# VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning

**Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang,**
**Jianfeng Gao, Lijuan Wang, Zicheng Liu**
Microsoft Corporation
{xiaowh, keli, lijuanw, leizhang, jfgao, zliu}@microsoft.com, yinxi.whu@gmail.com

## Abstract

It is highly desirable yet challenging to generate image captions that can describe novel objects which are unseen in caption-labeled training data, a capability that is evaluated in the novel object captioning challenge (nocaps). In this challenge, no additional image-caption training data, other than COCO Captions, is allowed for model training. Thus, conventional Vision-Language Pre-training (VLP) methods cannot be applied. This paper presents VIsual VOcabulary pre-training (VIVO) that performs pre-training in the absence of caption annotations. By breaking the dependency of paired image-caption training data in VLP, VIVO can leverage large amounts of paired image-tag data to learn a visual vocabulary. This is done by pre-training a multi-layer Transformer model that learns to align image-level tags with their corresponding image region features. To address the unordered nature of image tags, VIVO uses a Hungarian matching loss with masked tag prediction to conduct pre-training.

We validate the effectiveness of VIVO by fine-tuning the pre-trained model for image captioning. In addition, we perform an analysis of the visual-text alignment inferred by our model. The results show that our model can not only generate fluent image captions that describe novel objects, but also identify the locations of these objects. Our single model has achieved new state-of-the-art results on nocaps and surpassed the human CIDEr score.

## Introduction

Image captioning is a long-standing task in artificial intelligence (Farhadi et al. 2010; Kulkarni et al. 2013; Kuznetsova et al. 2012; Mitchell et al. 2012; Yang et al. 2011; Fang et al. 2015). The task is challenging in that it requires visual perception and recognition, and natural language generation grounded in perception and real-world knowledge (Kuznetsova et al. 2012; Yang et al. 2011). With recent progress in computer vision (He et al. 2017; Ren et al. 2015), natural language processing (Devlin et al. 2018; Radford 2018; Vaswani et al. 2017), and vision-language understanding (Li et al. 2020; Sharma et al. 2018; Zhou et al. 2020a), the performance on image captioning has been substantially improved on public benchmarks like COCO (Chen et al. 2015) and Flickr30k (Young et al. 2014). However, models trained on such datasets with limited visual concepts generalize poorly to in-the-wild images (Tran et al. 2016).

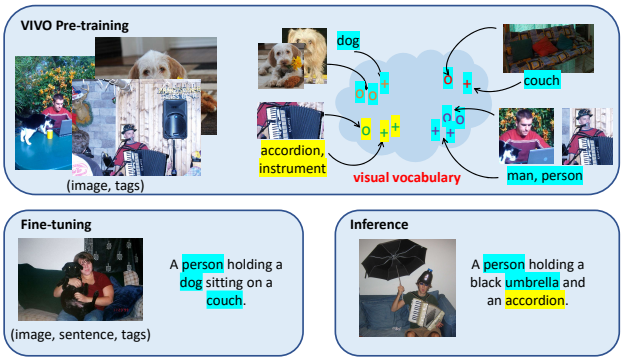



Figure 1: VIVO pre-training uses paired image-tag data to learn a rich visual vocabulary where image region features and tags of the semantically similar objects are mapped into vectors that are close to each other. Fine-tuning is conducted on paired image-caption data that only cover a limited numbers of objects (in blue). During inference, our model can generalize to describe novel objects (in yellow) that are learnt during VIVO pre-training.

To improve image captioning in the wild, the nocaps benchmark (Agrawal et al. 2019) is developed to evaluate Novel Object Captioning (NOC)[1] at scale. The training data for nocaps is the COCO dataset consisting of image-caption pairs and the Open Images dataset (Kuznetsova et al. 2020) containing bounding boxes and image-level tags. The test data consists of images selected from Open Images, containing nearly 400 objects that are not or rarely seen in the COCO dataset. This raises the challenge of how to generate captions that describe novel objects unseen in the paired image-caption training data. A common strategy is to resort to alternative data sources without caption supervision. Prior works on NOC (Lu et al. 2018; Wu et al. 2018) propose to generate template sentences that can be filled in with detected visual concepts for NOC. However, the relationship between image and text is not fully explored in their frameworks. We will show that the performance of NOC can be



[1] We use "NOC" to represent the task of novel object captioning and "nocaps" to refer to the nocaps benchmark.

significantly improved by pursuing image-text aligned representation learning.

In this paper, we present VIsual VOcabulary (VIVO) pre-training that leverages large amounts of vision data without caption annotations to learn a rich visual vocabulary for NOC. As shown in Figure 1, we define visual vocabulary as a joint embedding space where image region features and tags of semantically similar objects are mapped into vectors that are close to each other, *e.g.*, "person" and "man", "accordion" and "instrument". Once the visual vocabulary is pre-trained, we can fine-tune the model using image-caption pairs for caption generation. Note that the dataset used for fine-tuning only covers a small subset of the most commonly occurred objects in the learnt visual vocabulary. Nevertheless, our model can generalize to any images that contain similar scenes (*e.g.*, people sitting in couch in Figure 1) with novel objects unseen in the fine-tuning dataset, like "accordion", thanks to the pre-trained visual vocabulary.

The VIVO pre-training method is motivated to learn the cross-modality semantic alignment, similarly as in conventional Vision-Language Pre-training (VLP) methods. However, unlike existing VLP models which are pre-trained using image-caption pairs, VIVO is pre-trained on image-tag pairs. To the best of our knowledge, VIVO is the first VLP method that does not rely on caption annotations. Thus, it opens the possibility of leveraging, for VLP, many existing vision datasets originally developed for image tagging or object detection tasks like ImageNet (Deng et al. 2009), Open Images (Kuznetsova et al. 2020), Objects365 (Shao et al. 2019), etc. Moreover, we can also leverage large amounts of images, paired with machine-generated tags as weak supervision signals, for VLP.

VIVO pre-training aims to learn a joint representation of visual and text input. We feed to a multi-layer Transformer model an input consisting of image region features and a paired image-tag set. We then randomly mask one or more tags, and ask the model to predict these masked tags conditioned on the image region features and the other tags. Given that tags are not ordered, we employ the Hungarian matching loss (Stewart, Andriluka, and Ng 2016; Carion et al. 2020) for tag prediction optimization. Extensive experiments show that VIVO pre-training significantly improves the captioning performance on NOC. In addition, our model can precisely align the object mentions in a generated caption with the regions in the corresponding image.

In summary, we make the following contributions.

- We propose a new VIVO pre-training method that leverages large amounts of vision data without caption annotations for vision-language representation learning.
- We develop a Hungarian matching loss with masked tag prediction to conduct pre-training with image-tag pairs.
- With a single model, our method achieves the new state-of-the-art result on the nocaps benchmark and surpasses the human CIDEr score.

## Prior Work

**Image Captioning** Prior works on image captioning have focused on exploring different model structures and learning methods for different applications. For example, Song et al. (2019); Wang, Chen, and Hu (2019); Gao et al. (2019); Huang et al. (2019); Pan et al. (2020); Guo et al. (2020); Cornia et al. (2020) explore different attention mechanisms in captioning modeling. Other works improve the performance with reinforcement learning (Rennie et al. 2017; Li, Chen, and Liu 2019; Yang et al. 2020) or adversarial learning (Chen et al. 2019; Dognin et al. 2019). Different applications such as dense captioning (Johnson, Karpathy, and Fei-Fei 2016; Yin et al. 2019; Li, Jiang, and Han 2019), grounded captioning (Ma et al. 2020; Zhou et al. 2020b), image captioning with reading comprehension (Sidorov et al. 2020) have been studied. However, all these methods assume that most of the visual objects in test data are seen in training data. Thus, they do not work well for NOC, where the objects presented in test images are often unseen in the caption-annotated training data.

**Novel Object Captioning (NOC)** NOC requires a model to generate image captions that describe novel objects that are unseen in the paired image-caption training data. Since the task setting resembles that in real-world applications, it draws growing interest in the research community. The early works, such as Deep Compositional Captioner (Hendricks et al. 2016) and Novel Object Captioner (Venugopalan et al. 2017), propose to use unpaired image and sentence data to transfer knowledge among semantically similar visual concepts. Empirical evaluation on the COCO dataset by holding out 8 novel object categories suggests that these methods might be applicable to NOC.

Recent studies propose to explicitly leverage the object detection results for NOC. Yao et al. (2017) use LSTM-C with a copying mechanism to assemble the detected novel objects for caption generation. Neural Baby Talk (Lu et al. 2018) and Decoupled Novel Object Captioner (Wu et al. 2018) generate template sentences that are later filled in with visual concepts recognized by object detectors. Similarly, Constrained Beam Search (Anderson et al. 2017) is exploited to generate captions that contain detected novel objects (Agrawal et al. 2019).

None of the aforementioned methods for NOC fully exploits the relationship between image and text, which we argue is crucial to the quality of generated captions. In this study, we pre-train a Transformer model to learn a visual vocabulary where object tags are aligned with their corresponding image feature representations in a semantic space.

**Vision and Language Pre-training** Motivated by BERT (Devlin et al. 2018), many VLP methods have been proposed to learn vision-language representations by pre-training large-scale Transformer models (Lu et al. 2019; Tan and Bansal 2019; Su et al. 2019; Chen et al. 2020; Zhou et al. 2020a; Li et al. 2020). Most existing VLP methods are developed for understanding tasks such as image-text retrieval and visual question answering. Only a few of them (Zhou et al. 2020a; Li et al. 2020) can be applied to image captioning. But these methods use paired image-caption data for pre-training, and are not applicable to NOC. In this study, we break the dependency on image-caption pairs in VLP for the first time. The proposed VIVO pre-training learns vision-

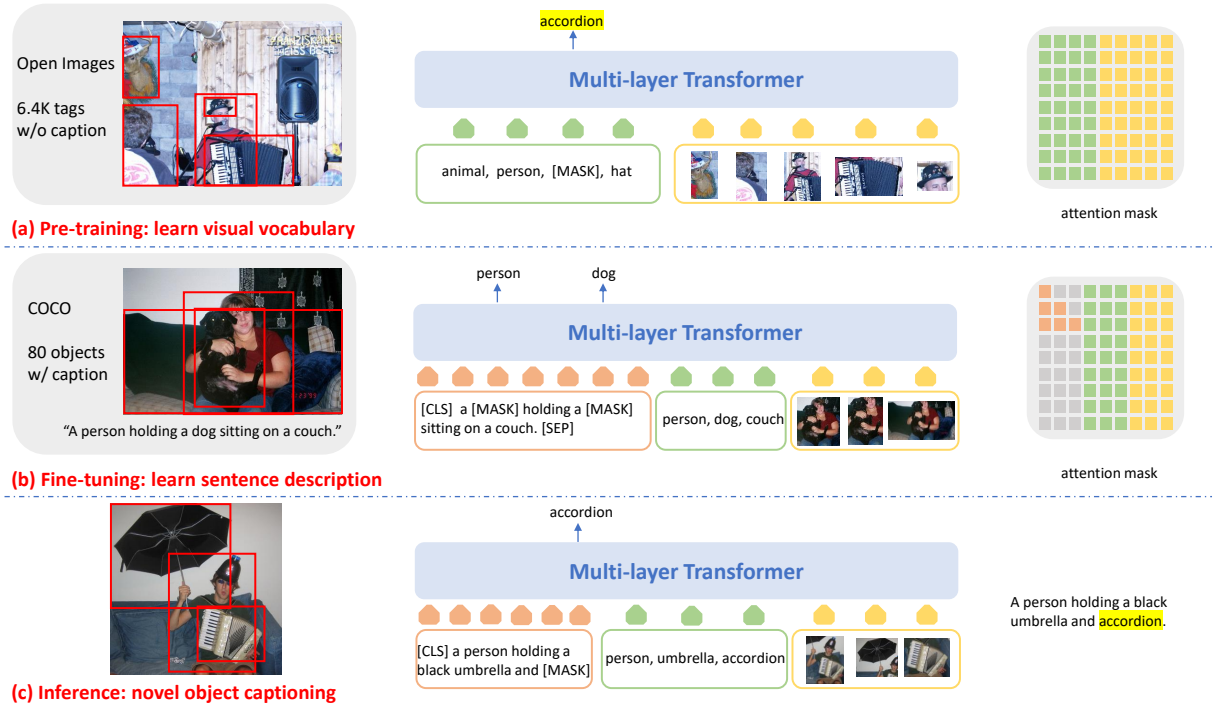



Figure 2: The proposed two-stage training scheme. (a) In VIVO pre-training, we train a Transformer-based model on image-tag pairs for tag prediction, where it learns cross-modal representations for rich visual concepts. (b) In fine-tuning, we train the same model on limited image-caption pairs to learn how to generate captions conditional on the image and tags. (c) During inference, given the image and detected tags, our model is applied iteratively to generate a sequence of words describing novel objects in an auto-regressive manner.

language alignment on image-tag pairs, improving the image captioning results on both NOC and the general image captioning task.

## Proposed Method

Recent image captioning models have achieved impressive results on the tasks where large amounts of paired image-caption training data is available. But they generalize poorly to images in the wild, where there are a wide variety of visual objects that are unseen in the caption corpora for training. For example, the models trained on COCO Captions can faithfully describe images containing objects such as "people", "dogs", or "a couch", but fail to generate a reasonable caption for any image containing "an accordion" since the object is unseen in COCO Captions.

To address this problem, we propose a weakly supervised learning approach to pre-training image captioning models on image-tag pairs that, compared to image-caption pairs, are of larger amounts and contain many more diverse visual objects. Our approach uses a two-stage training scheme that consists of VIVO pre-training and fine-tuning. Figure 2 illustrates our approach using an example. First, in the pre-training stage (Figure 2(a)), an image captioning model learns to label image regions using tags (*e.g.*, "person", "accordion") using image-tag pairs as training data, where the object "accordion" is included. Then in fine-tuning (Figure 2(b)), given image-caption pairs and their corresponding object tags detected (*e.g.*, "person" and "dog"), the model learns to map an image to a sentence conditioned on the detected objects, *e.g.*, "[A] holding [B] ...", where [A] and [B] could attend to object tags. While the sentences are learned from image-caption pairs, the object tags may refer to novel visual objects that are unseen in image-caption pairs (but seen in image-tag data in this example). Thus, our model achieves the compositionality generalization, allowing for zero-shot generalization to novel objects for image captioning. As shown in Figure 2(c), at inference time the model is able to recognize objects (*e.g.*, "person", "accordion") and compose familiar constituents in a novel way to form a caption "a person holding an accordion".

The model architecture is shown in Figure 3. It consists of multiple Transformer layers to encode the input into a feature vector and a linear layer with softmax to generate the text description of the visual objects in the image. In what follows, we describe in detail the way the model is pre-trained and fine-tuned.

### VIVO Pre-training

We pre-train the Transformer model on a large-scale dataset with abundant tags, *e.g.*, the Open Images training set with $6.4K$ classes of image-level tags. Unlike many existing VLP methods that rely on image-caption pairs, VIVO pre-training is conducted solely on image-tag pairs, which are much eas-

ier to collect by either human labeling or auto tagging. The training objective is to predict the missing (masked) tags given a bag of image-level tags and image regions. We denote the training set as $\mathbb{D} = \{\mathbf{I}_i, \mathbf{G}_i\}_{i=1}^{N}$ with $N$ images and their corresponding tags, where $\mathbf{G}_i = \{g_{ij}\}_{j=1}^{L_i}$ is a set of $L_i$ image-level tags that are associated with the image $\mathbf{I}_i$. These tags are textual labels of the visual objects presented in the image, *e.g.*, "person", "cat", "dinning table", etc. In the rest of the paper, we omit the subscript $i$ for simplicity.

We use a multi-layer Transformer model to learn a joint representation for both vision and language domains. The input to the Transformer model consists of image region features $\mathbf{V}$ and tag tokens $\mathbf{T}$, where $\mathbf{V} = \{\mathbf{v}_k\}_{k=1}^{K}$ are extracted from image $\mathbf{I}$ using a detector trained on Visual Genome dataset (Anderson et al. 2018), and $\mathbf{T} = \{t_j\}_{j=1}^{T}$ are tokenized tags in $\mathbf{G}$. During training, some tokens are randomly masked out for the model to predict.

The main difference between a caption and a set of tags is that words in the caption are ordered while tags are not ordered. This unordered nature may result in ambiguity in tag prediction when two tags are masked out simultaneously. For example, if the masked tokens are "dog" and "cat", we can predict each token in either position without restricting to the original position or order in the input. To resolve this issue, we propose to use the Hungarian matching loss (Stewart, Andriluka, and Ng 2016; Carion et al. 2020) to formulate the tag prediction as a set-matching problem.

We denote the set of $M$ masked tokens as $\tilde{\mathbf{T}} = \{t_m\}_{m=1}^{M}$ where $t_m$ is the token id in the vocabulary, and the prediction probabilities of the corresponding representations in the final layer of Transformer as $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{M}$ where $\mathbf{p}_i$ is the classification probabilities for the $i$-th masked position. Since the target tokens in $\tilde{\mathbf{T}}$ are unordered, we need an one-to-one mapping from $\tilde{\mathbf{T}}$ to $\mathbf{P}$ such that the prediction for each masked position is assigned one of the target tokens. Once such an assignment $\alpha$ is known, the loss is defined as:

$$L(\tilde{\mathbf{T}}, \mathbf{P}, \alpha) = \sum_{i=1}^{M}(-\log(\mathbf{p}_i(t_{\alpha(i)}))) \quad (1)$$

where $\alpha$ is a permutation of the $M$ indices, i.e., $\alpha(i)$ is the index of the target token assigned to the $i$-th prediction. Since the assignment is unknown, we want $\alpha$ to be the best possible mapping between $\tilde{\mathbf{T}}$ and $\mathbf{P}$. Formally, we define such best possible $\alpha$ to be the one that minimizes the following total cost among all the valid[2] permutations:

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{M} C(\mathbf{p}_i, t_{\alpha(i)}), \quad (2)$$

where $C(\mathbf{p}_i, t_m) = 1 - \mathbf{p}_i(t_m)$ is the cost function of assigning the target $t_m$ to the $i$-th prediction. The reason why we use $C(\mathbf{p}_i, t_m)$ instead of $-\log(\mathbf{p}_i(t_{\alpha(i)}))$ as in (1) is that it is bounded. Now we can compute the final loss as $L(\tilde{\mathbf{T}}, \mathbf{P}, \hat{\alpha})$, where $L$ is defined in (1) and $\hat{\alpha}$ is defined in (2).

[2]For a tag tokenized into multiple tokens, the order of tokens within the tag cannot be changed.

As shown in Figure 2 (a), we use bi-directional attention mask in VIVO pre-training. In order to predict a missing tag, the model will have to resort to image region features and the other tags. So it learns a joint representation containing information from both image regions and textual tags. This facilitates the cross-modality alignment between representations of image regions and tags.

### Fine-tuning and Inference

After pre-training, the Transformer model is fine-tuned on a dataset where both captions and tags are available, *e.g.*, the COCO set annotated with tags from 80 object classes and captions. The tags can also be automatically generated using a pre-trained tagging or detection model. Given image regions and tags, the model learns to predict the conditional caption sentence where some positions are randomly masked out. More specifically, the input to the model during fine-tuning is a triplet of image region features $\mathbf{V}$, a set of tags $\mathbf{T}$ and a caption $\mathbf{C}$, where $\mathbf{V}$ and $\mathbf{T}$ are constructed in the same way as described in pre-training, and $\mathbf{C}$ is a sequence of tokens. During fine-tuning, we randomly mask out some of the tokens in a caption sentence for prediction, and optimize the model parameters using the cross-entropy loss. To make the model generate captions from left to right at inference time, during fine-tuning we apply the uni-directional attention mask on a caption sequence to prevent the positions from attending to subsequent positions.

During inference, we first extract image region features and detect tags from a given image. Then the model is applied to generate a sequence, one token at a time, until it outputs the end of sentence token or reaches the maximum length. At each step the model is auto-regressive, consuming the previously generated tokens as additional input when generating the next.

In the next section, we present extensive experimental results, showing that our model can generate captions to describe novel objects and that the alignment between image regions and tags, learned from VIVO pre-training, is crucial to the model's superior performance on NOC.

## Experiments

### Experimental Settings

**Datasets** We use the Open Images V5 challenge training set, which has $1.7M$ images, for VIVO pre-training. We select 500 classes[3] from bounding box annotations and $6.4K$ classes from human verified image-level labels. The joint image-tag pairs, containing $6.4K$ unique classes in total, are used in VIVO pre-training. In the fine-tuning stage, the training data is the COCO training set of $118K$ images, each with 5 captions. We evaluate our model on the validation and test sets of nocaps, which consist of $4.5K$ and $10.6K$ images from the Open Images validation and test sets, respectively.

**Implementation Details** We use the object detector from UpDown (Anderson et al. 2018) to extract image region fea-

[3]Only 500 out of 600 objects are used in the challenge set, as we further refine the labels by removing classes that are "parts" (*e.g.*, human eyes).

tures, which are concatenated with scaled bounding boxes to form a 2054-dimension vector (2048D for the visual features and 6D for the bounding box encoding including top-left and bottom-up corners as well as the box's width and height). We use an object detector trained on the Open Images dataset to detect object tags for all datasets. For pre-training and fine-tuning, we also add the ground-truth tags from the training sets. No ground-truth tags are used on the nocaps validation and test sets. The Transformer model is initialized using BERT-base (Devlin et al. 2018) where we add a linear layer to transform the image region features to the vectors with same size as the word embeddings.

In VIVO pre-training, we use a maximum of 50 image regions and 15 tag tokens per image. The model is trained for $160K$ iterations (about 100 epochs) with a batch size of 1024 and a learning rate of $5 \times 10^{-5}$. In fine-tuning, we set the maximum caption length to 40 and the maximum tag length to 30. The model is trained for 30 epochs with a batch size of 256 and a learning rate of $5 \times 10^{-5}$, optimized using the cross-entropy loss. To further boost the performance, we perform the SCST optimization (Rennie et al. 2017) with a learning rate of $2 \times 10^{-6}$ for 5 epochs. During inference, we use greedy decoding to generate image captions with a maximum length of 20.

## Novel Object Captioning

We compare our method with UpDown (Anderson et al. 2018; Agrawal et al. 2019) and OSCAR[4] (Li et al. 2020), which holds the state-of-the-art result on the nocaps benchmark. The training data for the baselines is the COCO dataset. Following prior settings, we also report the results after our model is optimized using SCST (Rennie et al. 2017) and generates captions using Constrained Beam Search (CBS) (Anderson et al. 2017).

The evaluation results on nocaps validation and test sets are shown in Table 1. By leveraging VIVO pre-training on the Open Images dataset, our method has achieved significant improvement compared to all prior works. Our plain version (VIVO) already outperforms UpDown+ELMo+CBS and OSCAR by a large margin. It is worth noting that CBS brings absolute gains of 17.8% and 15.5% for UpDown and OSCAR, respectively, but it only improves VIVO by 3.8%. This suggests that our model is more capable of generating captions with novel objects without explicitly adding any constrains. Our best results are new state-of-the-art and surpasses the human CIDEr score on the overall dataset.

To quantitatively evaluate how well the model can describe novel objects, we also calculate the F1-score following Hendricks et al. (2016), where all the objects mentioned in the generated caption sentences are compared against the ground-truth object tags. Table 2 shows the comparison with OSCAR on the nocaps validation set. We see that VIVO improves OSCAR in F1-scores substantially especially for out-of-domain objects. This again verifies the effectiveness of VIVO pre-training in learning to recognize novel objects for NOC.

[4]We compare with OSCAR base whose model size is the same as ours. In fact, our model with 12 layers and hidden size of 768 even outperforms the OSCAR large model.

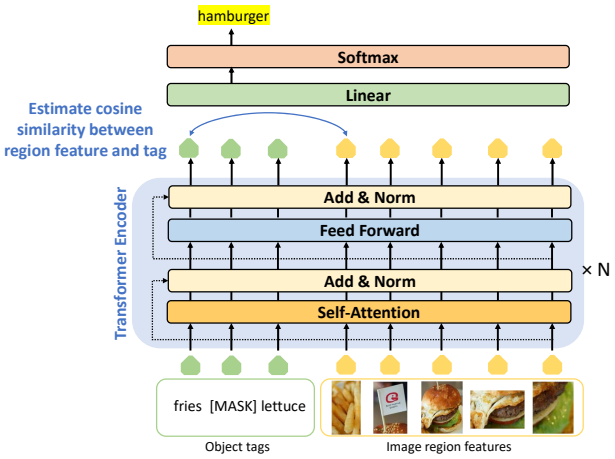



Figure 3: Overview of our VIVO pre-trained Transformer model. Our model consists of multiple Transformer encoder layers followed by a linear layer and a softmax layer. We use masked tag prediction to conduct pre-training. To analyze the visual-text alignment, we use the outputs of the last layer of the encoder layers to estimate the cosine similarity between the image region and tag.

Although object tags are used in both VIVO pre-training and fine-tuning stages, we show that the model's capability of generating captions that precisely describe novel objects at inference time attributes largely to pre-training. We compare the distribution of object tags on COCO and nocaps, which are generated by the object detector trained on the Open Images dataset and used for fine-tuning and inference, respectively. As shown in Table 3, COCO has a long-tail distribution where 415 out of 568 categories amounts only to 2.43% of all the tags. The under-representation of novel objects makes the trained model statistically unlikely to generate plausible captions that describe these novel objects. Therefore, our VIVO pre-training, which mitigates the data imbalance issue by leveraging diverse tags in image-tag pairs, is crucial to improving model's generalization property, as empirically demonstrated on NOC.

## Visual-Text Alignment

To further understand the effects of VIVO pre-training in learning visual vocabulary, which aligns image regions with object tags, we show how the novel object tags can be grounded in image regions in Figure 4. Given the images from the Open Images validation set, we extract image region features using the same object detector from UpDown and generate captions from the captioning model with VIVO pre-training. After identifying the novel objects in the generated captions, as shown in Figure 3, we feed the novel object tags, together with the extracted image region features, to the VIVO pre-trained Transformer model. The output of the last encoder layer is used as the contextualized representation of the corresponding input. We then calculate the cosine similarity between representations of each pair of image region and object tag. We highlight the pairs with high scores in Figure 4. The result shows that our model can precisely

| method | in-domain | | near-domain | | out-of-domain | | overall | |
|---|---|---|---|---|---|---|---|---|
| | CIDEr | SPICE | CIDEr | SPICE | CIDEr | SPICE | CIDEr | SPICE |
| Validation Set | | | | | | | | |
| UpDown (Agrawal et al. 2019) | 78.1 | 11.6 | 57.7 | 10.3 | 31.3 | 8.3 | 55.3 | 10.1 |
| UpDown + CBS | 80.0 | 12.0 | 73.6 | 11.3 | 66.4 | 9.7 | 73.1 | 11.1 |
| UpDown + ELMo + CBS | 79.3 | 12.4 | 73.8 | 11.4 | 71.7 | 9.9 | 74.3 | 11.2 |
| OSCAR (Li et al. 2020) | 79.6 | 12.3 | 66.1 | 11.5 | 45.3 | 9.7 | 63.8 | 11.2 |
| OSCAR + CBS | 80.0 | 12.1 | 80.4 | 12.2 | 75.3 | 10.6 | 79.3 | 11.9 |
| OSCAR + SCST + CBS | 83.4 | 12.0 | 81.6 | 12.0 | 77.6 | 10.6 | 81.1 | 11.7 |
| VIVO | 88.8 | 12.9 | 83.2 | 12.6 | 71.1 | 10.6 | 81.5 | 12.2 |
| VIVO + CBS | 90.4 | 13.0 | 84.9 | 12.5 | 83.0 | 10.7 | 85.3 | 12.2 |
| VIVO + SCST + CBS | **92.2** | 12.9 | **87.8** | 12.6 | 87.5 | 11.5 | **88.3** | 12.4 |
| Human | 84.4 | **14.3** | 85.0 | **14.3** | **95.7** | **14.0** | 87.1 | **14.2** |
| Test Set | | | | | | | | |
| VIVO + SCST + CBS | **89.0** | 12.9 | **87.8** | 12.6 | 80.1 | 11.1 | **86.6** | 12.4 |
| Human | 80.6 | **15.0** | 84.6 | **14.7** | **91.6** | **14.2** | 85.3 | **14.6** |

Table 1: Evaluation on nocaps validation and test sets.

| model | in-domain | out-of-domain | entire |
|---|---|---|---|
| OSCAR (Li et al. 2020) | 39.5 | 15.7 | 20.7 |
| VIVO | **46.3** | **30.6** | **33.8** |

Table 2: Comparison of F1-scores (in %) on object classes of Open Images, evaluated on the nocaps validation set. There are 504 classes in total. 105 of them are in-domain, which are 80 common classes from COCO and 25 objects frequently appearing in COCO Captions. The remaining 399 classes are the out-of-domain objects.

| #occur in COCO (<=) | 0 | 10 | 100 | $1K$ | $10K$ |
|---|---|---|---|---|---|
| #categories | 194 | 274 | 415 | 522 | 563 |
| percentage in COCO | 0.0 | 0.14 | 2.43 | 15.62 | 64.01 |
| percentage in nocaps | 0.24 | 5.05 | 15.98 | 35.71 | 69.91 |

Table 3: Distribution of 568 object categories on COCO training images and nocaps validation images. Each column is a subset of object categories whose number of occurrences are below the threshold. The percentage is calculated by dividing the counts of those objects by the total counts of all objects in the dataset.

align the mentions of these novel objects in captions with the corresponding image regions.

## General Image Captioning

VIVO pre-training does not require the paired image-caption data for model training as in conventional VLP methods. It opens up an opportunity to leverage additional data sources to improve image captioning models. To demonstrate the effectiveness of VIVO pre-training on general image captioning tasks, we trained two versions of OSCAR, following the setting in Li et al. (2020). The first OSCAR model is trained solely on Conceptual Captions (CC) (Sharma et al. 2018), as described in Li et al. (2020). The second OSCAR model is pre-trained using VIVO on Open Images (OI), and then fine-tuned on CC. As shown in Table 4, VIVO pre-training

| pre-training | BLEU4 | Meteor | CIDEr | SPICE |
|---|---|---|---|---|
| NO | 33.7 | 27.9 | 114.7 | 21.2 |
| OSCAR | 34.8 | **28.4** | 118.2 | 21.6 |
| OSCAR + VIVO | **34.9** | **28.4** | **119.8** | **21.7** |

Table 4: Evaluation on COCO Karpathy test split (Karpathy and Fei-Fei 2015). All results are based on single model with cross-entropy optimization.

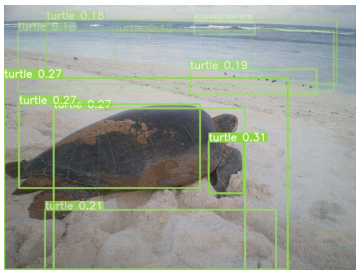
improves the model performance across all metrics evaluated on the COCO test set, especially in CIDEr score. We do observe, however, that the gain on the COCO benchmark is not as substantial as that on the nocaps benchmark. We conjecture that this is due to the COCO dataset containing only a small number of visual concepts and thus diminishing the benefit of learning a large visual vocabulary. It is also worth noting that using machine-generated image tags rather than human-written captions makes it possible to utilize potentially unlimited amounts of images, which we will pursue in our future work.

| Tag size | BLEU4 | Meteor | CIDEr | SPICE |
|---|---|---|---|---|
| 0 (w/o VIVO) | 18.3 | 24.2 | 69.6 | 11.3 |
| 500 classes | 20.6 | 25.4 | 76.5 | 11.9 |
| $6.4K$ classes | **21.2** | **25.4** | **77.8** | **12.0** |

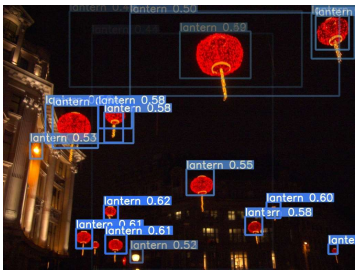
Table 5: Adding VIVO pre-training makes substantial improvement on NOC. Using more labels in pre-training also gives better results. All the models are fine-tuned on COCO and evaluated on the validation set of nocaps.

| Loss | BLEU4 | Meteor | CIDEr | SPICE |
|---|---|---|---|---|
| Mask only one token | 20.6 | 25.2 | 74.9 | 11.8 |
| w/o Hungarian matching | 21.0 | 25.4 | 75.8 | 11.8 |
| w/ Hungarian matching | **21.2** | **25.4** | **77.8** | **12.0** |

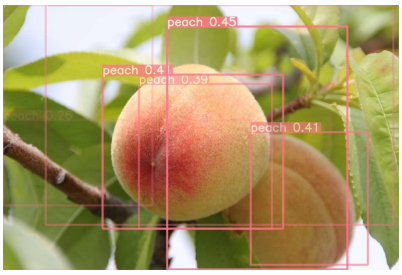
Table 6: Ablation study of the proposed Hungarian matching loss.



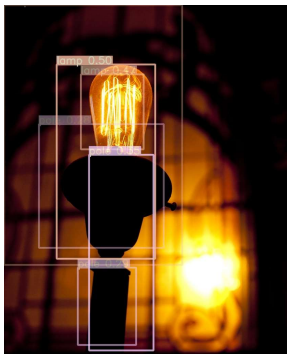
B: a large piece of art is displayed on the beach
V: a **turtle** that is laying down on the beach



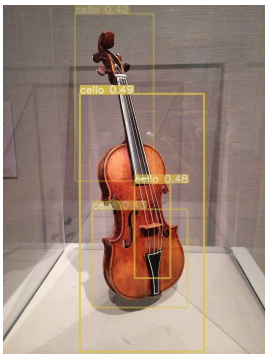
B: a group of four colored light up in the night sky
V: a bunch of red **lantern** lights on a street



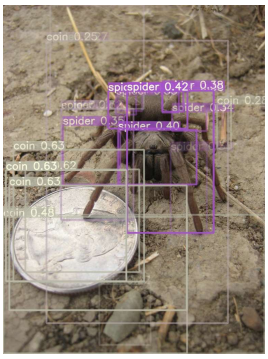
B: a close up of a fruit with leaves
V: a close up of a **peach** on a tree branch



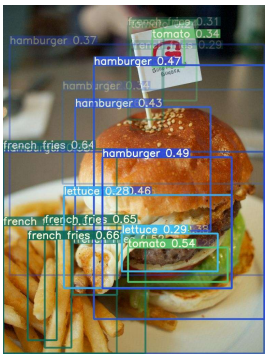
B: a street light with a yellow light in the background
V: a **lamp** that is on top of a **pole**



B: a small orange vase with a handle on a table
V: a **cello** is on display in a glass case



B: a spider sitting on top of a plate on a dirt ground
V: a **spider** sitting on the ground next to a **coin**



B: a hamburger and fries on a plate
V: a **hamburger** with **lettuce** and **tomato** on a plate with **french fries**

Figure 4: Image captioning results on nocaps. B: our baseline without adding VIVO pre-training. V: our approach with VIVO pre-training. Red text represents novel objects. For each image, we show the similarity scores of each image region to the novel objects appear in the captions. The bounding box color is brighter when the similarity is higher.

## Ablation Study

We select a subset of $10\%$ images from the Open Images training set to conduct an ablation study. We fine-tune with cross-entropy loss on the COCO dataset and report the performance on the nocaps validation set.

**Using a Larger Set of Tags** We investigate whether using a larger set of tags in pre-training improves performance of the downstream image captioning task. We select $500$ classes of objects, which are used to train the object detector, from the overall $6.4K$ classes of tags to conduct VIVO pre-training. As shown in Table 5, VIVO pre-training with $500$ classes significantly improves the performance on nocaps by $6.9\%$ compared to no pre-training. Expanding the labels to $6.4K$ classes can further improve the performance, although the gain is limited due to the increased diversity of objects presented in test images.

**Using Hungarian Matching Loss** We evaluate the effectiveness of the proposed Hungarian matching in VIVO pre-training to predict a set of tags. Training without Hungarian matching reduces the tag prediction to the standard masked language modeling task, which predicts the masked tokens in the same order as that in the input sequence. In addition, we also perform VIVO pre-training by masking only one token in input, which makes word order information not useful. The evaluation results on the nocaps validation set are in Table 6. We can see that masking only one token is not effective, and using Hungarian matching leads to the best model performance.

## Conclusions

We have presented a weakly supervised learning approach to training image captioning models in two steps. First, a Transformer-based model is pre-trained on large amounts of image-tag pairs to learn a visual vocabulary without the need of using image-caption pairs which are harder to obtain. Then, the model is fine-tuned on image-caption pairs to learn to incorporate information from the pre-trained visual vocabulary and compose image captions that can describe novel visual objects unseen in the training data of image-caption pairs.

Our experiments on the nocaps benchmark dataset demonstrate that our model achieves compositional generalization, allowing for zero-shot generalization to novel objects for image captioning. As a result, our best single model creates new state-of-the-art that surpasses the human CIDEr score on nocaps. A detailed analysis reveals that the generalization is attributed to a large degree to the visual vocabulary learned in model pre-training, which maps visual objects or regions with similar semantic meanings to feature vectors that are close to each other in a discrete semantic space.

Since our pre-training does not need paired image-caption data, one of our future works is to leverage large amounts of vision data, beyond image-tag pairs used in this paper, to significantly improve the quality of the visual vocabulary.

## Acknowledgements

We thank Jianfeng Wang, Ehsan Azarnasab, Lin Liang, Pengchuan Zhang, Xiujun Li, Chunyuan Li, Jianwei Yang, Yu Wang, Houdong Hu, Furu Wei, Dong Li for valuable discussions and comments.

## References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. nocaps: novel object captioning at scale. In *ICCV*.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2017. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *ECCV*.

Chen, C.; Mu, S.; Xiao, W.; Ye, Z.; Wu, L.; and Ju, Q. 2019. Improving image captioning with conditional generative adversarial nets. In *AAAI*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* .

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: Learning universal image-text representations. In *ECCV*.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-Memory Transformer for Image Captioning. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Dognin, P.; Melnyk, I.; Mroueh, Y.; Ross, J.; and Sercu, T. 2019. Adversarial semantic alignment for improved image captions. In *CVPR*.

Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; et al. 2015. From captions to visual concepts and back. In *CVPR*.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*.

Gao, L.; Fan, K.; Song, J.; Liu, X.; Xu, X.; and Shen, H. T. 2019. Deliberate attention networks for image captioning. In *AAAI*.

Guo, L.; Liu, J.; Zhu, X.; Yao, P.; Lu, S.; and Lu, H. 2020. Normalized and Geometry-Aware Self-Attention Network for Image Captioning. In *CVPR*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *ICCV*.

Hendricks, L. A.; Venugopalan, S.; Rohrbach, M.; Mooney, R.; Saenko, K.; and Darrell, T. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *ICCV*.

Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(12): 2891–2903.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Duerig, T.; et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV* .

Kuznetsova, P.; Ordonez, V.; Berg, A.; Berg, T.; and Choi, Y. 2012. Collective generation of natural image descriptions. In *ACL*.

Li, N.; Chen, Z.; and Liu, S. 2019. Meta learning for image captioning. In *AAAI*.

Li, X.; Jiang, S.; and Han, J. 2019. Learning Object Context for Dense Captioning. In *AAAI*.

Li, X.; Yin, X.; Li, C.; Hu, X.; Zhang, P.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *CVPR*.

Ma, C.-Y.; Kalantidis, Y.; AlRegib, G.; Vajda, P.; Rohrbach, M.; and Kira, Z. 2020. Learning to Generate Grounded Visual Captions without Localization Supervision. In *ECCV*.

Mitchell, M.; Dodge, J.; Goyal, A.; Yamaguchi, K.; Stratos, K.; Han, X.; Mensch, A.; Berg, A.; Berg, T.; and Daumé III, H. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-Linear Attention Networks for Image Captioning. In *CVPR*.

Radford, A. 2018. Improving Language Understanding by Generative Pre-Training.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. TextCaps: a Dataset for Image Captioning with Reading Comprehension. In *ECCV*.

Song, L.; Liu, J.; Qian, B.; and Chen, Y. 2019. Connecting Language to Images: A Progressive Attention-Guided Network for Simultaneous Image Captioning and Language Grounding. In *AAAI*.

Stewart, R.; Andriluka, M.; and Ng, A. Y. 2016. End-to-end people detection in crowded scenes. In *CVPR*.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*.

Tran, K.; He, X.; Zhang, L.; Sun, J.; Carapcea, C.; Thrasher, C.; Buehler, C.; and Sienkiewicz, C. 2016. Rich image captioning in the wild. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Venugopalan, S.; Anne Hendricks, L.; Rohrbach, M.; Mooney, R.; Darrell, T.; and Saenko, K. 2017. Captioning images with diverse objects. In *CVPR*.

Wang, W.; Chen, Z.; and Hu, H. 2019. Hierarchical attention network for image captioning. In *AAAI*.

Wu, Y.; Zhu, L.; Jiang, L.; and Yang, Y. 2018. Decoupled novel object captioner. In *ACM Multimedia*.

Yang, X.; Zhang, H.; Jin, D.; Liu, Y.; Wu, C.-H.; Tan, J.; Xie, D.; Wang, J.; and Wang, X. 2020. Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards. In *ECCV*.

Yang, Y.; Teo, C.; Daumé III, H.; and Aloimonos, Y. 2011. Corpus-guided sentence generation of natural images. In *EMNLP*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*.

Yin, G.; Sheng, L.; Liu, B.; Yu, N.; Wang, X.; and Shao, J. 2019. Context and attribute grounded dense captioning. In *CVPR*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2: 67–78.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J. J.; and Gao, J. 2020a. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*.

Zhou, Y.; Wang, M.; Liu, D.; Hu, Z.; and Zhang, H. 2020b. More Grounded Image Captioning by Distilling Image-Text Matching Model. In *CVPR*.