Efficient Vision-Language Pretraining with Visual Concepts and Hierarchical Alignment

Mustafa Shukor[1]
mustafa.shukor@sorbonne-universite.fr
Guillaume Couairon[1,2]
gcouairon@fb.com
Matthieu Cord[1,3]
matthieu.cord@sorbonne-universite.fr

[1] Sorbonne University, Paris, France
[2] Meta AI
[3] Valeo.ai, Paris, France

A Appendix

The Appendix is organised as follows; in section A.1, we give more details about the downstream tasks and how we finetune the model on them, then in section A.2, we elaborate on the implementation details used to train ViCHA. We do more ablation study in section A.3. We show the results of training ViCHA on 4M images in section A.4 and additional experiments on data filtering in section A.5. We discuss some aspects of the paper in section A.6 and conclude the appendix in section A.7 by showing some qualitative results on visual grounding, VQA and some limitations of VCs.

A.1 Downstream Tasks

We mainly follow the evaluation setup of ALBEF [19]. During finetuning, the image resolution is increased to 380. We use 4 GPUs for Image-Text retrieval and 2 GPUs for other tasks.

Image-Text Retrieval consists of finding the image that corresponds to a given text or vice-versa. We finetune and evaluate on MSCOCO [23] with Karpathy split [15] (113k/5k/5k as train/validation and test set) and Flickr30k [27] (29k/1k/1k). We extract the visual concepts from the training set of these datasets. We train using ITM and ITC objectives and evaluate by selecting top m examples from the global similarity at the output of the dual encoders then re-ranking these examples using the multimodal encoder. We finetune the model for 10 epochs with batch size 64 (16 per GPU), learning rate that decays from $1e-5$ to $1e-6$ using cosine annealing scheduler.

Visual Question Answering (VQA) consists of answering a question based on an image. We finetune on VQA v2 dataset [11] that contains images from COCO and is split as 83k/41k/81k images for train/validation/test. The visual concepts are extracted from the COCO training set. We consider the task as a generation problem constrained to the set of possible answers. We add a 6-layer transformer decoder initialized by the pretrained

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

multimodal encoder, which generates the answer based on the output of the multimodal encoder. We train on the train and validation set and did not include additional question and answers from Visual Genome as done in other work [3, 19], as it slows significantly the training. We finetune for 8 epochs, batch size 32 (16 per GPU), learning rate $2e-5$ warmed up to $1e-5$ and decayed using cosine scheduler to $1e-6$.

Visual Entailment (SNLI-VE) [35] is a visual reasoning task where given an image and a sentence the model should predict if the sentence is a contradiction, neutral or entailment to the image. We train on SNLI-VE [35] dataset, which is based on SNLI [1] and Flickr30K and consists of 30K images for training and 1k for evaluation and test. The visual concepts are extracted from the training set of Flickr30k. We regard this task as a three-way classification and we add a classification head on top of the [CLS] token of the mulltimodal encoder, consisting of 2 fully connected layers with ReLU activation in between. We train for 10 epochs with batch size 32 (16 per GPU) and initial learning rate 2e-5.

Natural Language For Visual Reasoning (NLVR^2) [33] is another visual reasoning task. The model should predict if a sentence describes the 2 input images. We consider the task as a binary classification task and we replicate each layer in the multimodal encoder so that the encoder can accept 2 images instead of 1, the replicated layers are initialized from the original ones and the linear projection of the keys and values of the cross attention layers are shared between the original and the replicate. The visual concepts are extracted from the sentences of the NLVR^2 training set. The prediction is obtained from a classification head on top of the multimodal encoder's [CLS] token. We pretrain the model on the pretraining corpus for additional 1 epoch on a classification task (text assignment) that consists of assigning the test to either one of the 2 images or non of them [19]. We finetune the model on [33] for 10 epochs, batch size 16 (8 per GPU) with initial learning rate 2e-5.

Visual Grounding Consists of matching a text query to the corresponding image region. Here we evaluate our approach in a weakly supervised manner where we only use the textual query/caption without using bounding boxes. During training, we train with ITC and ITM loss as we do for retrieval, and during inference we use Grad-CAM [30] to rank the proposals from [39]. We evaluate on the RefCOCO+ dataset [38] which contains 140K expression for 20K images collected from the training set of COCO. The model is trained for 5 epochs without random cropping and initial learning rate of $1e-5$. The batch size is set to 32 (16 per GPU).

A.2 Implementation Details

Here we describe additional implementation details to train ViCHA. The visual encoder is ViT-B/16 [6] initialized with DeiT pretrained on ImageNet [34], the text encoder is the first 6 layers of BERT-base [5] and the multimodal encoder is the last 6 layers of BERT-base. E_{vc} is the first 2 layers of BERT-base. We extract 15 concepts for each image and we set p_{vc}=30% for VCA. For H-ITC loss, we set $\lambda_{H-ITC} = 0.1$ and align the last 6 layers of E_v with all 6 layers of E_l. For U-MIM, we use 2-layer transformer encoder with 16 heads and set $\lambda_{MIM} = 1$. The hidden dimension is 768 and kept constant across all modules, the embedding dimension is 256. For the momentum model, the momentum is set to 0.995 and the queue size to 65536. The learned temperature is initialized to 0.07. The images are randomly cropped to 256×256

during pretraining. We use AdamW [24] with learning rate $1e-5$ that is warmed up during the first 2K iterations and then decreased using cosine schedule.

A.3 Ablation Study

In this section, we investigate the importance of different design choices. We follow other approaches [8, 36] and finetune only on Flickr30K and VQA. We compare against ALBEF* (Baseline) trained on the same setup described in Section in the main paper We keep the image resolution as in pretraining (256).

U-MIM	M-MIM	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10	RSUM	VQA test-dev Acc.
Baseline		85.8	97.4	98.5	70.2	89.9	94.0	535.8	71.1
✓		88.6	97.6	99.2	72.3	91.1	94.8	543.6	71.8
	✓	86.8	96.4	98.5	71.8	90.8	94.7	539.0	71.8
✓	✓	87.4	97.6	99.2	71.4	90.8	94.6	541.0	71.9

Table 1: Ablation study on MIM objective.

Masked Image Modeling: Table 1 shows the advantage of using MIM objective. Contrary to the results reported in other approaches [8], we find that MIM based on the MAE objective brings additional improvements in both cases (U-MIM +7.8% and M-MIM +3.2% RSUM). In addition, we found that the unimodal MIM outperforms the multimodal one. This may be due to the interchange between the text and image tokens as query to the multimodal decoder. However, using M-MIM with a decoder that takes the image tokens or both image and text tokens as query could lead to better performance. We tried also to combine both objectives, but we did not notice a significant improvement (+8.8% RSUM and +0.9% VQA). Nonetheless, these results reveal the importance of using MIM and SSL objectives for VLP, especially SSL with the unimodal encoders, as it seems improving the unimodal representation lead to better cross-modal alignment.

VCs corpus	VCs corpus size	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10	VQA test-dev Acc.
Baseline	-	85.8	97.4	98.5	70.24	89.9	94.02	71.15
4OD	6.1K	86.7	97.6	99.4	72.5	91.68	95.12	71.27
ViCHA	206.5K	87.6	97.6	99.3	74.0	91.56	95.34	71.57

Table 2: Ablation Study: Source of Visual Concepts.

Source of VCs: How the visual concepts are gathered could affect the performance. This is shown in Table 2, where we assess the source of VCs. We found that having a large pool of noisy concepts extracted from captions is better than using a smaller and cleaner pool of concepts extracted from the classes of several object detection datasets (*i.e*, Visual Genome, OpenImage 6, Objects365 and COCO) . However, the number of VCs can be reduced as discussed in Section A.6.

VCs Extraction Method	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10	VQA test-dev Acc.
Baseline	85.8	97.4	98.5	70.24	89.9	94.02	71.15
VinVL	88.4	98.3	99.3	73.22	92.56	95.54	72.75
CLIP	89.3	98.3	99.3	76.98	93.36	95.86	72.46

Table 3: Ablation Study: Extraction of Visual Concepts.

Visual Concepts Extraction Method: Here we show the importance of using our CLIP-based approach for VCs extraction compared to other approaches based on object detectors (*e.g.* VinVL [42]). Table 3 shows that VCs extracted with CLIP give better performance on retrieval and slight degradation on VQA. This might be due to the fact that CLIP excels in retrieval tasks and echos this bias to the extracted VCs. Importantly, with both approaches, we still get better results than the baseline. This validates that the improvement is not coming only from the large scale pretraining of CLIP.

E_{vc}'s # of layers	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10	VQA test-dev Acc.
Baseline	85.8	97.4	98.5	70.24	89.9	94.02	71.15
0	87.70	97.20	99.10	72.74	91.60	95.02	71.28
2	87.60	97.60	99.30	74.00	91.56	95.34	71.57
4	88.60	97.5	98.40	74.56	92.64	95.54	71.69

Table 4: Ablation study: size of the Visual Concepts Encoder.

Size of the Visual Concepts Encoder (VCE): Having a larger VCE can help to project the VCs to the image patch embedding space and embed VCs more effectively. This is supported by the results in Table 4, showing that increasing the VCE size helps to get better improvements. However, we favor to keep the model simple and use only 2-layers encoder.

p_{vc} %	Flickr30K (1K test set) TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10	VQA test-dev Acc.
100 % (No VCA)	87.6	97.6	99.3	74.0	91.6	95.3	71.6
50 %	88.5	97.5	99.0	75.6	93.3	95.4	72.2
30 %	89.3	98.3	99.3	76.98	93.36	95.86	72.46
10 %	89.1	98.4	99.6	75.56	92.78	96.3	72.43

Table 5: Ablation study: the percentage of Visual Concepts Augmentation (VCA) during pretraining.

VCA: Table 5 shows the effect of VCA. Reducing the percentage of randomly selected VCs brings significant improvements. This might help the model to avoid relying too much on them by showing different combination at each training step. Interestingly, reducing too much the number of VCs (10 % or 1 concept/image) impedes the performance.

Dataset	Corpus size	Examples
COCO	18K	man, people, street, pizza, laptop, baseball player, tennis court, dogs, rain, bicycles, computer keyboard, airport runway
VG	78.5K	person, sky, grass, giraffe, airplane, street light, license plate, coffee table, tennis shoes, clothes, sunlight, orange shirt
SBU	110K	house, water, beach, background, office building, kitchen window, netherlands, street signs, eiffel tower, mississippi river, architecture, computer desk

Table 6: VCs corpus for each dataset.

A.4 Experiments on 4M images

In this section, we train ViCHA on 4M images and compare it with other SoTA. We follow the same implementation details described in the main paper except the following (to have a fair comparison with other work [19]); we add CC3M [31], pretrain for 30 epochs using 8 GPUs, during Finetuning we increase the image resolution to 384 for VQA and NLVR^2, we use additional Visual Genome questions for VQA.

Results: Figure 7, shows that ViCHA outperforms state of the art models on most of the tasks (*i.e.*, VQA, NLVR-dev, F30K R@1 TR, MSCOCO R@1 TR) and competitive on other tasks. This validate the effectiveness of the proposed approach, however, the improvement is larger on low data regime (1M images) as shown in the main paper.

Method	# Pre-train Images	VQA		NLVR^2		SNLI-VE		Flickr30K (1K test set)						MSCOCO (5K test set)					
								TR			IR			TR			IR		
		test-dev	test-std	dev	test-P	val	test	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViLBERT [ ]	3M	70.55	70.92	-	-	-	-	-	-	-	58.2	84.9	91.5	-	-	-	-	-	-
l2-in-1 [ ]	3M	73.15	-	-	78.87	-	76.95	-	-	-	67.90	-	-	-	-	-	68.00	-	-
ERNIE-ViL [ ]	3.8M	73.18	73.36	-	-	-	-	86.7	97.80	99.00	74.44	92.72	95.54	-	-	-	-	-	-
ImageBERT [ ]	6M	-	-	-	-	-	-	87.0	97.6	99.2	73.1	92.6	96.0	66.4	89.8	94.4	50.5	78.7	87.1
Unicoder-VL [ ]	3.8M	-	-	-	-	-	-	-	-	-	-	-	-	62.3	87.1	92.8	46.7	76.0	85.3
UNITER [ ]	4M	72.70	72.91	77.18	77.85	78.59	78.28	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR [ ]	4M	73.16	73.44	78.07	78.36	-	-	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VILLA [ ]	4M	73.59	73.67	78.39	79.30	79.47	79.03	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
UNIMO$_B$ [ ]	4M+1.7M	73.79	74.02	-	-	80.00	79.10	89.70	98.40	99.10	74.66	93.40	96.08	-	-	-	-	-	-
ViLT [ ]	4M	70.94	-	75.24	76.21	-	-	83.5	96.7	98.6	64.4	88.7	93.8	61.5	86.3	92.7	42.7	72.9	83.1
ALBEF [ ]	4M	74.54	74.70	80.24	80.50	80.14	80.30	94.3	99.4	99.8	82.8	96.7	98.4	73.1	91.4	96.0	56.8	81.5	89.2
COTS [?]	4M	-	-	-	-	-	-	88.2	98.5	99.7	75.2	93.6	96.5	66.9	88.8	94.0	50.5	77.6	86.1
CODIS [ ]	4M	74.86	74.97	80.50	80.84	80.47	80.40	95.1	99.4	99.9	83.3	96.1	97.8	75.3	92.6	96.6	58.7	82.8	89.7
TCL [ ]	4M	74.90	74.92	80.54	81.33	80.51	80.29	94.9	99.5	99.8	84.0	96.7	98.5	75.6	92.8	96.7	59.0	83.2	89.9
MixGen [ ]	3M	74.51	74.79	80.23	80.94	80.05	80.05	94.8	99.4	100	82.4	96.3	98.0	74.2	92.8	96.4	57.3	82.1	89.0
ViCHA	4M	74.99	75.07	80.84	80.44	79.92	79.35	95.2	99.1	99.8	82.56	95.56	97.88	77.16	93.18	96.68	58.82	83.21	90.20

Table 7: Comparison with SOTA; we report the accuracy on VQA, NLVR^2, VE and R@K for Image-Text Retrieval. ViCHA ouperforms SoTA on most of the tasks.

A.5 Data Filtering

The growing size of image-text datasets scraped from the internet comes also with a growing amount of all kinds of accompanying noise. Even though, the presence of noise is not a major issue for extremely large datasets, the need for efficient data filtering techniques becomes inevitable when considering small to medium scale data regimes.

Many existing works focus on filtering the dataset from noise based on handcrafted filters [7, 4, 31]. Despite being effective, here we rather focus on selecting the best examples or image-text pairs for the underlying alignment task.

We propose a simple yet effective technique that selects the best image-text pairs based on the CLIP cosine similarity. This technique, can also cope with noisy and corrupted captions, as these captions have low similarity scores compared to the image. In addition, some captions

Model	dataset percentage %			Flickr30K (1K test set)						VQA
				TR			IR			test-dev
	COCO	VG	SBU	R@1	R@5	R@10	R@1	R@5	R@10	Acc.
ALBEF*	100	100	100	85.8	97.4	98.5	70.24	89.9	94.02	71.04
	100	50	100	86.1	97.3	99.1	72.7	91.92	95.34	71.58
	100	15	100	86.8	97.3	99.2	71.94	91.34	95.02	71.17
	100	50	50	88.5	97.6	99.1	72.38	91.52	95.12	71.38
	100	50	70	88.5	97.6	99.1	72.16	91.78	95.3	71.55
ViCHA	100	100	100	89.7	98.6	99.8	75.78	93.46	96.54	72.55
	100	50	100	89.4	98.1	99.7	76.3	93.6	96.6	72.36
	100	50	70	88.2	98.1	99.5	75.66	93.22	96.26	72.41

Table 8: To see later if we include all these results (with 15% of vg, ViCHA seems useless) CLIP-based data selection/filtering results: we ablate the choice of the selection percentage for each dataset.

are hard to align, especially when describing a small region of the input image that can be potentially removed with augmentation techniques (*e.g.* random crop). Moreover, it can filter wrong captions that might be fetched from the internet. Thus, we argue that noise filtering is not enough to have good captions, as they ignore the level of correspondence between the image and the caption. This data selection technique is similar to what have been proposed in [29], however they use a fixed and manually selected threshold, while here we select top p % pairs. We use the terms data selection and filtering interchangeably.

CLIP-based data selection: We use CLIP-ViT/B-16 to compute the image features, and the corresponding text encoder to compute the text features of the corresponding caption. We compute the cosine similarity for all pairs then we select the top p % of the pairs for training.

Experimental Results: Table 8 shows the results of applying this technique to ALBEF* and ViCHA. We consider COCO as a clean dataset as it is manually annotated for image captioning, and focus on Visual Genome (captions that describe small regions in the image) and SBU (noisy captions). For Visual Genome, we filter only the captions (e.g., 50 % of the captions for each image). In the case of ALBEF*, we notice that the using all the captions in the Visual Genome (VG) dataset harm the performance, as the results are significantly better when training only with 15 % and 50 % of the caption for each image. We obtain comparable performance when training on 50 % or 70 % of SBU. In the case of ViCHA, we have a comparable results to training on all data when training with 50 % VG and 70 % of SBU.

A.6 Discussion

Visual Concepts: We show (in the main paper) that VCs can capture high level, global and some aspects in the scene that can not be shown explicitly or detected by other techniques (*e.g.*, Object detectors as in OSCAR [21]).

However, our approach is far from being perfect. We show some limitations in Figure 3, where the VCs can be (a) redundant, (b) wrong, due to some biases in the dataset, (c) lacking the ability of capture local objects and (d) inherit the noise from the captions.

There are several reasons for these limitations; (a) The poor (naive) text/captions filtering techniques that we adopt which keep a lot of redundant (*e.g.* plural vs singular, prefixes/sufixes...) and noisy concepts, (b) the retrieval model (CLIP) that is trained for global representation alignment (matching class tokens) which hinders the ability to grasp local concepts. (c) the biases that can be included in the unknown training dataset of CLIP, which cause the model to capture wrong/hallucinated concepts, and finally (d) the limitation of the Scene Graph Parsing technique that we use.

These limitations can be addressed for example by; using some advanced text filtering techniques and eventually a model trained for more finegrained alignment on large unbiased dataset.

Model size: Another aspect that can help in low data regime, and we little investigate in this work (*e.g.* Table 4), is the size of the model. Recent work [13, 14, 17, 41] show that bigger models are more data/sample efficient. However, in this work we put forward model simplicity and keep this study for future work.

Comparability of existing methods Despite being common to compare different VLP models, a lot of these models are not comparable due to different design choices, such as;

- The model: size such as base vs large for transformers, single [19] vs double [8] multimodal decoders. Different unimodal encoders and initialization; Roberta [8] vs BERT [19], Swin [7] vs CLIP [8, 32] vs ViT [19].
- Pretraining dataset: number of images (200K vs 4M vs 14M...), number of image-caption pairs (*e.g.* for 4M images: 5M [9, 19, 36] vs 10M [8, 16]). The type/level of annotations used (*e.g.*, bounding boxes [7, 22, 40]).
- Number of epochs: *e.g.*, 10 [8] vs 30 [9, 19, 36].
- Batch size: which is important especially with contrastive learning (*e.g.*, 4096 [8, 16] vs 512 [19, 36]).

For this reason we argue that the best way to assess a given method is to re-implenment or re-train other models with the same setup as the proposed one.

A.7 Qualitative Results

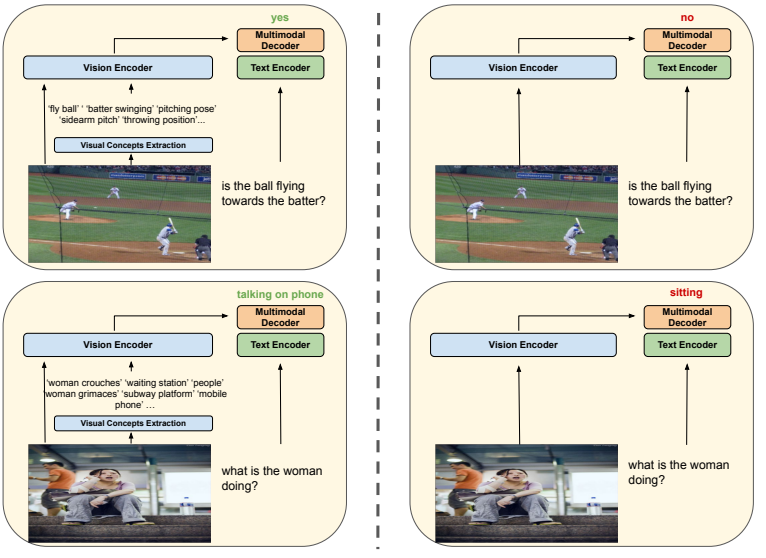


Figure 1: VQA results: comparison with and without using VCs

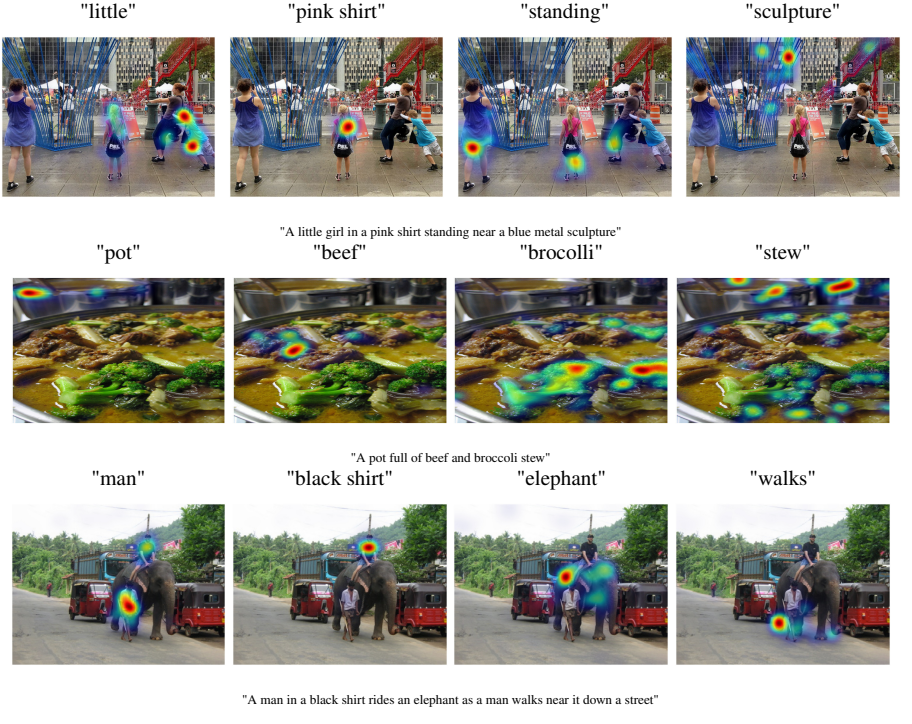
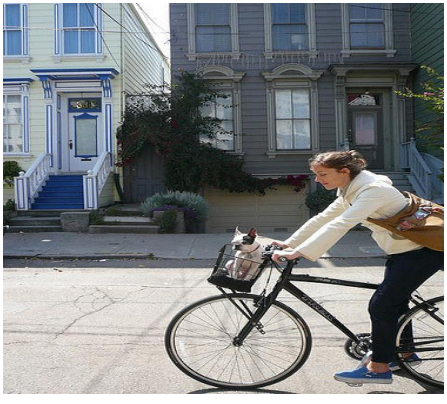
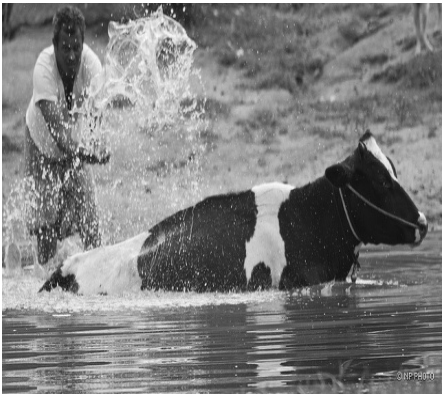


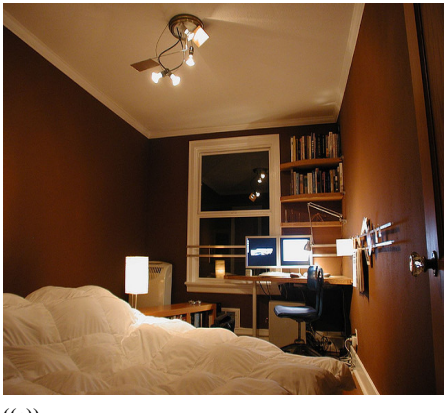
Figure 2: Visualization of the Grad-CAM corresponding to different words in the caption. The model is finetuned on RefCOCO+ and the images are from COCO test set. The model can accurately localize the objects mentioned in the caption, and also works well with specialized domains, such as food.



((a)) **Dataset**: COCO
Caption: 'a woman riding a bicycle by two buildings'.
VCs: 'bicycle riding', 'bicycle ride', 'bicycle rides', 'bicycle rack', 'bicycle basket', 'bike carriage', 'bicyclist', 'delivery bicycles', 'bicycle cart', 'biking', 'bicyclist rides', 'bicyclists', 'bicycle rider', 'bike riding', 'bicycler'



((b)) **Dataset**: COCO
Caption: 'A black and white cow and a man are in the water.'
VCs: 'bear splashes', 'dog swimming', 'bear swimming', 'water buffalo', 'animal swimming', 'holstein cow', 'border collies', 'bear diving', 'elephant swimming', 'calf', 'collie dog', 'border collie', 'retriever dog', 'bull dog', 'splash'



((c)) **Dataset**: COCO
Caption: 'A computer is sitting on a computer desk on the far side of the room.'
VCs: 'type bedroom', 'bedroom space', 'home bedroom', 'room combination', 'room arrangement', 'house bedroom', 'bedroom furniture', 'bedroom area', 'studio apartment', 'living someones room', 'someones bedroom', 'bedroom scene', 'bedroom setup', 'college dorm', 'room scene'

((d)) **Dataset**: SBU
Caption: 'Kevin is the cutest boy in the world. He BAKED Chris this cake and decorated it, penis and all!'
VCs: '4.1.2011from', '7/12/2009 view', 'rpop-2008', 'rpop-2007', 'birthday candles', '—cake', 'cake candles', 'trick candles', 'birthday candle', 'guitar cake', '2009.compliments', 'hat cake', '2/21/2010', 'boy baby cake', 'may,09'

Figure 3: Illustration of some limitations of VCs; (a) redundancy, (b) wrong VCs, (c) lacking local VCs and (d) noisy VCs.

References

[1] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

[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] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[4] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

[5] 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.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[7] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022.

[8] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

[9] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660, 2022.

[10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020.

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[12] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multi-modal data augmentation. *arXiv preprint arXiv:2206.08358*, 2022.

[13] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022.

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[17] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022.

[18] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[19] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] 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.

[21] 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.

[22] Zejun Li, Zhihao Fan, Huaixiao Tou, and Zhongyu Wei. Mvp: Multi-stage vision-language pre-training via multi-level semantic alignment. *arXiv preprint arXiv:2201.12596*, 2022.

[23] 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.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[26] 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.

[27] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[28] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

[29] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[31] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[32] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zf_Ll3HZWgy.

[33] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644.

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[35] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[36] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[37] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.

[38] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[39] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[40] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

[41] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022.

[42] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.