Since c describes the visual scene in I, we assume that concepts appearing in c are observable in I as well. This allows us to define ICMLM as a concept set recognition problem in images. More precisely, we use the pretrained BERT$_{\text{base}}$ model [13] as the textual embedding function ψ and define the learning protocol as follows. First, we segment the caption c into a sequence of tokens $(t_1, \ldots, t_T)$, and mask one of the tokens t_m, which belongs to the concept set. Masking is simply done by replacing the token t_m with a special token reserved for this operation, for instance BERT$_{\text{base}}$ [13] uses "[MASK]". Then, *contextualized* representations of the tokens are computed as $\mathbf{W} = \psi((t_1, \ldots, t_T))$. Meanwhile, the visual representation of the image I is computed by $\phi(I) = \mathbf{X}$. Since our goal is to predict the masked token by using both visual and textual representations, we need to merge them. A naive way to accomplish that is to (i) pool the representations of each modality into a global vector, (ii) aggregate (*i.e.* concatenate) these vectors, (iii) use the resulting vector to predict the label of the masked token. However, the representations obtained in this way could only focus on the global semantics, and the local information for both modalities might be lost during the pooling stage. To address this concern, we describe two possible designs for ICMLM relying on individual visual (in the spatial grid) and textual (in the sequence) features.

ICMLM$_{\texttt{tfm}}$. Here, we contextualize token representations among visual ones by fusing them in a data-driven manner (similar to [37]). Concretely, we spatially flatten and project $\mathbf{X}$ to the token embedding space, concatenate it with $\mathbf{W}$ and apply a transformer encoder module [54], `tfm`, on top of the stacked representations. Finally, as done in BERT$_{\text{base}}$ [13], the label of the masked token t_m can be predicted by feeding the representation of the *transformed* masked token into the pretrained token classification layer of BERT$_{\text{base}}$. We call this ICMLM flavor ICMLM$_{\texttt{tfm}}$(modules **(1)** + **(2)** + **(3)** in Fig. 2).

ICMLM$_{\texttt{att-fc}}$. Transformer networks employ a self-attention mechanism with respect to their inputs. Therefore they can learn the pairwise relationships of both the visual and the textual representations. This allows them, for instance, to fuse different domains quite effectively [37,51]. We also verify this powerful aspect of the transformers in our experiments, *e.g.* even a single-layered transformer network is enough to perform significantly well at predicting masked tokens on the MS-COCO dataset [36]. However, the rest of the caption is already a powerful cue to predict the masked token and this transformer-based architecture might rely too much on the text, potentially leading to weaker visual representations. As an alternative, we propose to predict the label of the masked token by using the visual features alone. Since the masked token is a concept that we want to recognize in the image, we divide the prediction problem into two sub-problems: localizing the concept in the image and predicting its label. To do that we define two additional trainable modules: `att` and `fc` modules that we describe in detail below. This ICMLM flavor is referred to as ICMLM$_{\texttt{att-fc}}$ (modules **(1)** + **(2)** + **(4)** in Fig. 2).

The goal of the `att` module is to create a 2D attention map on the spatial grid of the visual feature tensor $\mathbf{X}$ such that high energy values correspond to the location of the concept masked in the caption c. It takes as input the spatially-flattened visual features $\mathbf{X} \in \mathbb{R}^{H \cdot W \times d_{\text{x}}}$ and the textual features $\mathbf{W}$. First, $\mathbf{X}$ and $\mathbf{W}$ are mapped to a common d_{z}-dimensional space and then pairwise attention scores between visual and textual vectors are computed:

$$\tilde{\mathbf{X}} = \lfloor \text{norm}(\mathbf{X}\Sigma_x) \rfloor_+, \qquad \tilde{\mathbf{W}} = \lfloor \text{norm}(\mathbf{W}\Sigma_w) \rfloor_+, \qquad \mathbf{S} = \frac{\tilde{\mathbf{X}}\tilde{\mathbf{W}}^\top}{\sqrt{d_{\text{z}}}}, \tag{3}$$

where $\Sigma_x \in \mathbb{R}^{d_{\text{x}} \times d_{\text{z}}}$ and $\Sigma_w \in \mathbb{R}^{d_{\text{w}} \times d_{\text{z}}}$ are parameters to learn, norm is LayerNorm [1] and $\lfloor . \rfloor_+$ is ReLU operator. Note that $\mathbf{S}_{i,j}$ denotes the attention of visual vector i (a particular location in the flattened spatial-grid of the image) to textual vector j (a particular token in the caption). To be able to suppress attention scores of vague tokens such as “about” or “through”, we compute *soft maximum* of the textual attentions for each visual feature:

$$\mathbf{s}_i = \log \sum_{j=1}^{T} \exp \left(\mathbf{S}_{i,j} \right). \tag{4}$$

We note that operations in Eqs. (3) and (4) are performed for a single attention head and the multi-headed attention mechanism [54] can easily be adopted by learning a weighted averaging layer: $\mathbf{s}_i = \left[\mathbf{s}_i^1 | \cdots | \mathbf{s}_i^H\right] \Sigma_h + b_h$, where $\Sigma_h \in \mathbb{R}^H$ and $b_h \in \mathbb{R}$ are the parameters of the averaging layer, $\mathbf{s}_i^h$ is the aggregated textual attention score for the i^{th} visual feature coming from the h^{th} attention head, and $[.|.]$ denotes concatenation. Finally, attention probabilities are obtained by applying softmax, and used to pool $\mathbf{X}$ into a single visual feature $\hat{\mathbf{x}}$:

$$\mathbf{p}_{\mathtt{att}i} = \frac{\exp(\mathbf{s}_i)}{\sum_{j=1}^{H \cdot W} \exp(\mathbf{s}_j)}, \qquad \hat{\mathbf{x}} = \mathbf{X}^\top \mathbf{p}_{\mathtt{att}}, \tag{5}$$

where $\mathbf{p}_{\mathtt{att}} \in [0,1]^{H \cdot W}$ such that $\mathbf{p}_{\mathtt{att}}{}^\top \mathbf{1}_{H \cdot W} = 1$.

After localizing the concept of interest in image I by means of pooling $\mathbf{X}$ into $\hat{\mathbf{x}}$, we feed $\hat{\mathbf{x}}$ into the `fc` module, which consists in a sequence of fully-connected layers, each composed of linear transformation, LayerNorm and ReLU operator. Finally, we map the output of the `fc` module to the $\text{BERT}_{\text{base}}$'s token vocabulary $\mathbf{V}$ and compute prediction probabilities as follows:

$$p_{\mathbf{V}}\left(k | I, c, t_m\right) = \frac{\exp(\hat{\mathbf{v}}_k)}{\sum_j \exp(\hat{\mathbf{v}}_j)}, \tag{6}$$

where $\hat{\mathbf{v}}_k = \mathtt{fc}(\hat{\mathbf{x}})^\top \mathbf{V}_k$ and $\mathbf{V}_k \in d_{\text{w}}$ are the prediction score and the pretrained distributed representation of the k^{th} token in the pretrained candidate lexicon of $\text{BERT}_{\text{base}}$. As we compute dot-products between post-processed $\hat{\mathbf{x}}$ and the pretrained representations of the tokens in $\text{BERT}_{\text{base}}$'s vocabulary, it is possible to leverage the structure in $\text{BERT}_{\text{base}}$'s hidden representation space. Indeed, we observe that such probability estimation of a candidate token is more effective than learning a fully connected layer which projects $\mathtt{fc}(\hat{\mathbf{x}})$ onto the vocabulary.

Training $\text{ICMLM}_\star$ models. To train both model flavors, for each masked token t_m in all $(I,\, c)$ pairs in $\mathcal{D}$, we minimize the cross-entropy loss between the probability distribution over the $\text{BERT}_{\text{base}}$'s vocabulary as computed in Eq. (6) and the label of the masked token t_m (index of t_m in $\mathbf{V}$):

$$\ell_{\text{mlm}} = - \mathop{\mathbb{E}}_{(I,c) \in \mathcal{D}} \left[\mathop{\mathbb{E}}_{t_m \in c} \left[\log(p_{\mathbf{V}}(k | I, c, t_m)) \right] \right], \tag{7}$$

where k is the index of t_m in $\text{BERT}_{\text{base}}$'s vocabulary. The expectation over captions implies that there can be multiple concepts in a caption and we can mask and predict each of them separately. For $\text{ICMLM}_{\mathtt{tfm}}$, $\hat{\mathbf{x}}$ is computed by the `tfm` module, and it corresponds to the representation of the masked token. For $\text{ICMLM}_{\mathtt{att\text{-}fc}}$, $\hat{\mathbf{x}}$ corresponds to the output from the `fc` module.

We also note that ℓ_{tp} and ℓ_{mlm} are complementary, enforcing ϕ to focus on global and local semantics in images, respectively. Therefore, in both $\text{ICMLM}_{\mathtt{att\text{-}fc}}$ and $\text{ICMLM}_{\mathtt{tfm}}$ we minimize the weighted combination of ℓ_{tp} and ℓ_{mlm}:

$$\ell_{\text{icmlm}} = \ell_{\text{mlm}} + \lambda \ell_{\text{tp}}. \tag{8}$$

Table 1: **Proxy *vs.* target task performances.** We report top-1 and top-5 masked token prediction scores (as proxy, on VG and COCO) and mAP scores obtained using features from various layers (as target, on VOC-07), on validation sets. T-1/5: top-1/5 scores, C-$\star$: conv. layer from which features are extracted.

	Proxy			Target			Proxy			Target		
Method	Dataset	T-1	T-5	C-11	C-12	C-13	Dataset	T-1	T-5	C-11	C-12	C-13
$\text{BERT}_{\text{base}}$	VG	17.4	36.9	–	–	–	COCO	25.7	40.3	–	–	–
$\text{ICMLM}_{\texttt{tfm}}$	VG	**49.7**	**79.2**	71.3	75.8	80.5	COCO	**70.3**	**91.5**	70.2	74.2	77.5
$\text{ICMLM}_{\texttt{att-fc}}$	VG	41.1	71.3	**73.7**	**78.7**	**83.1**	COCO	59.4	83.4	**72.3**	**77.5**	**82.2**

4 Experiments

This section evaluates **(i)** how the performance on the masked language modeling (MLM) proxy task translates to target tasks (Sec. 4.1), **(ii)** how several types of supervision associated to a set of images (*i.e.* full, weak and self-supervision) compare to each other (Sec. 4.2), **(iii)** if the gains of ICMLM$_\star$ models are consistent across backbone architectures (Sec. 4.3), **(iv)** if ICMLM$_\star$ models attend to relevant regions in images (Figs 1 and 3). First, we introduce our experimental setup (remaining details are in the supplementary material).

Datasets. We train our models on the image-caption pairs of either the 2017 split of MS-COCO [36] (COCO) or the Visual Genome [33] (VG) datasets. COCO has 123K images (118K and 5K for train and val) and 5 captions for each image while VG has 108K images (we randomly split 103K and 5K for train and val) and 5.4M captions. We remove duplicate captions and those with more than 25 or less than 3 tokens. We construct several concept sets using the captions of COCO or VG, to be used as tags for $\text{TP}_{\text{Postag}}$ and as maskable tokens for ICMLM$_\star$ models (an ablative study is provided in the supplementary material). Note that depending on the concept set, the number of tags and the (image, caption, maskable token) triplets vary, therefore, we specify which concept set is used in all $\text{TP}_{\text{Postag}}$ and ICMLM$_\star$ experiments.

Networks. To be comparable with the state-of-the-art self-supervised learning method DeeperCluster [6], we mainly use VGG16 [49] backbones. We also evaluate ICMLM$_\star$ models using ResNet50 [26] in Sec. 4.3. Note that ICMLM$_\star$ models operate on a set of visual tensors, therefore, for TP$_\star$ and ICMLM$_\star$ models we remove the FC layers from VGG16. To compensate, we use 4-layered CNNs combined with global average pooling and linear layer for tag predictions as `tp` modules. For `tfm`, `att` and `fc` modules, we cross-validated the number of hidden layers and attention heads on the validation set of Pascal VOC-07 dataset, and found that 1 hidden layer (in `tfm` and `fc`) and 12 attention heads (in `tfm` and `att`) works well. While training ICMLM$_\star$ models we set $\lambda = 1$ in Eq. (8).

Target task. Once a model is trained, we discard its additional modules used during training (*i.e.* all but ϕ) and evaluate ϕ on image classification tasks, to test how well pretrained representations generalize to new tasks. To do that, following [6], we train linear logistic regression classifiers attached to the last three convolutional layers of the frozen backbones ϕ with SGD updates and data augmentation. We perform these analyses on the Pascal-VOC07 dataset [16] (VOC) for multi-label classification, and ImageNet-1K (IN-1K) [11] and Places-205 [63] datasets for large-scale categorization, using the publicly available code of [6] with slight modifications: We apply heavier data augmentations [8] and train the classifiers for more iterations, which we found useful in our evaluations.

Additional TP$_\star$ models. We note that the TP model defined in Sec. 3.1 can be used for predicting any type of image tags, with slight modifications. We use it to predict topics as proposed in [19] and denote this approach as TP$_{\text{LDA}}$. To do so, we only modify Eq. (2) to minimize binary cross-entropy loss instead, where K denotes the number of hidden topics. Similarly, we denote TP$_{\text{Label}}$ as the supervised approach which uses the annotated image labels as tags.

4.1 Ablative study on the proxy task

We first study the interplay between ICMLM and target tasks. To do so, we train several ICMLM$_\star$ models, and monitor their performance on both the proxy and target tasks, *i.e.* we report masked token prediction (MTP) scores on VG and COCO, and mAP scores on VOC, respectively. For reference, we also report MTP scores obtained by a single BERT$_{\text{base}}$ model, where masked tokens are predicted using only the remainder of the captions. In this study, we used the 1K most frequent nouns and adjectives in the captions as maskable tokens.

Results are shown in Tab. 1. We observe that ICMLM$_\star$ models significantly improve MTP scores compared to BERT$_{\text{base}}$ model, showing that visual cues are useful for MLM tasks. Moreover, ICMLM$_{\texttt{tfm}}$ is better than ICMLM$_{\texttt{att-fc}}$ on the proxy task, indicating that blending visual and textual cues, which is effectively done by the `tfm` module, is beneficial for MLM. However, ICMLM$_{\texttt{att-fc}}$ generalizes better than ICMLM$_{\texttt{tfm}}$ to VOC. We believe that, as ICMLM$_{\texttt{att-fc}}$ predicts masked tokens using visual cues only, it learns semantic concepts from the given training set better than ICMLM$_{\texttt{tfm}}$. A similar study which uses ResNet50 backbones [26] leads to similar observations (see the supplementary material).

4.2 Comparison of fully-, weakly- and self-supervised methods

Next, we compare the visual representations learned by different state-of-the-art fully-, weakly- and self-supervised learning (SSL) models. We do this by training the models explained below on COCO or VG, then using their backbones ϕ to perform the target tasks, *i.e.* image classification on VOC, IN-1K and Places-205.

Supervised. For reference, we report the results obtained by three supervised classifiers trained on different subsets of IN-1K: **(i)** "ImageNet" on the full IN-1K, **(ii)** "$\mathcal{S}$-ImageNet with 1K classes" on randomly-sampled 100 images per class, **(iii)** "$\mathcal{S}$-ImageNet with 100 classes" on 1K images for each of 100 randomly sampled classes. The latter two contain 100K images each *i.e.* the same order of magnitude as COCO or VG. For the models trained on these three subsets, we repeat the sampling 4 times and report their mean target task results. We also report TP$_{\text{Label}}$ which is trained to predict ground-truth labels.

Table 2: **Fully-, weakly- and self-supervised methods** trained with VGG16 backbones. We report mAP on VOC and top-1 on IN-1K and Places. For VOC, we report the mean of 5 runs (std. ≤ 0.2). We use pretrained models for ImageNet and DeeperCluster, and train other models from scratch. #I: number of images in training sets. C-$\star$: Conv. layer from which features are extracted. Red and orange numbers denote the first and second best numbers in columns. Blue numbers are not transfer tasks (*i.e.* they use the same dataset for proxy/target).

	Proxy tasks			*Target tasks*								
				VOC			**IN-1K**			**Places**		
Method	**Dataset**	**Supervision**	**# I**	C-11	C-12	C-13	C-11	C-12	C-13	C-11	C-12	C-13
ImageNet	IN-1K$_{\text{full}}$	Labels 1K classes	1.3M	77.5	81.0	84.7	59.8	65.7	71.8	43.0	43.5	47.3
$\mathcal{S}$-ImageNet	IN-1K$_{\text{sub}}$	Labels 1K classes	100K	69.3	72.4	74.1	50.5	52.5	53.8	40.9	41.6	41.1
$\mathcal{S}$-ImageNet	IN-1K$_{\text{sub}}$	Labels 100 classes	100K	67.4	69.6	70.5	47.4	48.4	46.3	39.3	39.3	35.8
TP$_{\text{Label}}$	COCO	Labels 80 classes	118K	72.4	76.3	79.9	50.4	50.6	49.9	44.5	45.0	44.5
DeeperCluster [6]	YFCC	Self -	96M	71.4	73.3	73.1	48.0	48.8	45.1	43.1	44.1	41.0
RotNet [18]	COCO	Self -	118K	60.3	61.1	58.6	41.8	40.1	33.3	39.5	38.4	34.7
RotNet [18]	VG	Self -	103K	59.9	60.9	59.2	39.5	38.4	34.7	39.7	38.9	34.9
TP$_{\text{LDA}}$ [19]	COCO	Text 40 topics	118K	70.6	73.9	76.3	48.7	48.4	46.7	43.7	44.1	43.0
TP$_{\text{Cluster}}$ (*Ours*)	COCO	Text 1K clusters	118K	71.5	74.5	77.0	49.5	49.8	48.1	44.1	44.6	43.7
TP$_{\text{Cluster}}$ (*Ours*)	COCO	Text 10K clusters	118K	72.1	75.0	77.2	50.2	50.3	48.7	45.1	45.3	44.2
TP$_{\text{Postag}}$ (*Ours*)	COCO	Text 1K tokens	118K	73.3	76.4	79.3	50.6	51.1	50.0	45.9	46.5	45.8
TP$_{\text{Postag}}$ (*Ours*)	COCO	Text 10K tokens	118K	73.6	77.0	79.4	51.2	51.7	50.5	46.1	47.0	46.1
ICMLM$_{\text{tfm}}$ (*Ours*)	COCO	Text sentences	118K	74.8	77.8	80.5	52.0	52.0	50.8	46.8	47.3	46.2
ICMLM$_{\text{att-fc}}$ (*Ours*)	COCO	Text sentences	118K	75.4	79.1	82.5	52.2	52.2	49.4	46.4	47.0	44.6
TP$_{\text{LDA}}$ [19]	VG	Text 40 topics	103K	71.5	74.6	77.7	49.3	49.2	47.8	44.4	44.9	44.0
TP$_{\text{Cluster}}$ (*Ours*)	VG	Text 1K clusters	103K	73.0	76.2	79.4	50.0	49.8	47.3	45.4	45.8	44.5
TP$_{\text{Cluster}}$ (*Ours*)	VG	Text 10K clusters	103K	73.9	77.8	81.3	50.8	50.7	48.5	46.2	46.9	45.6
TP$_{\text{Postag}}$ (*Ours*)	VG	Text 1K tokens	103K	72.9	76.4	79.6	49.9	49.8	49.1	46.0	46.5	46.4
TP$_{\text{Postag}}$ (*Ours*)	VG	Text 10K tokens	103K	73.5	76.9	80.1	50.9	51.3	50.0	46.1	46.7	46.7
ICMLM$_{\text{tfm}}$ (*Ours*)	VG	Text sentences	103K	75.5	79.3	82.6	52.4	52.2	51.1	47.3	47.8	47.5
ICMLM$_{\text{att-fc}}$ (*Ours*)	VG	Text sentences	103K	76.9	81.2	85.0	52.2	52.2	47.8	47.4	47.9	47.7

Weakly-supervised. We compare TP$_{\text{LDA}}$, TP$_{\text{Cluster}}$, TP$_{\text{Postag}}$ and ICMLM$_{\star}$ methods, for which image-level tags are extracted from the captions of COCO or VG. For TP$_{\text{LDA}}$ we use the publicly-available code of [19] to find 40 latent topics among all captions (the number of topics was validated on the validation set of VOC). Then, probabilities over caption topics define the tag labels for each image. For TP$_{\text{Cluster}}$, we cluster the captions (finding 1K or 10K clusters) and assign the cluster IDs of the captions associated to images as their tag labels. For TP$_{\text{Postag}}$, the tag labels are the most frequent 1K or 10K nouns, adjectives and verbs in the captions. For ICMLM$_{\star}$ models the maskable tokens are the most frequent 1K nouns, adjectives and verbs in the captions.

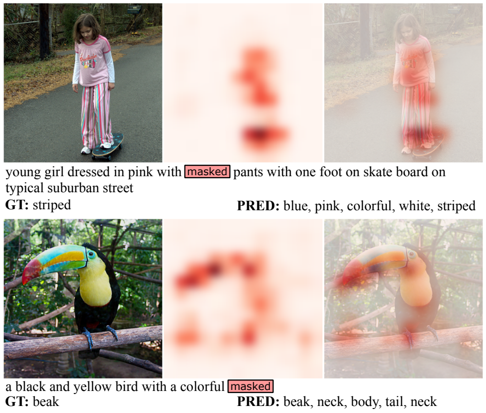
Self-supervised. For reference, we also provide results for two self-supervised approaches: RotNet [18] and DeeperCluster [6]. We train RotNet models from scratch on COCO or VG. For DeeperCluster, we use a model pretrained on the large-scale YFCC-100M dataset [53] (96M images).

Results are reported in Tab. 2. We observe the following. **(i)** We see that the good results of "ImageNet" are mostly due to its scale. Reducing it to 100K

Table 3: **Fully- and weakly-supervised methods** trained with ResNet50 backbones. We use the pretrained ImageNet model and train other models from scratch. We report mAP and top-1 obtained by linear SVMs (on VOC) and logistic regression classifiers (on IN-1K) using pre-extracted features (avg. of 5 runs, std. $\leq$ 0.2). Blue numbers are not transfer tasks.

Model	**Dataset**	**Sup.**	**VOC**	**IN-1K**
ImageNet	IN-1K	Labels	**87.9**	74.7
TP$_{\text{Label}}$	COCO	Labels	80.2	34.0
TP$_{\text{Postag}}$	COCO	Text	82.6	43.9
ICMLM$_{\texttt{tfm}}$	COCO	Text	87.3	**51.9**
ICMLM$_{\texttt{att-fc}}$	COCO	Text	87.5	47.9

Fig. 3: **Attention maps** for masked tokens produced by ICMLM$_{\texttt{tfm}}$ model with ResNet50 backbone trained on COCO (darker red means stronger attention).



young girl dressed in pink with masked pants with one foot on skate board on typical suburban street

GT: striped **PRED:** blue, pink, colorful, white, striped

a black and yellow bird with a colorful masked

GT: beak **PRED:** beak, neck, body, tail, neck

images, either by reducing the number of classes or the number of images per class significantly hurt the performance. Similarly, the supervised TP$_{\text{Label}}$, which uses an order of magnitude fewer categories and images performs far worse than ImageNet. **(ii)** The proposed TP$_{\text{Cluster}}$ outperforms the current state of the art for training with captions, TP$_{\text{LDA}}$ [19], for all three datasets. Exploiting both the structure and the semantics of captions with the BERT$_{\text{base}}$ language model, it improves over a topic model. However, we see that TP$_{\text{Cluster}}$ performs on par with or worse than TP$_{\text{Postag}}$, suggesting that the importance of individual tokens might be suppressed in global caption representations. This validates our motivation for proposing ICMLM in Sec. 3.2: models should leverage both global and local semantics in captions. **(iii)** We see that both ICMLM$_{\texttt{tfm}}$ and ICMLM$_{\texttt{att-fc}}$ improve over all TP$_\star$ baselines by significant margins. Moreover, on VOC evaluations ICMLM$_{\texttt{att-fc}}$ outperforms ICMLM$_{\texttt{tfm}}$ while on IN-1K and Places it performs on par with or worse than ICMLM$_{\texttt{tfm}}$. Note that we observe a similar outcome with ResNet50 backbones (Sec. 4.3). **(iv)** Surprisingly, for VOC and Places-205, at least one ICMLM flavor outperforms the full ImageNet pretrained model which we believe is a significant achievement. For IN-1K, such comparison does not make sense as, in this setting, the proxy and the target datasets are the same. Training on the target set clearly confers an unfair advantage w.r.t. other approaches.

4.3 Additional results with ResNet50

Some self-supervised proxy tasks might favor certain network architectures (*e.g.* see [32]). This section provides additional results where ICMLM$_\star$ models use ResNet50 [26] backbone architectures. To this end, we train TP$_{\text{Label}}$, TP$_{\text{Postag}}$ and ICMLM$_\star$ models on COCO and perform image classification on VOC and

IN-1K. To reduce computational costs, following [23], we train linear SVMs (on VOC) and logistic regression classifiers (on IN-1K) using image features pre-extracted from frozen backbones. Note that ResNet50 is a fully-convolutional network being more expressive compared to VGG16 (thanks to its residual connections and higher number of parameters). Consequently, in this analysis, we use a 2-layered MLPs as `tp` module, a single attention head, and $\lambda = 0.1$ in Eq. (8). We also move to a bigger concept set for $\text{TP}_{\text{Postag}}$ and $\text{ICMLM}_{\star}$ models, *i.e.* the 5K most frequent nouns, adjectives and verbs.

Results are shown in Tab. 3. We observe larger improvements of $\text{TP}_{\text{Postag}}$ over TP_{Label} and of $\text{ICMLM}_{\star}$ over $\text{TP}_{\text{Postag}}$. $\text{ICMLM}_{\star}$ outperforms $\text{TP}_{\text{Postag}}$ by at least **4.7**%, **4.0**% and TP_{Label} by at least **7.1**%, **13.9**% on VOC and IN-1K. These results indicate that more complex CNNs are better at suppressing noise in weak labels and at learning cross-modal representations. Besides, similar to our previous analyses, we see that $\text{ICMLM}_{\texttt{att-fc}}$ learns semantic concepts from the training set slightly better (see the VOC results). However, $\text{ICMLM}_{\texttt{tfm}}$ performs better on IN-1K, suggesting that the ResNet50 backbone learns more discriminative features when guided by the same language model.

Qualitative results. Our goal in ICMLM is to perform MLM task by *looking at* images. To see if they can attend to relevant parts in images, we visualize *attention maps* corresponding to the attention weights of visual features to masked tokens. Figs 1 and 3 present such visualizations produced by our $\text{ICMLM}_{\texttt{tfm}}$ model with ResNet50 backbone trained on COCO. We see that not only the model is able to detect possible concepts of interest, it can also understand which concept is asked in the captions (see the supplementary for more visualizations).

5 Conclusion

Until recently, carefully collected and manually annotated image sets have provided the most efficient way of learning general purpose visual representations. To address the annotation cost, weakly-, webly-, and self-supervised learning approaches have traded quality – a clean supervisory signal – with quantity, requiring up to hundreds of million images. Although, in some cases, large quantities of unlabeled data are readily available, processing such large volumes is far from trivial. In this paper, we seek for a cheaper alternative to ground-truth labels to train visual representations. First, starting from the observation that captions for images are often easier to collect compared to *e.g.* fine-grained category annotations, we have defined a new proxy task on image-caption pairs, namely image-conditioned masked language modeling (ICMLM), where image labels are automatically produced thanks to an efficient and effective way of leveraging their captions. Second, we have proposed a novel approach to tackle this proxy task which produces general purpose visual representations that perform on par with state-of-the-art self-supervised learning approaches on a variety of tasks, using a fraction of the data. This approach even rivals, on some settings, with a fully supervised pretraining on ImageNet. Such results are particularly relevant for domains where images are scarce but companion text is abundant.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016) 8
2. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: Proc. NeurIPS (2019) 4
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proc. ICLR (2015) 7
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR **3**(Jan) (2003) 5
5. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proc. ECCV (2018) 4
6. Caron, M., Bojanowski, P., Mairal, J., Joulin, A.: Unsupervised Pre-Training of Image Features on Non-Curated Data. In: Proc. ICCV (2019) 2, 4, 10, 11, 12
7. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. PAMI **40**(4) (2018) 3
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proc. ICML (2020) 2, 11
9. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: Proc. ICCV (2015) 3
10. Deng, C., Wu, Q., Wu, Q., Hu, F., Lyu, F., Tan, M.: Visual grounding via accumulated attention. In: Proc. CVPR (2018) 4
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: Proc. CVPR (2009) 2, 3, 11
12. Deshpande, A., Rock, J., Forsyth, D.: Learning large-scale automatic image colorization. In: Proc. ICCV (2015) 4
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: ACL (2019) 2, 3, 4, 6, 7, 8
14. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proc. ICCV (2015) 2, 4
15. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: Proc. ICCV (2017) 2
16. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results 11
17. Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: Proc. CVPR (2017) 2
18. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: Proc. ICLR (2018) 4, 12
19. Gomez, L., Patel, Y., nol, M.R., Karatzas, D., Jawahar, C.: Self-supervised learning of visual features through embedding images into text topic spaces. In: Proc. CVPR (2017) 2, 4, 5, 11, 12, 13
20. Gomez, R., Gomez, L., Gibert, J., Karatzas, D.: Chapter 9 - self-supervised learning from web data for multimodal retrieval. In: Multimodal Scene Understanding (2019) 4
21. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. IJCV (2017) 3
22. Gordo, A., Larlus, D.: Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In: Proc. CVPR (2017) 4

23. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: Proc. ICCV (2019) 2, 4, 14
24. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: Proc. CVPR (2017) 4
25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proc. CVPR (2020) 2, 4
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR (2016) 10, 11, 13
27. Hong, S., Yeo, D., Kwak, S., Lee, H., Han, B.: Weakly supervised semantic segmentation using web-crawled videos. In: Proc. CVPR (2017) 3
28. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear (2017), https://spacy.io 5
29. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proc. CVPR (2016) 4
30. Jenni, S., Favaro, P.: Self-supervised feature learning by learning to spot artifacts. In: Proc. CVPR (2018) 4
31. Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: Proc. ECCV (2016) 2, 3, 5
32. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting Self-Supervised Visual Representation Learning. In: Proc. CVPR (2019) 13
33. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017) 10
34. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: Proc. CVPR (2017) 4
35. Li, A., Jabri, A., Joulin, A., van der Maaten, L.: Learning visual n-grams from web data. In: Proc. ICCV (2017) 4, 5
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Proc. ECCV (2014) 8, 10
37. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proc. NeurIPS (2019) 4, 8
38. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proc. ECCV (2018) 2, 3, 7
39. Mahendran, A., Thewlis, J., Vedaldi, A.: Cross pixel optical flow similarity for self-supervised learning. In: Proc. ACCV (2018) 4
40. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv:1807.03748 (2018) 4
41. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: Proc. CVPR (2012) 2
42. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proc. CVPR (2016) 4
43. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proc. NAACL-HLT (2018) 2

44. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv:2001.07966 (2020) 2
45. Quattoni, A., Collins, M., Darrell, T.: Learning visual representations using images with captions. In: Proc. CVPR (2007) 2, 4, 5
46. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proc. NeurIPS (2015) 3, 4
47. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3) (2015) 3
48. Sariyildiz, M.B., Cinbis, R.G.: Gradient matching generative networks for zero-shot learning. In: Proc. CVPR (2019) 3
49. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proc. ICLR (2015) 10
50. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: Proc. ICLR (2020) 4
51. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proc. ICCV (2019) 4, 8
52. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Proc. EMNLP (2019) 4
53. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: The new data in multimedia research. arXiv:1503.01817 (2015) 3, 4, 12
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proc. NeurIPS (2017) 4, 8, 9
55. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 2
56. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proc. CVPR (2016) 4
57. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proc. CVPR (2018) 4
58. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. Proc. CVPR (2020) 3
59. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv:1905.00546 (2019) 2, 3
60. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Proc. ECCV (2016) 2, 4
61. Zhang, R., Isola, P., Efros, A.A.: Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In: Proc. CVPR (2017) 2, 4
62. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. PAMI (2017) 2
63. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proc. NeurIPS (2014) 11
64. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J.: Unified Vision-Language Pre-Training for Image Captioning and VQA. Proc. AAAI (2020) 4