Learning from Children:
Improving Image-Caption Pretraining via Curriculum

Hammad A. Ayyubi[1], Rahul Lokesh[1], Alireza Zareian[2], Bo Wu[3], Shih-Fu Chang[1]

[1] Columbia University   [2] Snap Inc.   [3] MIT-IBM Watson AI Lab

{ha2578, rl3164}@columbia.edu

Abstract

Image-caption pretraining has been quite successfully used for downstream vision tasks like zero-shot image classification and object detection. However, image-caption pretraining is still a hard problem – it requires multiple concepts (nouns) from captions to be aligned to several objects in images. To tackle this problem, we go to the roots – the best learner, children. We take inspiration from cognitive science studies dealing with children's language learning to propose a curriculum learning framework. The learning begins with easy-to-align image caption pairs containing one concept per caption. The difficulty is progressively increased with each new phase by adding one more concept per caption. Correspondingly, the knowledge acquired in each learning phase is utilized in subsequent phases to effectively constrain the learning problem to aligning one new concept-object pair in each phase. We show that this learning strategy improves over vanilla image-caption training in various settings – pretraining from scratch, using a pretrained image or/and pretrained text encoder, low data regime etc. Code available at: https://github.com/hayyubi/cur_vl.git.

1 Introduction

Recently, there has been a tremendous interest in employing image-caption pretraining for downstream vision tasks like zero-shot object classification (Radford et al., 2021) and zero-shot object detection (Zareian et al., 2021; Li et al., 2022). The idea is to learn a common semantic space where the visual embeddings of objects in images lie close to the textual embeddings of the concepts (objects' name/tag/label) in captions they refer to. This learned semantic space is later exploited for zero-shot object recognition by finding the concept embedding nearest to the objects' embeddings.

Despite the recent success, image-caption pretraining is a complex problem as it entails aligning
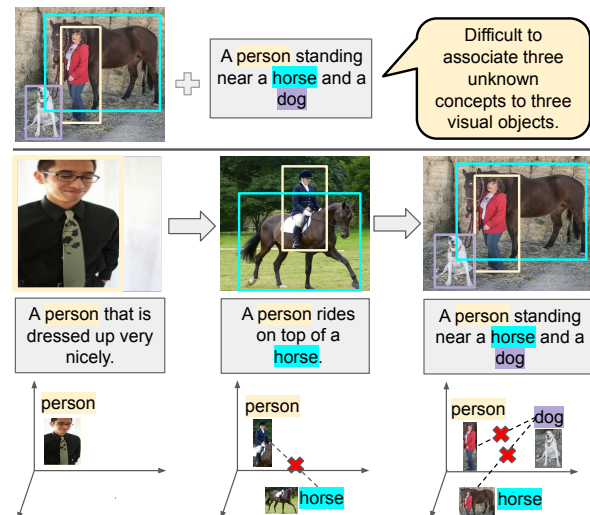

Figure 1: Top: Normal image-caption pertaining. Bottom: Proposed Curriculum Learning Framework. The curriculum eases the learning problem by requiring the model to align only one concept-object pair at a time.

multiple concepts in a caption with multiple objects in an image, as shown in fig. 1. Different methods have tried to solve this problem from various angles – CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) by using more data, ALBEF (Li et al., 2021) by using more complex network architecture, Florence (Yuan et al., 2021) and CoCa (Yu et al., 2022) by using more tasks and ERNIE-ViL 2.0 (Shan et al., 2022) by using more data augmentations (views).

We propose an alternative approach based on a novel learning strategy that is architecture agnostic and does not require any additional data or compute. We take inspiration from cognitive science research studying how children learn language (concepts) in early stages by just observing their surroundings (images). Specifically, we refer to two studies showing that children learn rapidly if the object of interest is unambiguous (Pereira et al., 2014) and by applying co-referential statistics across multiple scenes (Smith and Yu, 2008).

We implement these two ideas via a curriculum

learning approach (demonstrated in fig. 1):

1. We train the model in multiple phases of increasing difficulty with each phase containing one more concept in the caption than the previous one. Moreover, each phase contains only one new concept, the rest seen in prior phases.

2. In each phase, we leverage the concept-object association learned in prior phases to recognize the seen concepts and focus on aligning the new/unseen concept (section 2.2.2).

These two strategies effectively reduce the problem of aligning multiple object-concept pairs per training sample to aligning only one such pair.

To the best of our knowledge, no prior work has applied curriculum leaarning to image-caption pretraining in this way. Srinivasan et al. (2022) apply a curriculum based on the difficulty of negative samples in contrastive loss. Whereas, Liu et al. (2021) design the curriculum based on the granularity of text: from words to phrases to sentences.

Although our proposed approach can be applied to any multimodal network architecture, we pick OVR-CNN (Zareian et al., 2021) due to its simplicity. We pretrain it with the proposed curriculum learning approach and evaluate on the downstream task of zero-shot object detection. We demonstrate that curriculum learning outperforms vanilla image-caption pretraining on a variety of architectural settings – with and without a pretrained image encoder and/or a pretrained text encoder. We even show superior performance in low-data settings, suggesting our method can be leveraged in low-resource scenarios as well.

2 Method

We propose a curriculum learning framework to improve image caption pretraining. In this work, we apply it to OVR-CNN as its architecture is simpler and easier/faster to train/evaluate. We begin the description of our approach with a brief background on OVR-CNN. Next, we discuss how we modify it to implement the proposed curriculum learning framework.

2.1 OVR-RCNN Background

OVR-RCNN is a dual-encoder (separate visual encoder and text encoder) multimodal architecture. First, it pretrains the encoders using image-captions and later utilizes them for the downstream task of

object detection. We only discuss the pretraining procedure as we only utilize this component.

OVR-RCNN's visual encoder is ResNet-50 (He et al., 2016) and text encoder is either BERT (Devlin et al., 2019) or GloVE (Pennington et al., 2014). The visual encoder takes an image, I with $w \times h$ dimensions, as input and outputs a feature map of $w/32 \times h/32$ regions. Each feature map is a vector which is transferred to language space using a projection layer. This gives the visual embeddings, e_i^I, for each region i. The tokenized caption, C, is input to the text encoder which outputs an embedding e_j^C for each token j.

The token-image region pair is aligned via weak supervision. Specifically, a global alignment score between image and caption, $\langle I, C \rangle_G$ is calculated using a locally weighted average alignment score of image regions and tokens as follows:

$$\langle I, C \rangle_G = \frac{1}{n_C} \sum_{j=1}^{n_C} \sum_{i=1}^{n_I} a_{i,j} \langle e_i^I, e_j^C \rangle_L \quad (1)$$

where $\langle ., . \rangle_L$ is the dot product of two vectors, n_I and n_C are the number of image and caption tokens respectively, and

$$a_{i,j} = \frac{\exp \langle e_i^I, e_j^C \rangle_L}{\sum_{i'=1}^{n_I} \exp \langle e_{i'}^I, e_j^C \rangle_L} \quad (2)$$

The model is trained using contrastive learning by maximizing the global alignment score, $\langle I, C \rangle_G$, between positive image-caption pairs and minimizing it between negative pairs sampled from the same training batch.

$$\mathcal{L} = -\log \frac{\exp \langle I, C \rangle_G}{\sum_{\{I', C'\} \in \mathcal{N}_{I,C}} \exp \langle I', C' \rangle_G + \exp \langle I, C \rangle_G} \quad (3)$$

where, $\mathcal{N}_{I,C} = \{I, C' | C' \in \mathcal{B}_C\} \cup \{I', C | I' \in \mathcal{B}_I\}$ and $\mathcal{B}_C, \mathcal{B}_I$ are batch captions and batch images respectively. This learning objective aligns paired image and caption together and also provides weak supervision for image-regions and caption-tokens association.

2.2 Curriculum Learning Framework

OVR-CNN facilitates object-concept alignment through coarse image-region and concept alignment. However, as an object can span multiple image regions or multiple objects can span an image region, this strategy can be noisy. To eliminate this noise and focus on the contribution of our curriculum framework to object-concept alignment, we train the model using object region features instead of image region features. To this end, object
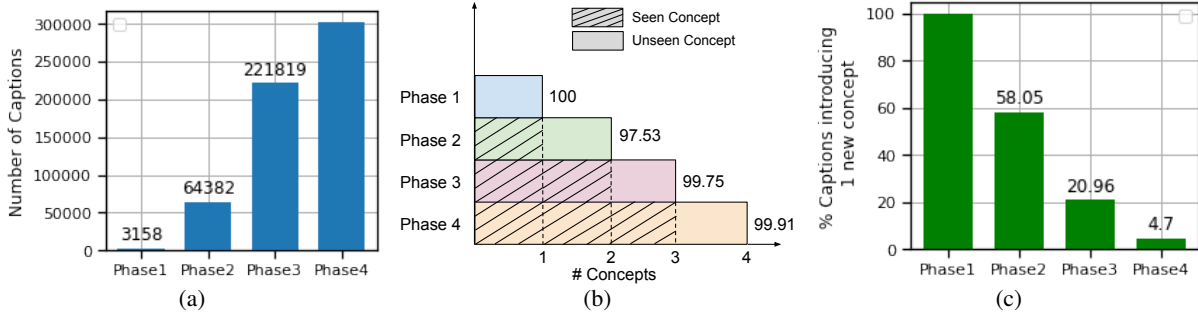
Figure 2: Curriculum Statistics. (a) #Captions Vs Phase (b) The number next to bar shows % of captions per phase with at least #shaded concepts seen previously. (c) % Captions per phase introducing 1 new concept

region bounding box is used to ROI pool (Girshick, 2015a) the image region features. The resulting feature vector e_o^I, for each object o, is used to replace e_i^I in eqs. (1) and (2).

2.2.1 Curriculum Design

The learning is divided into $1, 2, 3 \ldots k$ phases. Each phase p is trained with only those image-caption pairs having p concepts per caption. To divide the data into phases, we use spacy[1] to PoS (Part of Speech) tag the captions. Depending upon the number of nouns in each caption, the caption and its paired image is grouped into the corresponding phase. This strategy of designing the curriculum also imparts the data an additional property empirically – at most only one new concept is introduced per caption in each phase (as demonstrated in fig. 2b).

2.2.2 Curriculum Aware Alignment Loss

To recognize the concepts in captions previously seen in prior phases and focus on aligning the new/unseen concept, we formulate a novel Curriculum Aware Alignment Loss ($\mathcal{L}_{\mathcal{C}}$). Specifically, we first calculate the previously learned object-concept alignment, $a_{o,j}$ from modified eq. (2), using either the trained model from the last iteration ($\mathcal{L}_{\mathcal{CR}}$) or the trained model from the last phase ($\mathcal{L}_{\mathcal{CP}}$). Next, $a_{o,j}$ is used to compute:

$$a'_{o,j} = \frac{\exp\langle e_o^I, e_j^C\rangle_L \exp\left(-\max_o(a_{o,j}).\frac{t}{T}\right)}{\sum_{o'=1}^{n^I} \exp\langle e_{o'}^I, e_j^C\rangle_L \exp\left(-\max_o(a_{o,j}).\frac{t}{T}\right)}$$

where, t is the current iteration number and T is the total number of iterations in training.

For a concept j, which is already closely aligned to an object o, $\max_o(a_{o,j})$ is high. This leads to a low value of $a'_{o,j}$, resulting in less attention being paid to concept j in the current training iteration/phase. Vice versa for a concept that is not

well aligned with any object. $a'_{o,j}$ effectively redistributes the attention of learning to focus more on concepts that are not well aligned with any object. The term t/T has a low value in the beginning of training and gradually scales to 1 by the end. This allows the network to ignore prior knowledge in the beginning while utilizing it in the latter stages.

We use $a'_{o,j}$ to replace $a_{o,j}$ in modified eq. (1), and then use eq. (3) to compute $\mathcal{L}_{\mathcal{C}}$.

3 Experiments

3.1 Pretraining Dataset and Implementation Details

We use the COCO Captions dataset (Chen et al., 2015) for pretraining. It contains 118,287 images and 5x captions. To obtain bounding box regions for objects in images, we use COCO Objects (Lin et al., 2014) dataset as it uses the same set of images as COCO Captions. We divide the data into $k = 4$ phases using the strategy discussed in section 2.2.1. Figure 2a shows number of captions assigned to each phase. As shown in fig. 2b, the majority of captions in each phase have at least k-1 concepts previously seen, allowing the curriculum to introduce at most one new concept per training sample. Further, as more concepts are introduced with each passing phase, the percent of captions per phase actually introducing a new concept decreases (as depicted in fig. 2c). By phase 4, this percent reduces to < 5%. Additional phases of training may not contain enough captions actually introducing a new concept in the curriculum way, making these phases similar to regular image-caption training. Hence, we limit to 4 phases.

We train the model using SGD optimizer, with a batch size of 32 for 4 epochs in each phase, a learning rate of 0.005, a step scheduler, and the loss $\mathcal{L}_{\mathcal{CP}}$.

[1] https://spacy.io/usage

Model	Pretrained Visual BB	Language Backbone	Accuracy
OVR-CNN$_O$	✗	GloVE	$20.62_{\pm0.86}$
Ours	✗	GloVE	$\mathbf{21.64}_{\pm1.02}$
OVR-CNN$_O$	✗	BERT	$22.73_{\pm0.06}$
Ours	✗	BERT	$\mathbf{23.74}_{\pm0.48}$
OVR-CNN$_O$	✓	BERT	$34.46_{\pm0.11}$
Ours	✓	BERT	$\mathbf{35.49}_{\pm0.21}$

Table 1: Curriculum learning vs. baseline in various settings with/without pretrained encoders. BB: backbone

Model	P_1 Obj.	P_2 Obj.
OVR-CNN$_O$	49.36	16.61
Ours	**50.9**	**26.1**

Table 2: Phase wise top-5 accuracy. P_i Obj.: Phase i Objects.

Model	Accuracy
OVR-CNN$_O$	13.10
Ours	**17.45**

Table 3: Object bounding boxes from RPN.

3.2 Downstream Task, Dataset and Transfer

We evaluate the performance of the model on zero-shot object detection task on COCO Objects, val split (4836 images; 33374 instances of 65 object class). The task involves object bounding box predictions besides classifying these object regions to a label (concept). However, our method is aimed only at improving the alignment of object regions to a concept. As such, we eliminate any performance noise from bounding box predictions by only evaluating the classification accuracy of object regions given ground truth object bounding boxes.

Transfer to Downstream Task: We extract object features from image and object bounding box using visual backbone and use it to find the closest class label vector (obtained via language backbone).

3.3 Baseline and Evaluation

Our baseline is OVR-CNN, a regular image-caption pretrained model. However, since our method uses object region features instead of image patch features for multimodal alignment (section 2.2), we also pretrain OVR-CNN with object regions to obtain OVR-CNN$_O$. It is transferred to downstream task similar to our proposed model (section 3.2).

Our proposed curriculum framework outperforms the baseline in various settings, as shown in table 1. The accuracy numbers reported are averaged across three seeds. This demonstrates that our proposed learning strategy works across encoders trained from scratch or pretrained ones.

Performance Gain Analysis. We analyze model performance on object classes introduced during pretraining in phase 1 and phase 2 separately.
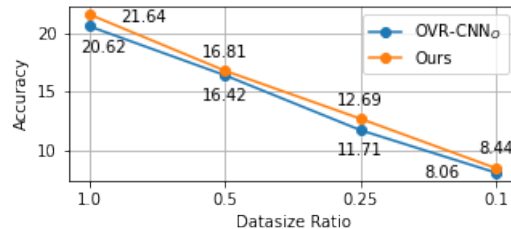

Figure 3: Performance comparison in low data setting.

Model	Cur.	Loss	Accuracy
OVR-CNN$_O$		$\mathcal{L}$	19.73
OVR-CNN$_O$		$\mathcal{L}_{\mathcal{CR}}$	20.65
Ours	✓	$\mathcal{L}$	21.02
Ours	✓	$\mathcal{L}_{\mathcal{CR}}$	21.58
Ours	✓	$\mathcal{L}_{\mathcal{CP}}$	**22.57**

Table 4: Ablation of proposed loss.

As reported in table 2, the improvement in phase 2 objects is ~10x. This illustrates that our curriculum strategy improves alignment of multiple concepts in a caption by focusing on one at a time.

Low Data Setting. Our model outperforms the baseline even if both uses 50%, 25% or 10% data (fig. 3), indicating its utility when data is scarce.

Region proposals instead of ground-truth object regions. We use a RPN model (Girshick, 2015b) trained class-agnostically on Visual Genome (Krishna et al., 2016) to generate object regions. The superior performance of our model against baseline, reported in table 3, demonstrates that our approach is effective even when ground-truth object regions are not available.

Loss Ablation. From table 4, we can conclude that our curriculum design works (Ours + $\mathcal{L}$ > OVR-CNN$_O$ + $\mathcal{L}$); our proposed curriculum aware loss works (Ours + $\mathcal{L}$ < Ours + $\mathcal{L}_{\mathcal{CR}}$) irrespective of curriculum (OVR-CNN$_O$ + $\mathcal{L}$ < OVR-CNN$_O$ + $\mathcal{L}_{\mathcal{CR}}$); curriculum aware loss works better when previous knowledge is taken from the last phase instead of the last iteration (Ours + $\mathcal{L}_{\mathcal{CP}}$ > Ours + $\mathcal{L}_{\mathcal{CR}}$).

Qualitative Analysis. We provide qualitative analysis as well to shed more insights into the cases where our approach works/doesn't work. From Figure 4, we find that our model performs better than OVR-CNN$_O$ in certain cases, especially when the objects are from Phase 2 – "snowboard", "cup", "skis" etc. This provides further evidence towards our claim that our approach improves the alignment of Phase 2 objects.

Comparison of traditional mAP metric for object detection As mentioned before, we have focused our experiments on evaluating object-
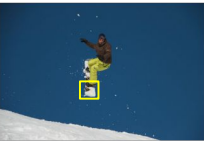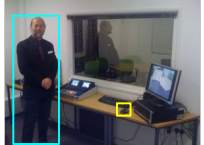
Image	Objects	OVR-CNN_o	Ours	Image	Objects	OVR-CNN_o	Ours
	Snowboard	✗	✓		Cup	✗	✓
	Mouse	✗	✓		Skis	✗	✓
	Person	✗	✓				
	Chair	✓	✗		Knife	✓	✗
	Bed	✓	✓		Cake	✓	✓

Figure 4: Qualitative analysis of cases where our approach works/doesn't work. The top two rows show samples where our model is successfully able to align the objects with the correct concept while OVR-CNN$_O$ makes mistakes. Interestingly, most of these objects are from Phase 2 – "snowboard", "cup", "skis" etc. The bottom row shows cases where our model makes mistakes. Note: only relevant, not all, objects are shown from each image.

Model	Base	Target	Generalized Base	Generalized Target	All
OVR-CNN	46.8	**27.5**	46.0	**22.8**	39.9
Ours	**48.89**	26.65	**48.18**	21.99	**41.33**

Table 5: Comparison of traditional mAP metric for object detection.

Model	Accuracy
OVR-CNN	**23.45**
Ours	23.33

Table 6: Training with image grid regions instead of object regions.

concept alignment rather than on traditional object detection mAP metric. This was done to avoid unnecessary performance noise arising from training a RPN, which is required for mAP evaluation. However, to test the limits of our model, we evaluate on this noisy mAP metric as well. We keep all the settings similar as Zareian et al. (2021), except we pretrain using our curriculum learning approach. The results are reported in Table 5. We find that our model performs better in the most generic Generalized ('All') set (41.33 vs 39.9), signifying the effectiveness of our approach even in this noisy setting. We further observe that we perform better in the base classes while lagging behind in the target classes. A deeper analysis shows that most of the Phase 2 objects, on which we make major improvements, lie in the base classes. This explains the improved performance on base classes and slight depreciation in target classes performance.

Training with image grid regions. Our curriculum based pretraining method was aimed at improving object-concept alignment by focussing on one object at time. To facilitate this, we pretrained directly with object regions. Image regions were not used to eliminate noise arising from an ob-

ject spanning multiple regions or multiple objects being present in the same image region (object presence noise). However, we further push the limits of our model to assess how it performs when trained with noisy image regions instead of object regions. The results are reported in Table 6. We find that our model performs slightly worse than OVR-CNN (23.33 vs 23.45). We attribute this performance degradation to the inherent object presence noise in image regions as discussed earlier.

4 Conclusion

We proposed a curriculum learning framework to improve image-caption pretraining, using the number of concepts in captions. We also designed a novel curriculum aware loss to focus learning on the unaligned concept in each phase. Our approach outperforms vanilla image-caption pretraining in various settings, including with/without pretrained encoders and small data. Further, we extensively analysed our model to study the contribution of each component.

Limitations

Although our proposed curriculum can be applied to any multimodal architecture, curriculum aware

loss requires modifications for use with dual encoder architectures that don't use cross-modal attention. Additionally, we use an off-the-shelf Part-of-Speech tagger to divide the data into different phases. As such, the correctness of this division is dependent on the quality of tagger. A poor tagger can negatively impact the curriculum design. Moreover, our approach doesn't apply to possible image-captions dataset which contain only short captions, containing possibly only one noun.

Acknowledgement

References

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ross Girshick. 2015a. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*.

Ross Girshick. 2015b. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded language-image pre-training. In *IEEE Computer Vision and Pattern Recognition*. arXiv.

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

Tongtong Liu, Fangxiang Feng, and Xiaojie Wang. 2021. Multi-stage pre-training over simplified multimodal pre-training models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2556–2565, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

A. F. Pereira, L. B. Smith, and C. Yu. 2014. A bottom-up view of toddler word learning. In *Psychonomic bulletin & review*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Bin Shan, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Ernie-vil 2.0: Multi-view contrastive learning for image-text pre-training.

Linda B. Smith and Chen Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. In *Cognition*, volume 106, pages 1558–1568.

Tejas Srinivasan, Xiang Ren, and Jesse Thomason. 2022. Curriculum learning for data-efficient vision-language alignment.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021. Florence: A new foundation model for computer vision.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 2021. Open-vocabulary object detection using captions. In *IEEE Computer Vision and Pattern Recognition*.

A For every submission:

☑ A1. Did you describe the limitations of your work?
After conclusion.

☒ A2. Did you discuss any potential risks of your work?
Our work uses well-established datasets which have not been known to be risky.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☑ A4. Have you used AI writing assistants when working on this paper?
We used ChatGPT to improve the sentence structures and correct grammatical mistakes in all sections of the paper.

B ☑ Did you use or create scientific artifacts?

Section 3.3

☑ B1. Did you cite the creators of artifacts you used?
Section 1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 1; the code for the model we used is released publicly along with the paper.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The code we used is in public domain. We plan to publicly release our original artifacts.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The datasets we have used is public and have been extensively analysed by prior research.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Our method covers only such domain as the data used (which has been previously analysed).

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.1 and 3.2

C ☑ Did you run computational experiments?

Section 3

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Our model is a relatively small model which doesn't incur much computational cost.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discuss this partly in Section 3. By and large, we use the same hyperparameters as used by the baseline.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3.3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 2.2.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.