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 [10]	COCO [14]	CC [2]	SBU [19]
Image	108K	113K	3.1M	875K
Text	5.4M	567K	3.1M	875K

Table 1. Statistics of the VL pre-training datasets.

1. Pre-training VL Corpus

As previous works in [23], we carry out the pre-training of ViTCAP on the aggregation of several common datasets, which include COCO [14], Conceptual Caption [2], SBU Captions [19], and Visual Genome [10]. We have the detailed statistics of the aggregated datasets in Table 1. In total, we use 4.2 millions of images and 9.9M captions for the pre-training. Following [16], we de-duplicate images that exist in both pre-training corpus and COCO Karpathy testing splits for fair comparisons.

2. 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.

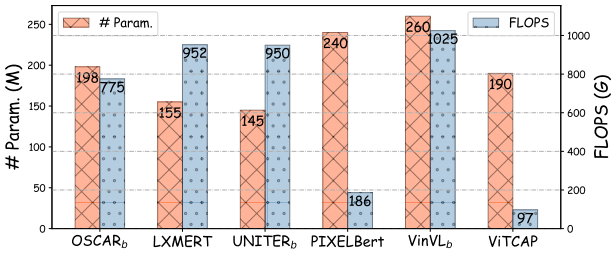


Figure 1. Inference speed in FLOPs (in G), number of parameters (in M) of multiple VL models and ViTCAP.

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 2). 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 [1] that has been pre-trained on Visual-Genome dataset [10] (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 1, 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 [15] 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 [17, 24] and object de-

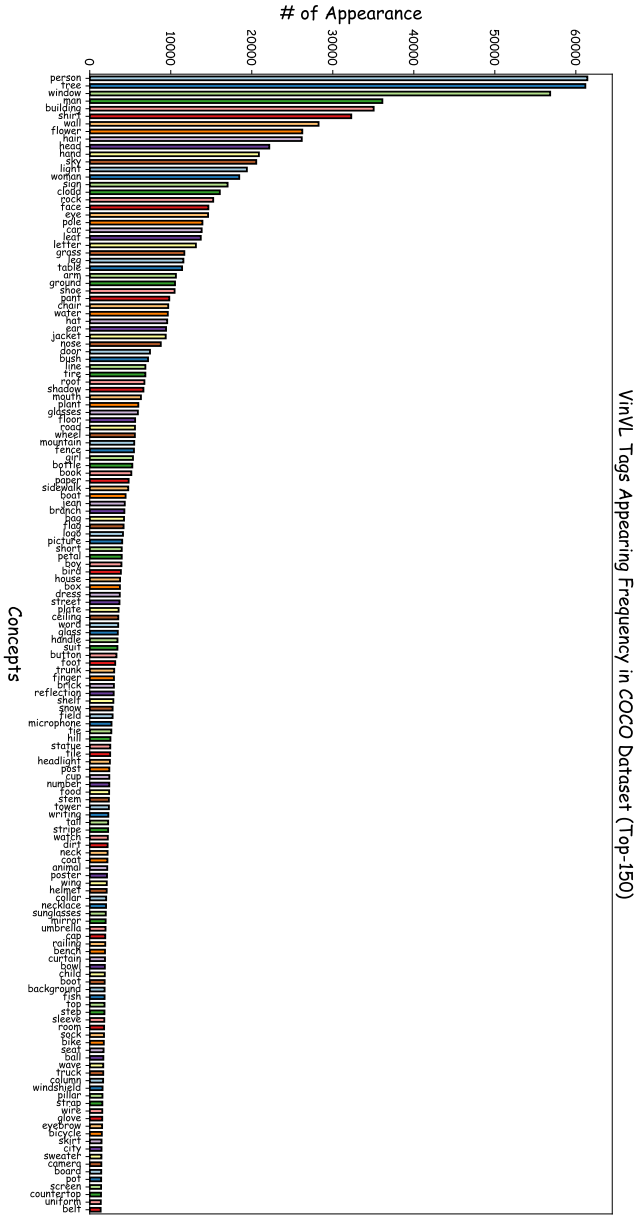


Figure 2. Top-150 most frequently appeared semantic concepts produced by VinVL's object detector. The produced tags are severely long-tail distributed and certain concepts dominates across all samples. This arises the necessity to apply focal loss as countermeasure.

tection [12, 20], etc. As countermeasures, there exist works designing advanced losses [13, 24] re-weighting different samples. In Table 2, 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 [11]. "$_{\text{Tag}}$" 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 2. 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 3. 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 3, 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 [3], 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 4. 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 4, we list the results of training using GT semantic concepts encoded as tokens, GT concepts mixed with predicted concepts, and fully predicted concepts.

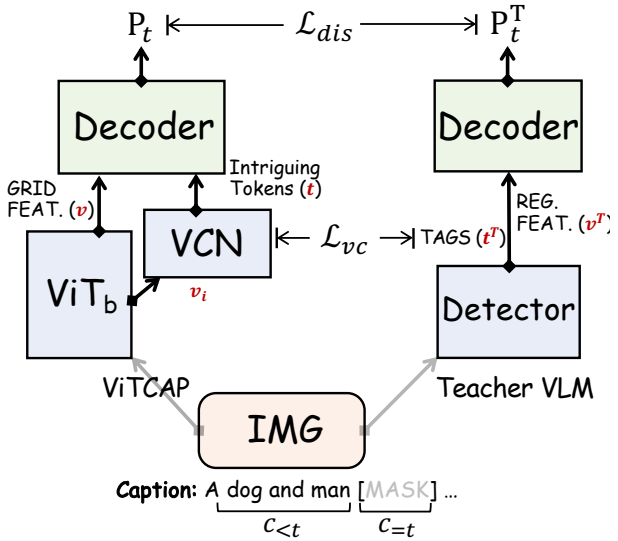


Figure 3. The overall training paradigm of ViTCAP can be understood as the knowledge distillation procedure where a detector-based Teacher VLM to assist the training of ViTCAP as a knowledge distillation paradigm. The CTN branch in ViTCAP learns to predict the semantic concepts as conceptual tokens for 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 5. 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 [11]: 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 5. 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 [9], 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 6. Performance of ViTCAP comparing with previous models under SMURF [7] metric. Note that this results is evaluated using ViTCAP without pre-training.

3. Discussions

Qualitative Examples. We demonstrate more qualitative examples of the attention maps produced by ViTCAP together with their predicted semantic concepts in Figure 4.

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 5 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 [4, 6, 21, 22] and video [5, 8, 18].

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 3. 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.

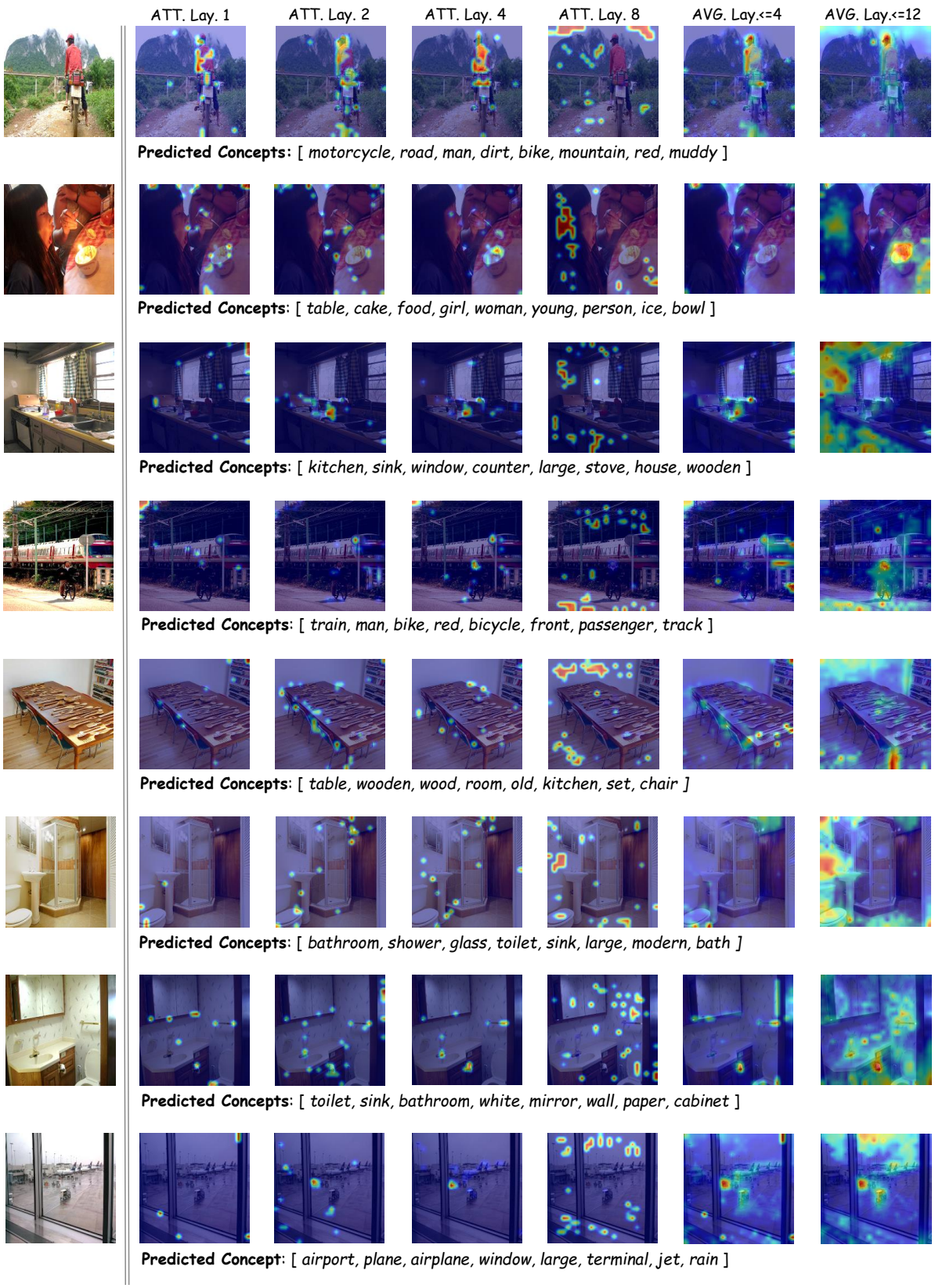
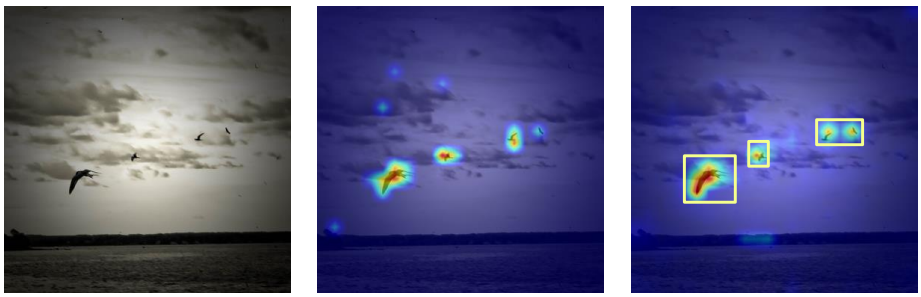
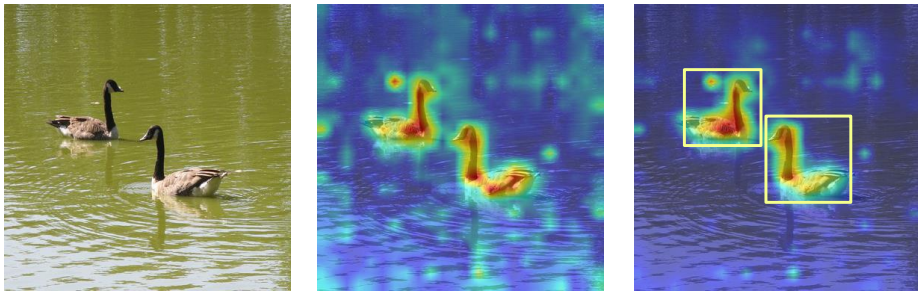


Figure 4. ViTCAP produced class-agnostic attention maps and their associated semantic concepts of random images from COCO caption dataset. We exhibit attention maps of 1, 2, 4, 8th transformer blocks of ViTCAP and the mean-average attention maps of first 4 and the entire 12 transformer blocks (last two 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 5. From left to right, we show the original image, average attention maps of the front 4 and 8 transformer blocks.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] Zhiyuan Fang, Shu Kong, Zhe Wang, Charless Fowlkes, and Yezhou Yang. Weak supervision and referring attention for temporal-textual association learning. *arXiv preprint arXiv:2006.11747*, 2020.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.