arXiv:2112.05230v2 [cs.CV] 31 Mar 2022

Injecting Semantic Concepts into End-to-End Image Captioning

Zhiyuan Fang♠, Jianfeng Wang♡, Xiaowei Hu♡, Lin Liang♡, Zhe Gan♡,
Lijuan Wang♡, Yezhou Yang♠, Zicheng Liu♡
♠Arizona State University, ♡Microsoft Corporation
{zy.fang, yz.yang}@asu.edu
{jianfw, xiaowei.hu, lliang, zhe.gan, lijuanw, zliu}@microsoft.com

Abstract

Tremendous progresses have been made in recent years in developing better image captioning models, yet most of them rely on a separate object detector to extract regional features. Recent vision-language studies are shifting towards the detector-free trend by leveraging grid representations for more flexible model training and faster inference speed. However, such development is primarily focused on image understanding tasks, and remains less investigated for the caption generation task. In this paper, we are concerned with a better-performing detector-free image captioning model, and propose a pure vision transformer-based image captioning model, dubbed as ViTCAP, in which grid representations are used without extracting the regional features. For improved performance, we introduce a novel Concept Token Network (CTN) to predict the semantic concepts and then incorporate them into the end-to-end captioning. In particular, the CTN is built on the basis of a vision transformer, and is designed to predict the concept tokens through a classification task, from which the rich semantic information contained greatly benefits the captioning task. Compared with the previous detector-based models, ViTCAP drastically simplifies the architectures and at the same time achieves competitive performance on various challenging image captioning datasets. In particular, ViTCAP reaches 138.1 *CIDEr scores on COCO-caption Karpathy-split,* 93.8 *and* 108.6 *CIDEr scores on nocaps and Google-CC captioning datasets, respectively.*

1. Introduction

The task of image captioning aims to generate human-readable descriptive text from an image. Recent studies have witnessed its great development which are primarily reflected in the aspects of more advanced cross-modal fusion architectures [11, 53, 58, 63, 66, 73, 75, 77, 81]; more expressive object-centric features [4, 79] & tags [18, 25, 38, 67] obtained from a pre-trained object detection model; or learning general ***V***ision and ***L***anguage (VL) representations from large image-text corpus [18, 38, 67, 71, 72, 82].

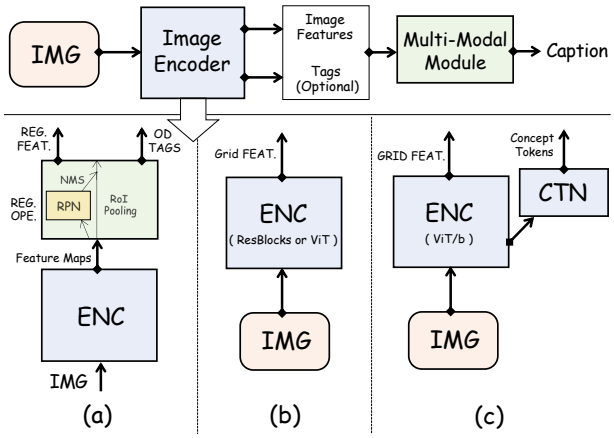


Figure 1. Comparisons of different image captioning models. Top: A general image captioning pipeline. Bottom: (a). Prevailing conventional models [25, 39, 79] which are based on an object detector to extract regional features. Object tags [38, 79] can be optionally used to assist the text generation through a multi-modal decoder network. This usually requires regional operations (REG. OPE.) that are time consuming. (b). To eliminate the detection module, a ResNet variant [22] or Vision Transformer [33] can be applied as substitution to output the grid feature [71, 72]. This replacement has been studied on the image understanding task recently but very few works focus on the generation task. (c). Our proposed ViTCAP, which is detector-free and incorporates a novel Concept Token Network to predict semantic concepts as tokens for the image captioning task.

Despite these significant advances, most of the mainstream captioning models [11, 53, 63, 81] rely heavily on a bulky object detector to provide regional visual representations for the multimodal interaction, as shown in Figure 1-a. In spite of the superior performance brought by the object features, the ensuing difficulties occur as they: 1) **lead to heavy computational load** due to the regional operations (*i.e.*, RPN, RoI Pooling, and NMS). These intermediate operations unavoidably cause training inefficiency and high inference latency at prediction stage [33, 67]; 2) **require**

box annotations and largely limit the flexibility in training and application. To address these challenges, there is an emerging trend that more recent works propose to eliminate the detector for the VL pre-training in an end-to-end fashion [28, 29, 33, 71, 74]. In such detector-free design, a general visual encoder serves as a substitute for the detector and from which the grid features are produced for later cross-modal fusion, as in Figure 1-b. Heretofore, the majority of these works mainly focus on the image understanding task, which is typically cast as a classification problem, and only a few of them shed light on the generation task. In [72], the image is encoded with ResNet [22] and the performance (117.3 CIDEr on COCO [72]) is still far from the state-of-the-art detector-based approach (129.3 CIDEr with VinVL-base [79]). The challenge remains uncharted and insufficiently investigated regarding *how to build a stronger detector-free image captioning model.*

Previous efforts [18, 25, 38, 67, 79] have demonstrated that the object tags play an important role in improving the captioning performance. Instead of gleaning the object tags from the detector, we introduce a novel fully **VI**sion **T**ransformer based image **CAP**tioning model, dubbed ViTCAP, with a lightweight Concept Token Network (CTN) that produces concept tokens (see Figure 1-c). ViTCAP is constructed on the basis of a vision transformer [13] as the stem image encoder. Our vision transformer backbone starts with encoding the image and produces grid features, on top of which the CTN branch is then applied to predict semantic concepts of images. We represent the semantic concepts at the token level instead of the tag level to avoid the tokenization. The multi-modal module then takes the input of both grid representations and Top-K concept tokens for decoding. During training, the CTN is optimized to predict the pseudo ground-truth concepts extracted from image captions via a simple classification task. We also investigate to adopt the object tags from the detector as the pseudo ground-truth, and empirically observe no further improvement. Overall, this straight-forward design allows the injection of semantic concepts into the multi-modal fusion module with abundant semantics, and is critical for the improved captioning performance.

Our ablative analysis suggests that, with no bells and whistles, simple vanilla transformer architecture based ViTCAP 1) significantly outperforms existing detector-free captioning models; 2) surpasses most detector-based models and 3) approaches the state-of-the-art detector-based models. In particular, ViTCAP achieves 138.1 CIDEr scores on COCO-caption Karpathy split [43], 108.6 on Google-CC [61], and 95.4 on nocaps [1] datasets.

To summarize our contributions:

- We present a detector-free image captioning model ViTCAP with fully transformer architecture, where it leverages grid representations without regional operations.
- We propose to inject semantic concepts into end-to-end captioning by learning from open-form captions. We find that our proposed concept classification training and concept tokens significantly benefit the captioning task.
- Extensive evaluations on multiple captioning datasets confirm the validity of our method. ViTCAP achieves competitive or even leading results amongst detector-based prior arts with clear inference-time advantages.

2. Related Work

Image Captioning aims to produce an open-form and human-readable textual description that summarizes the content of an image. Most previous captioning models unanimously [4, 15, 21, 25, 58, 63, 66, 73, 77] use detector based visual encoder like Faster-RCNN [57] to extract visual features, and apply decoders like RNN, LSTM or Transformer for caption generation. Existing efforts on image captioning are reflected from the perspective of novel architectures [11,53,81], more effective learning objectives [25, 48, 58], or large-scale VL pre-training [38, 79, 82], *etc*. Some recent works [4, 81] arrive at an empirical conclusion that a strong object detector is necessary, providing clean and unambiguous regional features for objects. Li *et al*. [25, 38] show that object tags output from the detectors play a critical role as anchoring points in VL tasks across modalities. Following this, [79] proposes to adopt a strengthened detector to obtain regional features and expanded object tags covering both entities and attributes for VL tasks. Nevertheless, object detectors hinder the VL models to be deployed on edge devices, known for their snail's pace at inference.

Efficient VL Models. Several recent efforts build efficient VL Models that either optimize the object detector for feature extraction with faster inference speed, or adopt non-detector image encoders. For instance, MiniVLM [67] first proposes an EfficientNet [62] based lightweight detector. [29] revisits grid features for VQA task with great performance and fast inference speed. [14, 28, 33, 71, 72] also inherit such detector-free design and use architecture like ResBlocks [22] for image encoding. On the other side, DistillVLM [18] introduces VL distillation that facilities VL pre-training & fine-tuning for small transformer architectures; [20] proposes to prune the transformer architecture and shows that close performance can be maintained at 50%-70% model sparsity.

3. ViTCAP

Existing image captioning models usually consist of an object detector module (**Detector**) to extract regional feature ($\boldsymbol{v}^T$) from the raw image ($\mathbf{I}$), and a multi-modal module (**MM**) to generate a textual description ($\boldsymbol{c}$). Several recent works [38, 79] show that the object tags ($\boldsymbol{t}^T$) extracted from the detector can serve as anchoring points across modalities, and are essential for various VL tasks. This procedure can

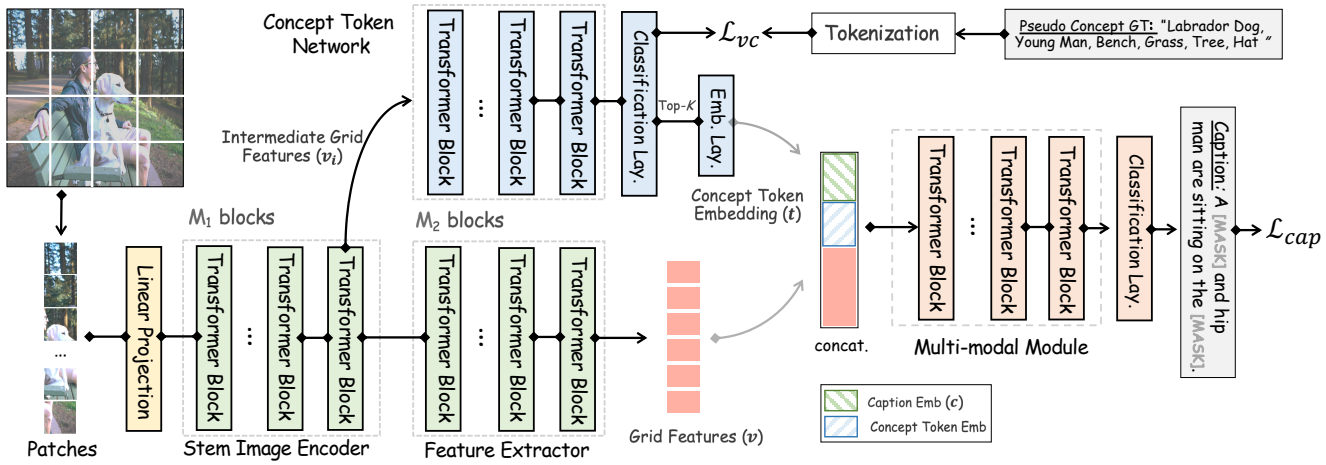


Figure 2. Architecture of our proposed ViTCAP image captioning model. ViTCAP is a detector-free image captioning model based on the vision transformer, where image patches are encoded into continuous embeddings as grid representations. The CTN branch roots from an intermediate block of the image encoder, and is a shallow transformer architecture (*e.g.*, 4 self-attention blocks). The CTN is trained via a classification task using object tags gleaned from the Teacher VLM's detector as pseudo-labels and the keywords parsed from image captions as the semantic concept ground-truth. During captioning, the CTN-produced concept tokens from the semantic concept vocabulary are then concatenated with the grid representations and fed into the multi-modal module for decoding. Best viewed in color.

be expressed as follows:

$$(\boldsymbol{v}^T, \boldsymbol{t}^T) = \textbf{Detector}(\mathbf{I}), \quad \boldsymbol{c} = \textbf{MM}(\boldsymbol{v}^T, \boldsymbol{t}^T). \quad (1)$$

Several VL models [28, 33, 72] obtain a great improvement in inference speed by using general image encoders without regional operations. However, these models are unable to utilize the image tags due to the absence of a detector.

In this work, we aim to build a detector-free captioning model with concept tokens containing rich semantics, coming from a novel Concept Token Network (CTN). An overview of ViTCAP is depicted in Figure 2. The raw image is firstly fed into the image encoder to generate the intermediate representations ($\boldsymbol{v}_i$) and the final grid representations ($\boldsymbol{v}$). A CTN branch then takes $\boldsymbol{v}_i$ as the input and predicts concept tokens ($\boldsymbol{t}$), followed by the multi-modal module that allows the interactions across modalities and generates caption ($\boldsymbol{c}$). We adopt the fully transformer [64] framework in all modules, but the image encoder and CTN modules are not architecture-specific. The overall pipeline can be summarized as:

$$(\boldsymbol{v}_i, \boldsymbol{v}) = \textbf{Encoder}(\mathbf{I}), \quad \boldsymbol{t} = \textbf{CTN}(\boldsymbol{v}_i), \quad \boldsymbol{c} = \textbf{MM}(\boldsymbol{v}, \boldsymbol{t}). \quad (2)$$

In the following, we first introduce how the vision transformer produces grid representations and our proposed CTN in Section 3.1, and the overall training losses in Section 3.2.

3.1. Model Structure

Vision Transformer. The transformer architecture and its instantiations (*e.g.*, BERT [12], GPT [7]) are well-known for their remarkable performances on natural language processing tasks, which are mostly attributed to the self-attention design. Recent efforts have advanced this to vision tasks, *i.e.*, Vision Transformer (ViT) [13]. We use ViT as the backbone of the image encoder to produce grid representations ($\boldsymbol{v}_i$ and $\boldsymbol{v}$). To be specific, the raw image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ is partitioned into N disjoint patches. The size of each patch is $P\times P\times 3$ and the number of patches N is $(HW)/P^2$. These patches are then flattened and projected into patch embedding of dimension d via a trainable linear projection layer. Concatenated with a special `[CLS]` token, these patch representations are added with learnable positional embeddings and then sent into M consecutive transformer blocks thereafter. To this end, we use the final representation as the grid features $\boldsymbol{v}$, and extract the output of the first M_1 blocks as the intermediate representations $\boldsymbol{v}_i$, which is the input of the Concept Token Network for concept predictions as detailed below.

Concept Token Network. The Concept Token Network (CTN) is composed of M_2 transformer blocks to process the intermediate features v_i. The output representation corresponding to `[CLS]` is used to predict the concept token with a multi-linear perceptual (MLP) network. The vocabulary of the concept token is identical with the one used for the captions. It is noted that we predict the concept in the token level rather than in the tag level, and thus the top-K ($K = 50$ in our experiments) tokens can be directly used by the multi-modal decoding module for auto-regressive decoding. In [38, 79], the object tags are predicted from the object detector, while we eliminate the detection module to remove the dependency of the box annotations. Another difference lies in the tag/concept vocabulary. The existing approaches apply the tag list from the dataset as the vocabulary which

are pre-defined and need an extra tokenization operation. Instead, our concept token vocabulary is shared with the one for captions and also removes the tokenization step.

Multi-Modal Fusion Module. Our multi-modal fusion module is a shallow network composed of multiple transformer blocks, and we follow [7, 55] to apply the `seq2seq` attention mask to generate the caption token in an auto-regressive way. First, the Top-K concept tokens' indices are mapped to token embeddings through an embedding layer l_c. Then, the module takes as input the concatenation of concept token embeddings ($\boldsymbol{t}$) and grid representations ($\boldsymbol{v}$) to generate the description, where we append a mask token `[MASK]` to the previous generated tokens (empty at very beginning) to predict the next token one by one. With the `seq2seq` attention mask, the generated token (including the appended `[MASK]` token) is able to access the preceding tokens and $(\boldsymbol{t}, \boldsymbol{v})$, while $(\boldsymbol{t}, \boldsymbol{v})$ has no access to the generated tokens. The generated caption token is also mapped through an embedding layer l_d. In experiments, we make the two embedding layers (l_c and l_d) shared to reduce the parameter size as the result is similar to two separate layers (see Appendix for results).

3.2. Model Training

The training of ViTCAP is composed of the CTN and the captioning training.

CTN is used to predict the image concepts. However, the widely-used VL pre-training dataset contains only the image descriptions without the tags. To address the issue, one can simply retrieve the concepts from the open-form captions (*e.g.*, by extracting nouns or adjective words as keywords) as the pseudo ground-truth concepts, or alternatively leverage a pre-trained object detector (*e.g.* on Visual Genome [34]) to produce the image tags (remove the bounding boxes). Empirically, we observe that by using caption extracted concepts lead to better results. We optimize the CTN to predict the target concepts via a multi-label classification task. Due to the extremely imbalanced semantic concepts distribution (certain concepts appear much frequently than the rest), we adopt the simplified asymmetric focal loss [6, 43, 44] which shows great performances handling sample imbalance problems for the multi-label classification task. The overall concept classification loss can be expressed as:

$$\mathcal{L}_{vc} = \mathbb{E}_{\boldsymbol{v}_i \sim D} f_\theta(p \mid \boldsymbol{v}_i), \tag{3}$$

$$f_\theta(p \mid \boldsymbol{v}_i) = \frac{1}{K} \sum_{k=1}^{K} \begin{cases} (1-p_k)^{\gamma_+} \cdot \log(p_k), & +, \\ p_k^{\gamma_-} \cdot \log(1-p_k), & -, \end{cases} \tag{4}$$

$p_k \in [0, 1]$ denotes the output probability for the k-th class and $\pm$ specifies whether the class is the pseudo ground-truth concept. Despite the rarity of positive samples, setting parameters $\gamma_+ < \gamma_-$ decouples its decay rates from the deluge of negative samples and emphasizes more the contribution of the positive. We set parameters $\gamma_+ = 0$ and $\gamma- = 1$ as [44] in our experiment.

For the captioning training, the multi-modal module takes the Caption-Concept Token-Feature triple $(\boldsymbol{c}, \boldsymbol{t}, \boldsymbol{v})$ as input, where $\boldsymbol{c} = \{\boldsymbol{c}_1, \dots \boldsymbol{c}_T\}$ are the masked input words after tokenization and we set the mask probability $= 15\%$. The masked tokens are replaced with the special token `[MASK]`. The prediction of masked token at the position t is conditioned on the preceding tokens ($\boldsymbol{c}_{<t}$), visual representations ($\boldsymbol{v}$) and the concept tokens ($\boldsymbol{t}$). We train our model parameters θ by minimizing the negative log-likelihood over the masked tokens:

$$\mathcal{L}_{cap} = -\mathbb{E}_{\boldsymbol{T} \sim D} \Big[\log \prod_{\hat{\boldsymbol{c}}_t \sim \boldsymbol{C}_M} P_\theta(\hat{\boldsymbol{c}}_t | \boldsymbol{c}_{<t}, \boldsymbol{t}, \boldsymbol{v}) \Big], \tag{5}$$

where $\boldsymbol{C}_M$ refers to the ground-truth set of the masked tokens.

Recent works [18, 45] reveal that by leveraging the knowledge distillation technique [24], the VL model can be improved compared to the non-distilled counterpart using a pre-trained Teacher VL model. In our training, we experiment with applying a trained detector-based captioning model as the Teacher (parameterized by θ_t), *i.e.*, VinVL [79], to assist the training of ViTCAP. Note that the Teacher model is a two-stage VL model adopting regional features and object tags from the detector, yielding discrepant visual features with ViTCAP, and hence the distillation objectives like attention-map loss and hidden-states loss are not directly applicable as in [18]. We adopt the classification distillation loss over the masked token probabilities between the predictions from the Student (P_θ) and Teacher (P_{θ_t}) models:

$$\mathcal{L}_{dis} = \mathbb{E}_{\boldsymbol{T} \sim D} \Big[\sum_{\hat{\boldsymbol{c}}_t \sim \boldsymbol{C}_M} \mathrm{KL}\Big(P_\theta(\hat{\boldsymbol{c}}_t), P_{\theta_t}(\hat{\boldsymbol{c}}_t) \Big) \Big], \tag{6}$$

where KL(,) is the Kullback–Leibler divergence. Overall, our final loss is then the combination of the terms:

$$\mathcal{L} = \mathcal{L}_{vc} + \mathcal{L}_{cap} + \mathcal{L}_{dis}. \tag{7}$$

4. Experiment

We now introduce the implementation details of ViTCAP and empirically verify the validity of our proposed training schema from different aspects. To highlight the generalizability of ViTCAP, we benchmark performances of ViTCAP and compare it with prior arts on multiple image captioning testbeds. We then exhaustively study the effect of our proposed concept tokens, the benefits of pre-training at scale, the effect of VL distillation, *etc*. In the end, we visualize the attention maps of ViTCAP and provide in-depth discussion.

4.1. Datasets

Pre-training Datasets. In our experiment, we aggregate image-text pairs from Google-CC [61], SBU Caption dataset [51], MS COCO [43] and Visual Genome dataset [34] to form the pre-training corpus. In total, our pre-training corpus contains 9.9M image-text pairs and 4.1M independent images, and we follow [47] to de-duplicate testing images exist in evaluating datasets. Details of the pre-training corpus can be found in the Appendix.

Evaluation Datasets. We report performances of ViTCAP on COCO captions (Karpathy split) [43], Google-CC [61], and nocaps [1] datasets. We follow Karpathy's split and use 113k, 5k and 5k images for training, validation and testing respectively on MS COCO dataset. As regards to Google-CC, we follow [61] and use its training split containing 3M image-text pairs for training, and report the performances on validation split with 16K image-text pairs. To test the generalization of ViTCAP, we also report the performances on nocaps dataset [1], a benchmark consisting of 166k human-generated captions describing 15k images in the wild collected from the OpenImages dataset [60].

4.2. Implementation Details

Architecture. Our ViTCAP is based on a Vision Transformer base (ViT/b) architecture consisting of $M = 12$ consecutive transformer blocks, with hidden size as 768, and 12 attention heads. In our experiment, we set the patch size as 16×16 and resize the shorter side of the image to 384. We use $M_1 = 8$ transformer blocks in Stem Image Encoder to extract the intermediate grid representations and use $M_2 = 4$ transformer blocks for the CTN branch. When enlarging the size of CTN and Feature Extractor to $M_2 = 12$ transformer blocks, it is equivalent to two independent networks for the computation of Concept Token/Embedding and Grid Feature respectively. We adopt this design with more learnable parameters in our ViTCAP with large scale pre-training (see ViTCAP* in Table 1). Data augmentations are applied on raw images before the linear projection as [13] including *ColorJitter*, *horizontal flipping*, *etc*.

Two-stage Training. Training both the CTN branch jointly with the captioning task jointly from scratch is challenging, we observe that using a pre-trained CTN with stable and consistent concept prediction throughout the training leads to superior captioning results. Thus in practice, we first conduct concept classification training for a good concept prediction, and then train the model with both tasks. Such strategy prevents the "*cold-start*" issue when the initially produced concepts are mostly random, impairing the captioning training. During the joint captioning & concept branch training, we reduce the learning rate for both the Stem Image Encoder and CTN branch by a factor of α ($\alpha = 10$) and keep the predicted concepts relatively consistent but still slowly adapted throughout the training.

- **Concept Classification.** The concept classification is conducted on an aggregated dataset with 4.1M images (see later section for details). To obtain the pseudo ground-truth concepts, we experiment with using the NLTK [46] toolkit to parse out the nouns and adjectives as the target concepts, or simply use all tokens in captions as targets for the classification task. For the detector-produced tags, we take advantage of a ResNeXt-152 C4 architecture based object-attribute detector that has been well-trained [79] to produce image tags as pseudo-labels for concept classification training. We only retain image tags with confidence score > 0.2 from the detector and acquire 50 tags at most per image. For classification training, the model is initialized from the ImageNet-21k [35] pre-trained checkpoint[1], and is optimized for 10 epochs using AdamW [56, 78] optimizer. The batch size is $1,024$. The initial learning rate is $5e-5$ and is linearly decayed to 0.

- **Captioning Training.** For the joint optimization, we apply the well-trained model after concept classification to initialize Stem Image Encoder, CTN and the feature extractor. The initial weights in the feature extractor are copied from the CTN branch, as the architecture for grid feature extractor is the same as the CTN branch. We set base learning rate $lr = 1e-4$, batch-size $= 512$ and train the model for 30 epochs using AdamW optimizer, and set weight decay$= 0.05$.

Evaluation. We evaluate the quality of the generated captions using the prevailing metrics including BLEU@4 [54], METEOR [5], CIDEr [65], ROUGE [41] and SPICE [3]. During inference, we use beam search (beam size $= 1$) for decoding. There exist many evaluating metrics studying the qualities of the generated captions, including SelfCIDEr [69], SMURF [19] and from different aspects [23, 30, 70]. In the Appendix, we conduct more studies studying the diction quality of our generations using SMURF [19] metric.

4.3. Main Results

We perform extensive comparisons of ViTCAP with the prior arts. Table 1 presents the captioning results on MS COCO dataset where the models are trained with cross-entropy loss or optimized with CIDEr as reward [58]. We compare ViTCAP with 1). "*detector w/o VLP*" models with complex architectural modifications. These models [11, 27, 53, 81] all come unanimously with heavy computational burdens and extra learnable parameters. 2). "*detector w. VLP*": prevailing detector-based VL models pre-trained with a large VL corpus and then fine-tuned on image captioning tasks. 3). "*detector-free*" methods: the end-to-end trainable

[1] https://github.com/lucidrains/vit-pytorch.

Methods	V. ENC.	# *I-T*	Cross-Entropy Loss					CIDEr Optimization				
			B@4	M	R	C	S	B@4	M	R	C	S
Detector w.o. VLP												
RFNet [31]	Ensemble	✗	35.8	27.4	56.5	112.5	20.5	36.5	27.7	57.3	121.9	21.2
BUTD [4]	F-RCNN$_{101}$	✗	36.2	27.0	56.4	113.5	20.3	36.3	27.7	56.9	120.1	21.4
LBPF [76]	F-RCNN$_{101}$	✗	37.4	28.1	57.5	116.4	21.2	38.3	28.5	58.4	127.6	22.0
SGAE [75]	F-RCNN$_{101}$	✗	36.9	27.7	57.2	116.7	20.9	38.4	28.4	58.6	127.8	22.1
AoANet [27]	F-RCNN$_{101}$	✗	37.2	28.4	57.5	119.8	21.3	38.9	29.2	58.8	129.8	22.4
M^2 Transfm. [11]	F-RCNN$_{101}$	✗	-	-	-	-	-	39.1	29.2	58.6	131.2	22.6
X-LAN [53]	F-RCNN$_{101}$	✗	38.2	28.8	58.0	122.0	21.9	39.5	29.5	59.2	132.0	23.4
RSTNet [81]	RESNeXt$_{152}$	✗	-	-	-	-	-	40.1	29.8	59.5	135.6	23.3
Detector-Free w.o. VLP												
ViTCAP (Ours)	ViT$_b$	✗	35.7	28.8	57.6	121.8	22.1	40.1	29.4	59.4	133.1	23.0
Detector w. VLP												
UVLP [82]	F-RCNN$_{101}$	4M	36.5	28.4	-	116.9	21.2	39.5	29.3	-	129.3	23.2
MiniVLM [67]	Eff-DET	14M	35.6	28.6	-	119.8	21.6	39.2	29.7	-	131.7	23.5
DistillVLM [18]	Eff-DET	7M	35.6	28.7	-	120.8	22.1	-	-	-	-	-
OSCAR$_b$ [38]	F-RCNN$_{101}$	7M	36.5	30.3	-	123.7	23.1	40.5	29.7	-	137.6	22.8
UNIMO$_b$ [37]	F-RCNN$_{101}$	9M	38.8	-	-	124.4	-	-	-	-	-	-
VL-T5 [10]	F-RCNN$_{101}$	9M	-	-	-	116.5	-	-	-	-	-	-
VinVL$_b$ [79]	RESNeXt$_{152}$	9M	38.2	30.3	-	129.3	23.6	40.9	30.9	-	140.4	25.1
Detector-Free w. VLP												
ViLT-CAP ♠	ViT$_b$	10M	33.7	27.7	56.1	113.5	20.9	-	-	-	-	-
E2E-VLP [72]	ResNet$_{50}$	6M	36.2	-	-	117.3	-	-	-	-	-	-
ViTCAP* (Ours)	ViT$_b$	10M	**36.3**	**29.3**	**58.1**	**125.2**	**22.6**	**41.2**	**30.1**	**60.1**	**138.1**	**24.1**

Table 1. Performance comparisons on COCO-caption Karpathy split [32], where B@4, M, R, C denote BLEU@4, METEOR, ROUGE-L, CIDEr and SPICE scores. All values are reported as percentages (%). We compare the ViTCAP with previous state-of-the-art detector-based baselines (without the VLP) in the first section, and detector-based baselines (with large scale pre-training) in the third section, and the detector-free methods with pre-training in the last section. V. ENC. denotes visual encoders for feature extraction; # *I-T* refers to the number of image-text pairs used in pre-training (in millions). ViTCAP* is a larger version of ViTCAP with more parameters. ♠ is the results we achieved using the ViLT [33] pre-trained checkpoint for image captioning task (see Appendix for more explanation).

image captioning models without object detector (with or without pre-training).

Without VLP. To compare fairly with the detector-based baselines without VLP, we adopt the VinVL tags as concept sources instead of the captions to guarantee that *no additional captions have been exploited* during the concept classification training. Note that the knowledge distillation objective is not applied for this experiment as it introduces extra knowledge from the pre-training of Teacher model. On COCO-caption Karpathy split, our ViTCAP achieves similar results and even surpasses most existing detector-based methods, *i.e.*, CIDEr score 121.8, using caption extracted concepts. It is worth mentioning that the architectures of most existing detector-based methods are deliberately designed, *e.g.*, the self-attention module in X-LAN [53] has 2^{nd} interactions for multi-modal inputs, M^2 Transformer [11] has the multi-level representation of the relationships between image regions, *etc*. ViTCAP adopts the simplest vanilla transformer architecture without any bells and whistles. This proves the effectiveness of our proposed learning paradigm. The ablations in the later section comprehensively explore the benefits of CTN and the knowledge distillation technique.

With VLP. We observe a clear performance gain of ViTCAP after the large scale pre-training (3.0 higher CIDEr scores), better than most detector-based VL methods: *e.g.*, 125.2 *vs*. 123.7 (OSCAR$_b$), and 0.8 higher than UNIMO$_b$, 8.7 higher than VL-T5 when pre-trained on similar VL corpus. This conclusion is further supported by results of other metrics. ViTCAP approaches the state of the art, only 2.3 lower than VinVL in CIDEr scores after CIDEr optimization, considering the fact that VinVL used ResNeXt$_{152}$-based object detector. Compared with detector-free baselines, ViTCAP outperforms all existing works with an obvious discrepancy: 11.7 CIDEr scores higher than the ViLT-CAP [33] and 7.9 higher than E2E-VLP [72].

4.4. Ablative Study

We now comprehensively study ViTCAP's performance gain from different aspects, *i.e.*, knowledge distillation, the effect of concept tokens, and large-scale pre-training.

Concept Source	COCO Captioning				
	B@4	M	R	C	S
✗	33.9	27.8	56.4	114.8	21.3
BUTD [4]	35.0	28.2	56.9	117.4	21.3
VinVL [79]	35.6	28.6	57.4	119.7	21.8
CAPTION	35.6	28.7	57.6	120.9	21.8
VinVL → CAP.♠	35.9	28.6	57.6	121.3	21.9
CAPTION♠	35.7	28.8	57.6	121.8	22.1

Table 2. Adopting various sources of semantic concept leads to different performances. "CAPTION" represents the baseline extracting keywords from open-form captions; "♠" is the baseline using all words in captions as target concepts; "BUTD" and "VinVL" represent using the object tags produced by the object detector from [4] and [79] as target semantic concepts, respectively. "VinVL → CAP." represents adopting detector tags [79] during first stage of concept classification and using caption extracted tags during the second stage.

Methods	Cross-Entropy Loss				
	B@4	M	R	C	S
ViT/B	33.9	27.8	56.4	114.8	21.3
ViT/B$_{+\text{KD}}$	35.4	28.5	57.5	120.0	21.7
ViT/B$_{+\text{CTN-TAG}}$	35.2	28.0	57.0	117.1	21.4
ViT/B$_{+\text{OD-TAG}}$	34.3	28.2	57.4	117.4	21.7
ViTCAP$_{+\text{CTN-TOK}}$	35.7	28.8	57.6	121.8	22.1
ViTCAP$_{+\text{CTN-TOK+PRE+KD}}$	36.3	29.3	58.1	125.2	22.6

Table 3. Comparisons of ViTCAP with or without knowledge distillation, large-scale pre-training and with CTN. Performances are reported on COCO-caption Karpathy split optimized by cross-entropy loss. $_{+\text{OD-TAG}}$ indicates the result using the detector produced off-the-shelf tags as [38]. $_{+\text{CTN-TOK}}$ is the result of ViTCAP using the initialization after first-stage concept classification. KD and PRE are results obtained with masked token classification distillation and pre-training at scale respectively.

Semantic Concept Sources. We study the effects of different semantic concept sources, *i.e.*, from object detectors [4, 79], captions-extacted concepts, and the combination of them. Table 2 lists the performances of ViTCAP on the COCO caption dataset with various semantic concepts sources. Open-form captions are the most accessible source to directly obtain semantic concepts, although these descriptions can sometimes be noisy, inaccurate and incomplete. "CAPTION" in Table 2 is the result using nouns and adjectives parsed from captions using NLTK [46] toolkit as target concepts. This leads to an obvious improvement over the baseline (without CTN): CIDEr 120.9 *vs*. 114.8. We also attempt to leverage all tokens from the captions as concept targets in case of omitting essential words during parsing (see "CAP.♠"), which brings further incremental improvement and yield best result. Although using all tokens in the caption might inevitably introduce more noisy or irrelevant words, *e.g.*, connection and stop words, it also broadens the semantic concepts vocabulary as some rare entities/attributes might be missed using just keywords.

We then experiment with using the detectors in [79] and [4] to produce image-level tags as target concepts. We observe that using the detector of VinVL yields better performances than BUTD, *i.e.*, 119.7 *vs*. 117.4 CIDEr scores. This is mainly because of the more diverse collection of semantic concepts involved in [79] than BUTD [4]. The second last row is the experiment where the model is firstly trained using VinVL tags on large scale dataset (in the first stage), and then using the caption tokens during the second stage of captioning. This indicates that, when no captions are attainable, it is also viable to leverage detector-produced tags to improve the performance.

Effect of Different Modules. In Table 3, we show in details the independent performance gains from each design, *viz.*, with or without concept tokens, masked token distillation loss, pre-training and the combinations of them. We report the result of the baseline model which reaches CIDEr scores 114.8 on COCO caption dataset. With the aim of isolating the performance gain from concept tokens, we first decode the image-level semantic concepts and store them as offline tags for the captioning task. We then follow [38] to tokenize them and concatenate the tag embedding with visual features for captioning task. This allows us to directly compare the effect of CTN-produced concepts with detector tags without the concept classification initialization. Adopting the explicit tags predicted by the CTN leads to obvious improvements: 2.3 higher CIDEr and 1.3 higher BLEU@4 scores, reaching similar results with that using VinVL's detector tags directly (see ViT/B$_{+\text{OD-TAG}}$): 117.4 *vs*. 117.1 CIDEr scores. This proves that our generated semantic concepts play a significant role in the captioning task and have a similar effect as the VinVL's detector tags. Next, we apply the pre-trained weights after the concept classification to initialize the ViTCAP for the captioning task, and find further improvement (see ViTCAP$_{+\text{CTN-TOK}}$). This proves that both the predicted concept tokens and the concept classification training are beneficial for captioning tasks. For the knowledge distillation experiment, we use the VinVL-base [79] optimized on COCO-caption dataset as the Teacher and keep it frozen during distillation. The application of KD on masked token prediction (ViT/B$_{+\text{KD}}$) is also evidently helpful: there is an over 5.0 CIDEr scores improvement over the baseline. Note that the KD objective is only applied in the downstream for the ViTCAP baseline after VLP for fair comparison with previous works. Finally, by pre-training the ViTCAP with large scale VL corpus continuously contributes to the results.

Performances on other Benchmarks. To evaluate the generalizability of ViTCAP, we continue to expand the testbeds to other challenging captioning benchmarks, *i.e.*, Google-CC [61] and nocaps [1] datasets. For the Google-CC dataset,

Methods	CC-3M dev CIDEr
FRCNN [8]	89.2
Ultra [8]	93.7
ViLT-CAP [33]♠	83.8
VinVL [79]♠	103.4
CC-3M [9]	100.9
CC-12M [9]	105.4
ViTCAP	**108.6** +3.2

Table 4. Performances of ViTCAP model on Conceptual Captions (Google-CC 3M dev-split) [61] benchmark. We compare with the baseline methods FRCNN [8], Ultra [8] and [9]. The ViLT-CAP♠ and VinVL represent our reproduced results with pre-trained checkpoint from [33] and [79].

Methods	nocaps validation set							
	in-domain		near-domain		out-of-domain		overall	
	C	S	C	S	C	S	C	S
Human	84.4	**14.3**	85.0	**14.3**	**95.7**	**14.0**	87.1	**14.2**
UpDown [1]	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	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
OSCAR [38]	79.6	12.3	66.1	11.5	45.3	9.7	63.8	11.2
OSCAR + CBS	83.4	12.0	81.6	12.0	77.6	10.6	81.1	11.7
VIVO [25]	90.4	13.0	84.9	12.5	83.0	10.7	85.3	12.2
VIVO + CBS	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4
ViTCAP	**99.3**	13.2	90.4	12.9	78.1	11.9	89.2	12.7
ViTCAP + CBS	98.7	13.3	**92.3**	13.3	95.4	12.7	**93.8**	13.0
Δ	+6.5	+0.4	+4.5	+0.7	+7.9	+1.2	+5.5	+0.6

Table 5. Performances of ViTCAP in nocaps validation split. We compare our ViTCAP with previous state-of-the-art models at **"in-domain"**, **"near-domain"** and **"out-of-domain"**. Results are reported with constrained beam search (CBS) decoding [2].

we train the ViTCAP on the training split, which consists of ∼3.3M image-caption pairs, and test it on the dev split. We follow the same training protocols as previously mentioned and optimize the ViTCAP for 120 epochs. Following previous works, we evaluate the performances using the CIDEr metric and Table 4 shows the results of ViTCAP compared with previous captioning models. In particular, ViTCAP achieves the state-of-the-art results CIDEr 108.6 scores (without the knowledge distillation), surpassing all detector-based captioning models. CC-12M is the model trained with 12M image-caption pairs [9]. Again, when evaluating on nocaps dataset, ViTCAP shows promissing results across all in-domain, near-domain, and out-of-domain splits. For example, ViTCAP achieves 98.7 and 93.8 CIDEr scores on in/out-domain splits, 6.5 and 5.5 higher than the VIVO [25], which exploits OpenImage [36] dataset to learn semantic concepts for captioning task. The great generalization ability of ViTCAP can be partly ascribed to its ability to recognize expansive semantic concepts extracted from the open-form captions. Compared to predicting the pre-defined tags as in the detector, the usage of caption extracted concepts largely expands the concept vocabulary. This provides the ViTCAP with robust and broad concept tokens, which is essential for the images with novel concepts.

Qualitative Examples. We show visualization examples of the attention maps from ViTCAP in Figure 3 together with their generated concepts&captions. Interestingly, we observe obvious correlations between the attended regions across different layers and predicted concepts. For example, "dog" is notably highlighted according to the mean-averaged attention maps, yet the "man" is more attended in shallower transformer blocks. We conjecture that instead of relying on an object detector to glean object locations, training the detector-free VL model properly via image-text supervisions might potentially lead to a strong grounding model.

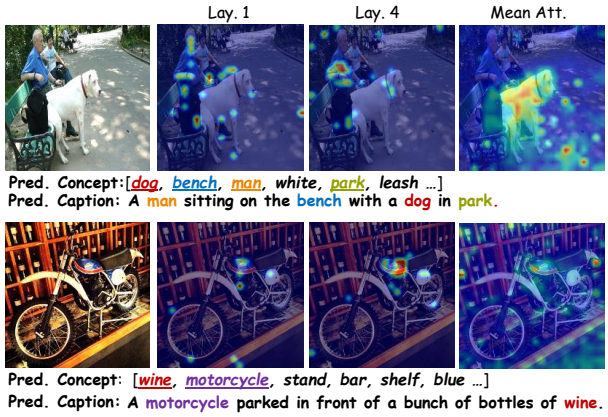


Figure 3. Visualization of the attention maps from ViTCAP and its produced concepts&captions. "____" refers to the concepts appear in captions. Best viewed in color.

5. Conclusion

In this paper, we propose the ViTCAP, a detector-free image captioning model in the full transformer architecture fashion. Compared with existing captioning models, ViTCAP can be trained in an end-to-end fashion without intermediate regional operations using grid representations. Our proposed Concept Token Network learns broad semantic concepts and encodes them as the concept tokens that largely benefit the captioning task on a series of challenging captioning benchmarks. Extensive experiments indicate that ViTCAP achieves competing performances, approaching most detector-based models. We anticipate that ViTCAP will lead to more future works in building efficient Vision and Language models.

Acknowledgement. This work was supported by the National Science Foundation under Grant CMMI-1925403, IIS-2132724 and IIS-1750082.

Supplementary Materials

In this supplementary materials, we provide additional details about experimental settings, and then further compare effect of different semantic concept sources, more ablative studies regards training, different architectural instantiations, and further showcase more qualitative examples of predicted semantic concepts.

Source	VG [34]	COCO [43]	CC [9]	SBU [51]
Image	108K	113K	3.1M	875K
Text	5.4M	567K	3.1M	875K

Table 6. Statistics of the VL pre-training datasets.

6. Pre-training VL Corpus

As previous works in [79], we carry out the pre-training of ViTCAP on the aggregation of several common datasets, which include COCO [43], Conceptual Caption [9], SBU Captions [51], and Visual Genome [34]. We have the detailed statistics of the aggregated datasets in Table 6. In total, we use 4.2 millions of images and 9.9M captions for the pre-training. Following [47], we de-duplicate images that exist in both pre-training corpus and COCO Karpathy testing splits for fair comparisons.

7. Ablative Studies

This section further presents additional ablative studies about ViTCAP, which includes: some examples and basic statistics about semantic concepts, the effect of different concept sources, results of different concept classification losses, different other training strategies.

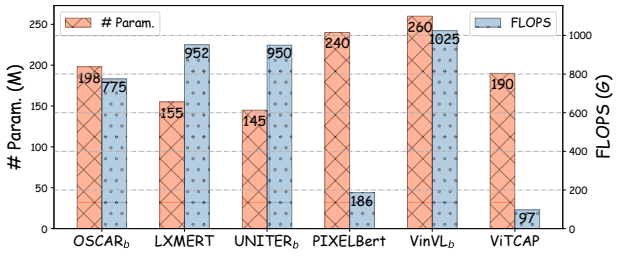


Figure 4. Inference speed in FLOPs (in G), number of parameters (in M) of multiple VL models and ViTCAP.

Examples and Stats of Concepts. In practice, we experiment with utilizing semantic concepts gleaned from 1). open-form image captions by language parsing (or simple as using all tokens as classification ground-truth) or 2). an object detector.

As previously mentioned, we notice that the concepts from both sides are all severely long-tailed distributed (an example of the detector-produced concept distribution is shown in Figure 5). Notably, certain concepts appear more frequently across the whole COCO training split, *e.g.*, "`person`", "`tree`", "`window`" obviously exist far more frequent than the remaining. We also resort to different object detectors to acquire high-quality semantic concepts, *i.e.*, a ResNet$_{101}$ base Faster-RCNN [4] that has been pre-trained on Visual-Genome dataset [34] (denoted as BUTD), and a ResNext$_{152}$ based modified Faster-RCNN detector with broader categories of the visual attribute as detection targets (denoted as VinVL). These detector-produced image-level tags are actually accurate with less noise than in captions, but they also require a pre-defined categorical dictionary with a fixed set of concepts. This largely limits the scope of their applications.

In Figure 4, we present the inference speed and the number of learnable parameters of prevailing detector-based VL models compared with ViTCAP Notably, with on-par parameters, ViTCAP consumes only $\sim$ 10% FLOPs of the prevailing VL models (97G for ViTCAP *vs.* 1,025G for VinVL).

More About Concept Sources. Open-form captions are the most ideal source to obtain semantic concepts as they naturally carry abundant semantic concepts with no vocabulary limitation. Notwithstanding that most of these descriptions can be noisy, inaccurate, and incomplete. In practice, we leverage different ways to extract the concepts from them by 1) using the NLTK [46] toolkit and parsing out only the nouns and adjectives as the semantic concepts for the classification task (see "CAPTION" baseline in main paper); 2) we also simply attempt to leverage all tokens from the captions as concept targets in case of omitting essential words during parsing (see "♠" in main paper). We first extract these tags as "*off-the-shelf*" annotations for the concept classification task and then apply the initialization of ViTCAP after the first stage of training for the joint captioning training. Note that we conduct and compare all these ablations without VL pre-training. It is beneficial to further adopt the concept classification loss during the joint training, as the semantic concepts in the COCO-caption dataset vary with the concept classification dataset. Also, captions in these two domains might vary from the aspect of textual styles: for example, length of captions, the use of synonyms, cognate and conjugate words, or various tenses.

Concept Classification Training. We now study the effect of different losses for the concept classification task, namely binary cross-entropy loss and focal loss, and the effect of the initialization after the classification training. The extremely imbalanced sample distribution usually leads to sub-optimal classification performances, as also studied in previous works like face recognition [49, 80] and object de-

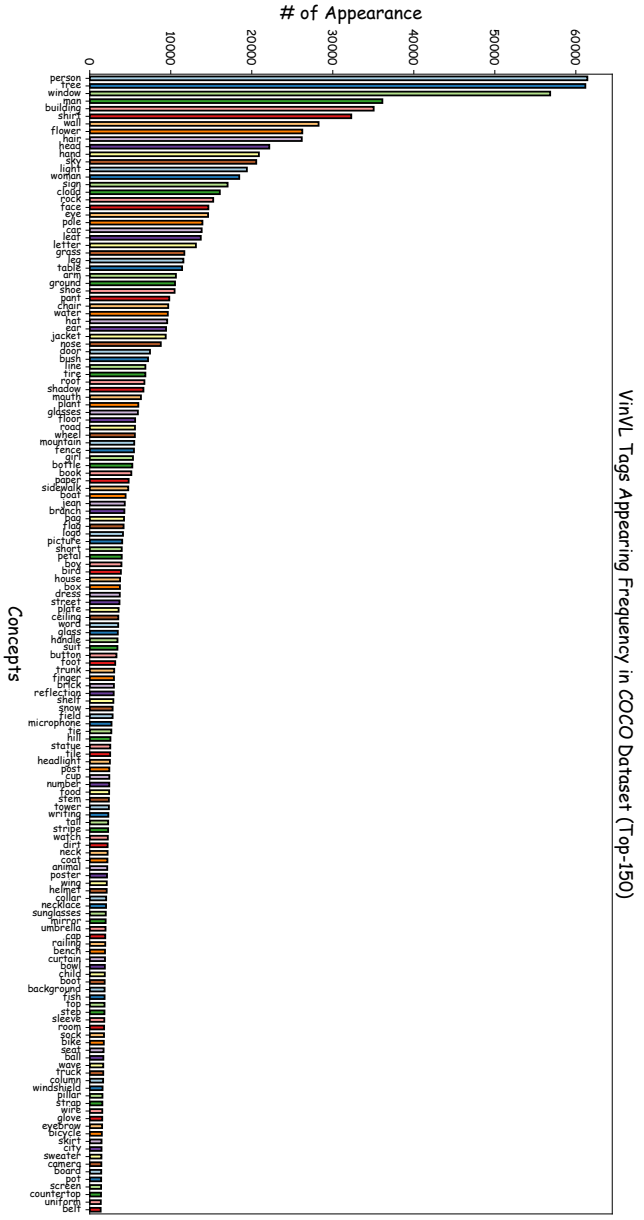


Figure 5. Top-150 most frequently appeared semantic concepts produced by VinVL's object detector. The produced tags are severely long-tail distributed and certain concepts dominates across all samples. This arises the necessity to apply focal loss as countermeasure.

tection [40, 52], etc. As countermeasures, there exist works designing advanced losses [42, 80] re-weighting different samples. In Table 7, we list the performances of ViTCAP using different losses. In specific, the top-two rows are the baseline results 1). Baseline: vanilla Encoder-Decoder architecture without CTN branch, and 2). Encoder-Decoder architecture using VinVL's OD tags as [38]. "$_{\text{Tag}}$" denotes the results are reported using concepts as the offline tags

	COCO Captioning					
	EPOCH	B@4	M	R	C	S
Baseline	-	33.9	27.8	56.4	114.8	21.3
VinVL-Tag	-	35.4	28.1	57.2	117.7	21.3
BCE_{Tag}	10	33.9	27.9	56.5	115.0	21.4
$\text{FOCAL}_{\text{Tag}}$	10	35.2	28.0	57.0	117.1	21.4
$\text{FOCAL}_{\text{Tag+Init}}$	10	36.0	28.4	57.5	120.5	22.0
$\text{FOCAL}_{\text{Init}}$	10	35.0	28.2	57.1	118.0	21.6
$\text{FOCAL}_{\text{Tag+Init}}$	40	35.9	28.4	57.6	121.1	22.1

Table 7. Performances of ViTCAP using focal loss, binary classification loss as concept classification training target.

without concept classification & its initialization. We observe that by applying the BCE loss trained offline concepts as offline tags, the results are only incrementally improved over the baseline, and it still shows a great performance gap *w.r.t.* the VinVL's tag. Notably, using focal loss obviously improves the quality of produced concepts, reaching 117.1 CIDEr scores. To this end, we apply the concept classification pre-trained initialization, and this further improves the performances to a great extent. It is discernible that the experiment "$_{\text{Init}}$" gives worse result than the "$_{\text{Tag+Init}}$". This validates that both the concept classification task and the predicted concepts are helpful for the captioning task. Results show that they are complementary to each other.

Tokenization	COCO Captioning				
	B@4	M	R	C	S
Caption Tokenizer	35.5	28.5	57.5	119.7	21.8
Classifier Tokenizer	35.6	28.4	57.4	119.8	21.8
Independent Tokenizer	35.9	28.5	57.6	120.1	21.9

Table 8. Performances of ViTAP using different strategies for concept tokenization.

Representing Concepts as Tokens. There are multiple ways to encode the predicted concepts as continuous embedding for the decoding stage. We study three different ways of encoding and present the results in Table 8, namely, 1). use the tokenizer for captioning, 2). use the concept classifier's tokenizer (in concept classification, we simply use the BERT tokenizer to encode the semantic concepts), 3). use an independent and untrained tokenizer. Though in practice, all three tokenizers are implemented based on the BERT tokenizer [12], the embeddings from the three are entirely different. From the results, we observe a fairly negligible performance gap: using an independent tokenizer only yields a 0.4 higher CIDEr score. Though adopting an independent tokenizer yield the best result, it introduces additional parameters and thus we choose to share the tokenizer for captioning instead.

	COCO Captioning				
	B@4	M	R	C	S
GT Concepts	35.5	28.4	57.3	119.1	21.7
GT + PRED. Concepts	35.2	28.5	57.3	119.2	21.8
PRED. Concepts	36.1	28.6	57.6	120.6	21.7

Table 9. Performances of VitCAP using either ground-truth concepts for captioning, the concept network predicted concept tokens or the mixture of them during training.

We experiment with different ways to train with the concept tokens. In Table 9, we list the results of training using GT semantic concepts encoded as tokens, GT concepts mixed with predicted concepts, and fully predicted concepts.

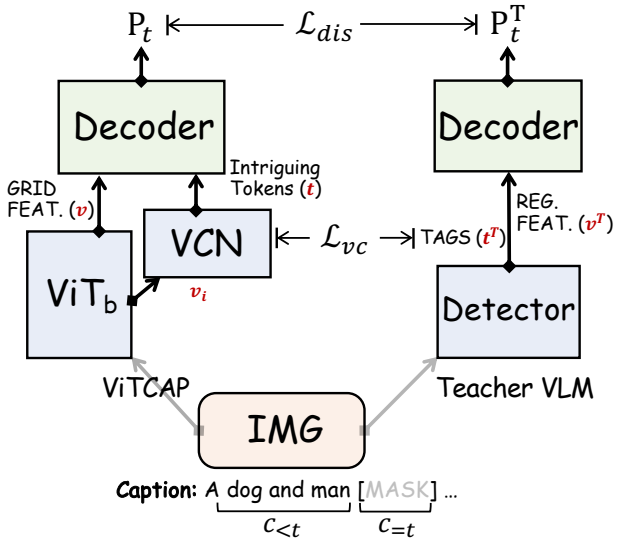


Figure 6. The overall training paradigm of ViTCAP can be understood as the knowledge distillation procedure where a detector-based Teacher VLM to assist the training of ViTCAP as a knowledge distillation paradigm. The CTN branch in ViTCAP learns to predict the semantic concepts as conceptual tokens for captioning.

We find that by using the predicted concepts for training leads to optimal results. This is mostly because the pre-trained CTN can already produce reasonable concepts at the captioning fine-tuning stage.

ViTCAP Architecture. To give a more detailed explanation of the architecture of ViTCAP: it consists of a stem image encoder with 8 transformer blocks (shared for both grid feature extractor and CTN), a CTN branch with 4 transformer blocks, and a grid feature extractor with 4 transformer blocks, the multi-modal module is also a 4 transformer blocks module. When $M_1 = 12$, the model can be understood as consisting of two parallel branches, with one for concept prediction and one for grid representation. We does find that minimizing the shared blocks can bring extra performance gains

Architecture	COCO Captioning				
	B@4	M	R	C	S
SIN-TOW$_{32\times32}$	32.5	27.1	55.4	109.5	20.2
+EFF. OD-Tags	32.8	27.4	55.5	110.9	20.6
+VinVL-Tags	33.5	27.8	56.1	114.6	21.1
ENC-DEC$_{32\times32}$	33.4	27.5	56.0	112.1	20.6
+EFF. OD-Tags	33.8	27.9	56.4	114.6	21.3
+VinVL-Tags	34.4	27.9	56.6	115.8	21.1
+ViTCAP-Tags	34.0	27.7	56.3	114.2	20.8
SIN-TOW$_{16\times16}$	33.8	27.8	56.2	113.9	21.0
+EFF. OD-Tags	33.8	27.9	56.4	114.6	21.3
+VinVL-Tags	34.3	28.2	56.7	117.4	21.7
ENC-DEC$_{16\times16}$	33.9	27.8	56.4	114.8	21.3
+VinVL-Tags	35.4	28.1	57.2	117.7	21.3
+ViTCAP-Tags	35.2	28.0	57.0	117.1	21.4
ViTCAP	35.7	28.8	57.6	121.8	22.1

Table 10. We compare different instantiations of ViTCAP with architectural variations of ViT based captioning model: single-tower (SIN-TOW), encoder-decoder structure (ENC-DEC), two-tower ViTCAP, and ViTCAP with various numbers of sharing blocks in stem image encoder. All experiments are conducted without VL pre-training and are trained by cross-entropy loss.

but this inevitably increases the model size very obviously. We only adopt this two-tower design in the experiment with large scale pre-training where we follow a two-step training schema as OSCAR [38]: we first leverage the CTN to predict the semantic concepts of all pre-training images; Then, we use these concepts as the off-the-shelf tags (similar as the object detector tags) for the pre-training.

Architectural Variations. We then experiment with different architectural variations of ViTCAP and report their performances on COCO-caption in Table 10. The baseline models include single-tower (SIN-TOW) that shares the ViT backbone for both modalities; Encoder-decoder (ENC-DEC) that use a ViT as visual encoder and 4 separate transformer blocks as modal fusion. This is similar to [33], however, we modify it by using seq-to-seq attention maps for the captioning training which prevents the model from seeing bidirectional context; Two-tower (TWO-TOW) uses an independent ViT/b architecture as a conceptual token network and another architecture as the visual encoder.

More Evaluations. In addition to previous benchmarks, we also use the recently proposed rule-based SMURF metric which demonstrates SOTA correlation with human judgment and improved explainability. SMURF is the first caption evaluation algorithm to incorporate diction quality into its evaluation. We observe that our method preserves both semantic performance and the descriptiveness of terms used in the sentence.

Methods	SMURF
w/ only periods removed	
VinVL	0.66
M^2 Transformer	0.49
X-Transformer	0.51
ViTCAP	0.55
w/ all punctuation removed	
VinVL	0.59
M^2 Transformer	0.42
X-Transformer	0.46
ViTCAP	0.49

Table 11. Performance of ViTCAP comparing with previous models under SMURF [19] metric. Note that this results is evaluated using ViTCAP without pre-training.

8. Discussions

Qualitative Examples. We demonstrate more qualitative examples of the attention maps produced by ViTCAP together with their predicted semantic concepts in Figure 7.

Can ViTCAP Ground Concepts? Interestingly, we observe that the attention maps produced from transformer blocks closely relate to the concepts and various layers have different focuses while the averaged attention maps cover broad holistic regions. We present more visualizations in Figure 8 which contain a single object per image for more direct analysis. The topmost row is a picture with multiple "`wild gooses`" and all regions of them are highlighted according to the attention maps. Despite so, it seems ViTCAP suffers from identifying the clear borders of the object that it may only recognize part of the objects, *e.g.*, ViTCAP only highlights the part of the "`traffic light`" and the "`tie`". This indicates the potential application of ViCAP for weakly supervised textual grounding tasks for the image [16, 17, 59, 68] and video [26, 50].

VL Distillation Schema. Our distillation schema can be indeed viewed as an extension of the VL distillation schema, where the Student model not only mimics the predicted masked token probability but also learns from the Teacher OD's object tags. As is shown in Figure 6. Note that our distillation technique is only applied on the ViTCAP with VL pre-training, as the teacher VL model contains knowledge acquired from large-scale pre-training and so it is unfair to compare the ViTCAP with other methods without VL pre-training.

Detector Tags *vs*. Caption Extracted Concepts. Empirical studies show that the caption extracted concepts lead to better ViTCAP. We conjecture that this is mainly because the captions contain much broader image concepts contained in open-form texts, yet the detector tags are pre-defined with much more limited vocabulary. However, perfectly aligned image-text pairs are not always attainable considering that most existing image-level annotations are collected from the Web. These image captions can be as noisy as alt text or short phrases, from which the extracted concepts only cover part of the image content. Thus in practice, it is also an important aspect to explore the feasibility of adopting the non-caption-extracted concepts, *e.g.*, from an object detector as a substitution. This provides a flexible source of the concepts.

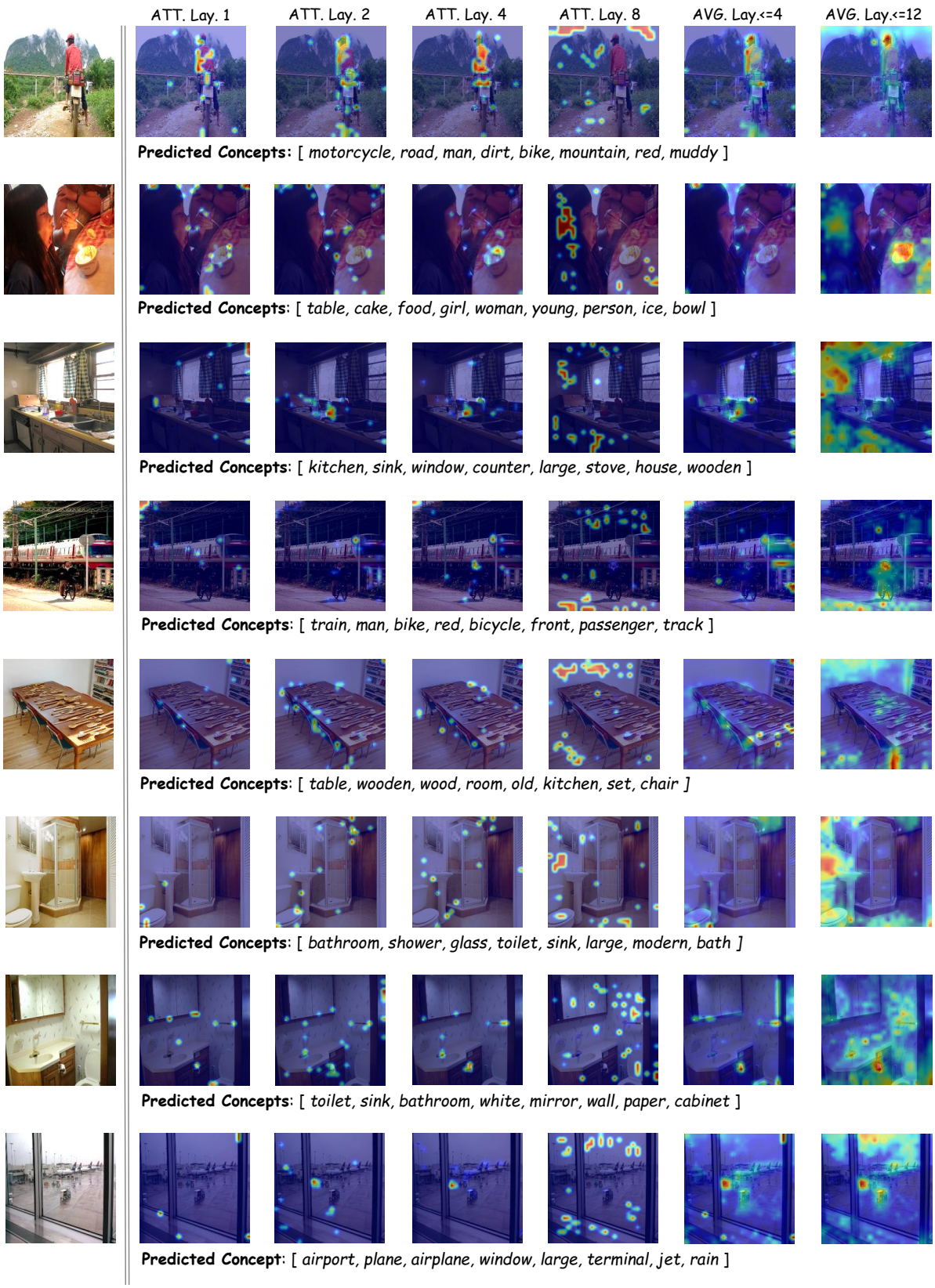
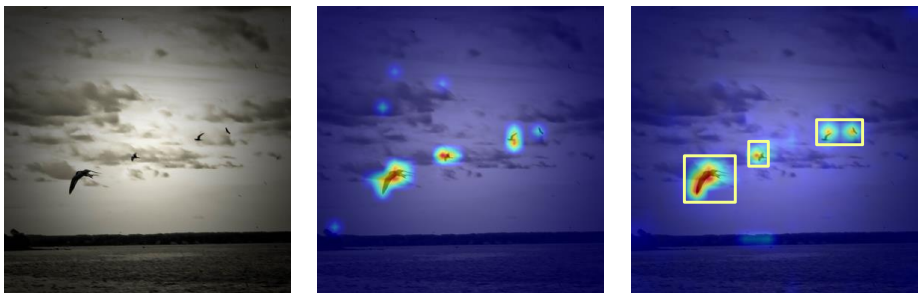
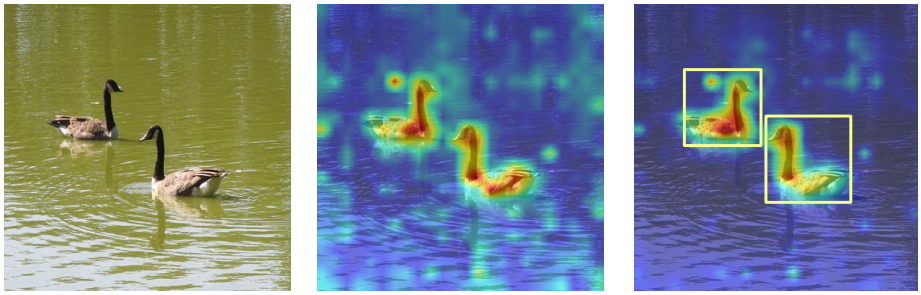


Figure 7. ViTCAP produced class-agnostic attention maps and their associated semantic concepts of random images from COCO caption dataset. We exhibit attention maps of 1, 2, 4, 8th transformer blocks of ViTCAP and the mean-average attention maps of first 4 and the entire 12 transformer blocks (last two columns).



Predicted Concept:[*bird, cloudy, beach, etc*]

Predicted Concept:[*traffic, light, red, pole, etc*]



Predicted Concept:[*water, geese, pond, lake, etc*]

Predicted Concept:[*man, tie, black, shirt, young, etc*]

Figure 8. From left to right, we show the original image, average attention maps of the front 4 and 8 transformer blocks.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*, 2016.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[6] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020.

[7] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[8] Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. *arXiv preprint arXiv:1909.02097*, 2019.

[9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[10] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*, 2021.

[11] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv:2111.02387*, 2021.

[15] Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2commonsense: Generating commonsense descriptions to enrich video captioning. *Conference on Empirical Methods in Natural Language Processing*, 2020.

[16] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6378–6388, 2019.

[17] Zhiyuan Fang, Shu Kong, Tianshu Yu, and Yezhou Yang. Weakly supervised attention learning for textual phrases grounding. *arXiv preprint arXiv:1805.00545*, 2018.

[18] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[19] Joshua Feinglass and Yezhou Yang. Smurf: Semantic and linguistic understanding fusion for caption evaluation via typicality analysis. *arXiv preprint arXiv:2106.01444*, 2021.

[20] Zhe Gan, Yen-Chun Chen, Linjie Li, Tianlong Chen, Yu Cheng, Shuohang Wang, and Jingjing Liu. Playing lottery tickets with vision and language. *arXiv preprint arXiv:2104.11832*, 2021.

[21] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5630–5639, 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[25] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. *arXiv e-prints*, pages arXiv–2009, 2020.

[26] De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding" it": Weakly-supervised reference-aware visual grounding in instructional

videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5948–5957, 2018.

[27] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019.

[28] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[29] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020.

[30] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. Tiger: Text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050*, 2019.

[31] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018.

[32] Karpathy. Karpathy/neuraltalk: Neuraltalk is a python+numpy project for learning multimodal recurrent neural networks that describe images with sentences.

[33] Wonjae Kim, Son Bokyung, Kim Ildoo, and Wonjae Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *International Conference on Machine Learning*, 2021.

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

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[36] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[37] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[38] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[39] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6054–6063, 2019.

[40] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10991–11000, 2020.

[41] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[42] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[44] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[45] Yongfei Liu, Chenfei Wu, Shao-yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*, 2021.

[46] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[47] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[48] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2018.

[49] Yuhao Ma, Meina Kan, Shiguang Shan, and Xilin Chen. Learning deep face representation with long-tail data: An aggregate-and-disperse approach. *Pattern Recognition Letters*, 133:48–54, 2020.

[50] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019.

[51] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.

[52] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 864–873, 2016.

[53] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10971–10980, 2020.

[54] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[55] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[56] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[57] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[58] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[59] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[60] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019.

[61] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[62] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[63] Yao Ting, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2621–2629, 2019.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[65] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[66] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[67] Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Minivlm: A smaller and faster vision-language model. *arXiv preprint arXiv:2012.06946*, 2020.

[68] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14090–14100, 2021.

[69] Qingzhong Wang, Jia Wan, and Antoni B Chan. On diversity in image captioning: Metrics and methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[70] Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14050–14059, 2021.

[71] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[72] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning. *arXiv preprint arXiv:2106.01804*, 2021.

[73] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[74] Ming Yan, Haiyang Xu, Chenliang Li, Bin Bi, Junfeng Tian, Min Gui, and Wei Wang. Grid-vlp: Revisiting grid features for vision-language pre-training. *arXiv preprint arXiv:2108.09479*, 2021.

[75] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.

[76] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.

[77] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[78] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

[79] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. *arXiv preprint arXiv:2103.15358*, 2021.

[80] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017.

[81] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474, 2021.

[82] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.