Supplementary Material for Unified Contrastive Learning in Image-Text-Label Space

Jianwei Yang[1*] Chunyuan Li[1*] Pengchuan Zhang[1*] Bin Xiao[2*]
Ce Liu[2] Lu Yuan[2] Jianfeng Gao[1]
[1]Microsoft Research at Redmond, [2]Microsoft Cloud + AI
{jianwyan,chunyl,penzhan,bixi,liuce,luyuan,jfgao}@microsoft.com

A. Validation dataset details

In addition to the training datasets listed in our main submission, we list in Table 2 the statistics for all the validation datasets used in our experiments. Similar to the Table 1 in our main submission, we calculate the vocabulary size for each dataset, which is typically more than the number of concepts (classes).

B. Experiment details

B.1. Training on image classification data

This part mainly explains the detailed experiment setups for Sec. 4.1 in our main submission.

Model architecture. We employ two representative architectures, ResNet [1] and Swin Transformer [4] to build the visual encoder. The globally pooled feature from last visual encoder layer is used as the visual feature. For language encoder, we use a 12-layer Transformer [7] with hidden dimension of 512 following [5]. Features from visual and textual encoder are projected to the same dimension of 512, using two linear projection layers.

Training protocol. For optimization, we use SGD [6] for all CNN models, while AdamW [3] for all models with Transformers on either vision or language side. We set the learning rate to 0.4 and 0.002, weight decay to 1e-4 and 0.05 for SGD and AdamW optimizer, respectively. All models are trained for 500 epochs with a batch size of 4096. We use same set of data augmentation and regularization as in [4], but do not use MixUp [10] and CutMix [9] except for the last column in Table 2 of our main submission. For all training, we used a cosine learning rate schedule, with 5 epochs and 20 epochs warmup for ResNet and Swin Transformer, respectively.

B.2. Training on image-text-label space

Training protocol for Sec. 4.2.1. We use Swin-Tiny as the visual encoder and follow the training settings in Section 4.1 mostly to train the models on the joint of image-label and image-text pairs. However, we notice there is a severe imbalance between image-label and image-text data as shown in Table 1 in our main submission (*e.g.*, there are around 1.3M images in ImageNet-1K while above 12M images in GCC-12M dataset). To ensure that the model training is not biased to the dominant image-text pairs, we develop a balanced data sampler for two data types. More specifically, at each epoch, we randomly sample a subset of image-text pairs that has the equal or similar size to that of image-label data. In this case, the model sees half image-label data and half image-text data at each iteration for a balanced learning. We keep the number of training epochs the same as 500, so the effective number of training epochs on the image-text dataset is roughly 500×(size of image-label dataset)/(size of image-text pair dataset). For example, the model learns from GCC-12M for around 40 epochs. We find this balanced sampling strategy is very important to achieve the reported performance in our main submission.

Training protocol for Sec. 4.2.2. We followed the training protocol in CLIP [5] for fair comparison. Specifically, we merely used random crop for dataset augmentation in all model trainings. All models including the baseline models are trained for 32 epochs, with batch size 4096, initial learning rate 1e-3 and weight decay 0.1. We also used a cosine learning rate scheduler with 5000 warmup iterations.

C. More results

C.1. Results over separate datasets

In Figure 1, we show the zero-shot classification on 14 datasets by adding different image-caption pairs into the ImageNet-1K, *i.e.* the methods compared in Table 5 in the main text. UniCL takes the advantages of learning rich concept coverage from image-text pairs: On most of the datasets, it outperforms the baseline, especially on fine-grained classification tasks such as Food101 and OxfordPets.

*equal contribution

Table 1. Statistics

Dataset	#Concepts	Vocab. Size	Train size	Test size	Evaluation metric	Source link	Linear Probe	Zero-shot
Food-101	102	139	75,750	25,250	Accuracy	Tensorflow	✓	✓
CIFAR-10	10	10	50,000	10,000	Accuracy	TensorFlow	✓	✓
CIFAR-100	100	100	50,000	10,000	Accuracy	TensorFlow	✓	✓
SUN397	397	432	19,850	19,850	Accuracy	Tensorflow	✓	
Stanford Cars	196	291	8,144	8,041	Accuracy	Stanfold Cars	✓	
FGVC Aircraft (variants)	100	115	6,667	3,333	Mean-per-class	FGVC website	✓	
VOC2007 classification	20	20	5,011	4,952	11-point mAP	voc2007	✓	✓
Describable Textures	47	47	3,760	1,880	Accuracy	TensorFlow	✓	✓
Oxford-IIIT Pets	37	53	3,680	3,669	Mean-per-class	Oxford-IIIT Pet	✓	✓
Caltech-101	102	122	3,060	6084	Mean-per-class	TensorFlow	✓	✓
Oxford Flowers 102	102	147	2,040	6,149	Mean-per-class	TensorFlow	✓	✓
MNIST	10	10	60,000	10,000	Accuracy	TensorFlow	✓	
FER 2013 *	8	12	32,298	3,589	Accuracy	Kaggle fer2013	✓	✓
STL10	10	10	5,000	8,000	Accuracy	TensorFlow	✓	
GTSRB *	43	85	26,728	12,630	Accuracy	GTSRB website	✓	
PatchCamelyon	2	6	294,912	32,768	Accuracy	TensorFlow	✓	✓
UCF101 *	101	153	9,537	3783	Accuracy	TensorFlow	✓	
Hateful Memes	2	2	8,500	500	ROC-AUC	FaceBook	✓	✓
EuroSAT	10	20	5,000	5,000	Accuracy	TensorFlow		✓
Resisc45	45	59	3,150	25,200	Accuracy	TensorFlow		✓
Rendered-SST2	2	2	6,920	1,821	Accuracy	OpenAI		✓

Table 2. Statistics of datasets used in zero-shot and linear probe. * indicates dataset whose train/test size we obtained is slightly different from Table 9 in [5]. ✓indicates the dataset is used in this setting. The datasets are chosen based on the criterion if we can reproduce the numbers reported from [5] and their availability.

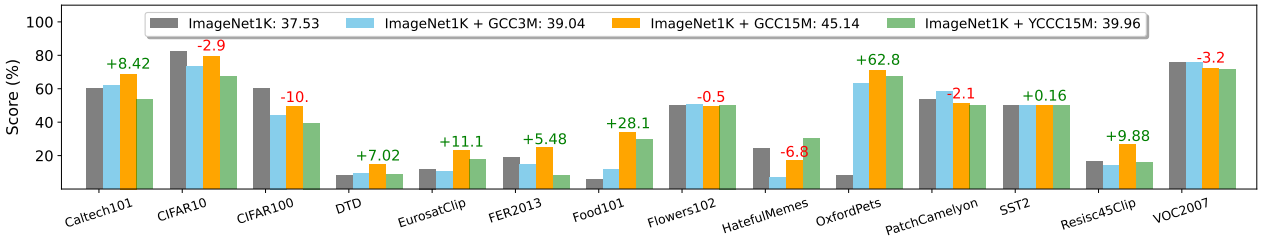

Figure 1. Zero-shot classification on 14 datasets by adding different image-caption pairs into the ImageNet-1K. The averaged scores of each method is reported in the legend. The gain between UniCL on mixed data (ImageNet1K+GCC15M) and image-label data (ImageNet1K) is shown.

C.2. Results with larger vision backbone

In our main submission, we used Swin-Tiny as the visual backbone to study how our UniCL perform when trained on the combination of image-label data ImageNet-21K and image-text pairs YFCC-14M in Table 6. Here, we investigate whether increase the capacity of the vision backbone can further improve the representation learning.

As shown in Table 3, we observe consistent trend as in Table 6 of our main submission. Though using similar amount of image-text-label corpus, combining two type of data can *significantly* improve the zero-shot recognition performance on both ImageNet-1K (**8.6** points) and other 14 datasets in average (**11.0** points). When using the full set of ImageNet-21K and YFCC-14M, both performance can be further improved significantly. These results suggest that our method is agnostic to different model sizes and thus a generic learning paradigm for visual-semantic representations. For comparison, we also list the numbers for Swin-Tiny models after each "/". Clearly, increasing the visual encoder size brings substantial gains in all cases, and particularly significant for the combination of both data types.

Training Data	Method	Zero-shot	
		ImageNet-1K	14 datasets
YFCC-14M	CLIP	32.4/30.1	37.5/36.3
ImageNet-21K	UniCL	29.9/28.5	42.4/37.8
YFCC-14M(half)+ImageNet-21K(half)	UniCL	41.0/36.4	48.5/45.5
YFCC-14M+ImageNet-21K	UniCL	**43.8/40.5**	**52.2/49.1**

Table 3. Ablation studies on the training datasets and tasks. We use Swin-Base [4] as the vision backbone. Each model is pre-trained with 32 epochs following CLIP [5]. Numbers before and after each "/" are with Swin-Base and Swin-Tiny, respectively.

Training Data	Method	Object Detection	
		box mAP	mask mAP
YFCC-14M	CLIP	39.9	37.3
ImageNet-21K	UniCL	41.4	38.6
YFCC-14M(half)+ImageNet-21K(half)	UniCL	41.9	39.0
YFCC-14M+ImageNet-21K	UniCL	**43.1**	**40.0**

Table 4. Object detection transfer learning with different models. We use the pretrained Swin-Tiny models listed in Table 6 of our main submission as the vision backbone.

C.3. Transfer to object detection

In the Table 5 of our main submission, we mainly studied whether image-text pairs can bring benefits to object detection transfer learning compared with the models solely trained on image-label data. As we demonstrated in Table 6 of our main submission, image-label data can help to learn more discriminative representations, and thus benefits ImageNet-1K finetuning and linear probing. Here, we further study whether the learned representations can generalize to object detection task as well. Specifically, we use the Swin-Tiny models pretrained in Table 6 as the vision backbones and train a Mask R-CNN model with $1\times$ schedule following the default settings in Swin Transformer [4] based on Detectron2 [8]. In Table 4, we can see combining two data types with similar amount clearly improve the object detection performance by around 2 points for both box and mask mAP, compared with the CLIP-based model trained on YFCC-14M. This further validates our note that representations learned from pure image-text pair data usually lack the discirminative ability required by transfer learning to image recognition and object detection. As expected, using the full set (last row) brings further around 1 point improvement for both metrics. Along with the reported numbers in Table 5 of our main submission, these results together imply that adding image-text pairs to image-label data and the other way around can universally help to learn a better visual representations compared with the individual counterparts. Adding image-text pairs data can enrich and smoothen the semantic space which may implicitly prompt distinctive representations for the concepts in COCO object detection, while adding image-label data directly imposes the pressure to learn more discriminative representations.

D. More analysis

D.1. Concept distribution

The concepts residing in the training data is arguably crucial to the model learning. Both CLIP [5] and ALIGN [2] exhaustively collect hundreds of millions of image-text pairs to cover as many visual concepts as possible. Though the datasets used in our experiments are at much smaller scale, we are still interested in the concept distributions of different datasets. In Fig. 2, we show the occurrences of top 1000 concepts in GCC-3M, GCC-12M and YFCC-14M. Along with the remaining concepts that do not show here, all three datasets have extreme long-tail distributions. For example, the most frequent concept "view" in GCC-12M appears over 185,363 times, while the 10k-th concept "candle holder" only appears 501 times, knowing that there are more than 584k concepts in the whole set.

Dataset	GCC-3M	GCC-12M	YFCC-14M
GCC-3M	100%	46.5%	50.2%
GCC-12M	46.5%	100%	37.9%
YFCC-14M	50.2%	37.9%	100%

Table 5. Overlap ratio of top 10k concepts among three image-text pair datasets, GCC-3M, GCC-12M and YFCC-14M. The matrix is symmetric.

Interestingly, we find the overlap of most common concepts across three datasets is lower than what we expect. Table 5 shows the overlap ratios of top 10k concepts among three datasets. These relatively lower overlapping indicates the sufficient diversities and complementary among them.

D.2. Concept coverage

Given the concept distributions above, we further investigate the concept coverage between training datasets and validation datasets. In Table 6, we calculate the coverage ratio to be the percentage of concepts mentioned by the pretraining data, including ImageNet-1K, ImageNet-21K, GCC-3M, GCC-12M and YFCC-14M. Coverage ratios equal or larger than **50**% are highlighted.

Accordingly, for image-label dataset ImageNet-1K, it has some overlaps with CIFAR-100 (24.0%) and Caltech-101 (24.5%). This may explain why the zero-shot performance on these two datasets shown in Fig. 1 is relative higher. In contrast, we also notice that even with less or no coverage, the model pretrained on ImageNet-1K with our method still attain reasonably good zero-shot performance on datasets like CIFAR-10, Flowers102, Oxford Pet, *etc*.

Similarly, for ImageNet-21K, it covers a certain proportion of concepts in the validation sets, such as CIFAR-10, CIFAR-100, Caltech-101, *etc*, and we did observe high zero-shot recognition performance on them in the Table 6 of our main submission. Nevertheless, for other datasets like *Hateful Memes*, *PatchCamelyon*, there are zero concept overlaps, while our model still realizes reasonable performance. This indicates that our model is not just memorizing the concepts appearing in the training datasets, but also learns to understand the underlying structures of different concepts, which has been also demonstrated in Fig. 5 of our main submission.

Finally, we find image-text pairs data have higher coverage of concepts than image-label datasets almost on all validation sets. Among the three image-text pair datasets, GCC-12M has relatively higher coverage than the other two

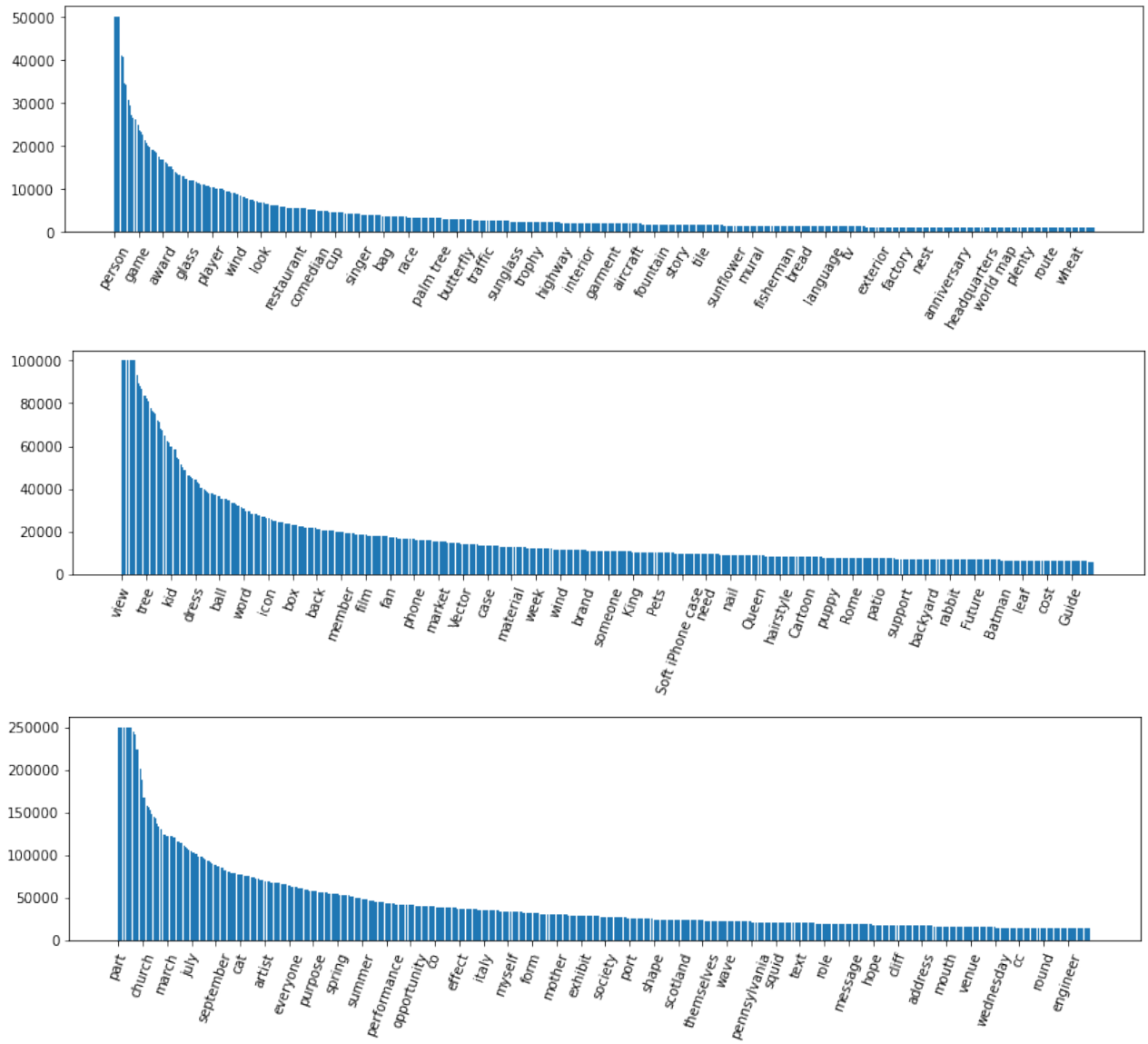

Figure 2. From top to bottom, bar charts are top 1000 most frequent concepts in GCC-3M, GCC-12M and YFCC-14M, respectively. We trim the heights of most frequent concepts for better display. For clarity, we display the concept name for every 25 concepts.

datasets. This may also explain why we observe better performance in the comparisons shown in Table 5 of our main submission. However, we also notice that higher concept coverage does not necessarily means better zero-shot performance. For example, even though all of these three datasets have a fully coverage of concepts in CIFAR-10 and CIFAR-100, adding them into the pretraining hurts the performance as shown in Fig. 1. We suspect there might be some significant gaps in the image domain between the pretraining and validation datasets even though they share common semantic concepts. Moreover, images in image-text pairs usually contain multiple objects, the coverage of concepts does not necessarily means the model can learn to grounding the concepts to the specific image contents. How to better leverage the image-text pair data and build a more gounded visual understanding worth further studies.

D.3. Concept visualizations

In Fig. 3, we further show the concept embeddings for two models as in Fig. 4 in our main submission. Fig. 3 left shows the model trained only on ImageNet-1K while right shows the model trained jointly with ImageNet-1K and GCC-15M. The model trained with two type of data understand the novel concepts from ImageNet-21K much better than the left one. For example, the left model put "porthole" and porcupine close to each other but the former is a circular window and latter is an animal. In contrast, the model at right side can easily find the "porcuponefish" as the

Dataset			ImageNet-1K		ImageNet-21K		GCC-3M		GCC-12M		YFCC-14M	
Name	#Concepts	Vocab. Size	Cover.	#Img/C.	Cover.	#Img/C.	Cover.	#Img/C.	Cover.	#Img/C.	Cover.	#Img/C.
ImageNet-1K	1,000	1,233	**100**%	1300	0%	0	45.3%	247.0	**78.5**%	851.1	**69.3**%	1930.8
Food-101	102	139	4.0%	1300.0	20.8%	650.0	21.8%	39.8	**58.4**%	250.8	**67.3**%	408.8
CIFAR-10	10	10	0.0%	0.0	**90.0**%	650.0	**100.0**%	6175.4	**100.0**%	19969.8	**100.0**%	32998.9
CIFAR-100	100	100	24.0%	1300.0	**65.0**%	650.0	**95.0**%	3928.4	**99.0**%	15628.5	**99.0**%	18303.2
SUN397	397	432	5.0%	1300.0	28.5%	650.0	48.1%	818.9	**65.5**%	2355.4	**66.5**%	7043.2
Stanford Cars	196	291	0.0%	0.0	0.0%	0.0	0.0%	0.0	0.0%	0.0	0.0%	0.0
FGVC Aircraft (variants)	100	115	0.0%	0.0	0.0%	0.0	0.0%	0.0	22.0%	4.1	0.0%	0.0
VOC2007 classification	20	20	0.0%	0.0	75.0%	650.0	**85.0**%	14721.6	**85.0**%	19934.8	**85.0**%	31448.8
Describable Textures	47	47	0.0%	0.0	4.3%	650.0	14.9%	8.9	27.7%	53.2	36.2%	181.7
Oxford-IIIT Pets	37	53	5.4%	1300.0	13.5%	650.0	10.8%	80.9	**64.9**%	134.0	37.8%	169.0
Caltech-101	102	122	24.5%	1300.0	43.1%	650.0	**66.6**%	1633.8	**83.3**%	5249.7	**87.3**%	5017.7
Oxford Flowers 102	102	147	10.0%	1300.0	40.2%	650.0	17.6%	53.2	**50.0**%	194.3	**65.7**%	422.7
MNIST	10	10	0.0%	0.0	0.0%	0.0	40.0%	0.8	**100.0**%	46.0	**90.0**%	68.8
FER 2013 *	8	12	0.0%	0.0	8.3%	650.0	25.0%	5.9	41.7%	29.2	41.7%	11.5
STL10	10	10	0.0%	0.0	100%	650.0	**100.0**%	8778.6	**100.0**%	28547.6	**100.0**%	45587.5
GTSRB *	43	85	0.0%	0.0	0.0%	0.0	2.3%	12.7	2.3%	52.9	2.3%	551.3
PatchCamelyon	2	6	0.0%	0.0	0.0%	0.0	0.0%	0.0	**50.0**%	143.0	**50.0**%	15.0
UCF101 *	101	153	0.0%	0.0	0.0%	0.0	0.0%	0.0	**51.5**%	66.4	0.0%	0.0
Hateful Memes	2	2	0.0%	0.0	0.0%	0.0	**50.0**%	79.5	**50.0**%	2742.5	**50.0**%	321.5
EuroSAT	10	20	0.0%	0.0	0.0%	0.0	20.0%	2946.6	30.0%	5266.3	30.0%	15458.7
Resisc45	45	59	8.9%	1300.0	26.7%	650.0	**71.1**%	3688.6	**75.6**%	7572.0	**80.0**%	26317.6
Rendered-SST2	2	2	0.0%	0.0	**50.0**%	650.0	**50.0**%	1.0	**100.0**%	114.0	**100.0**%	1259.0

Table 6. Statistics of concept coverage between training and validation dataests. * indicates dataset whose train/test size we obtained is slightly different from Table 9 in [5]. "Cover." denotes the coverage ratio of concepts in target dataset by the training dataet. For those with non-zero coverage ratio, we also list the average number of images for each concept. For ImageNet-1K and ImageNet-21K, we estimate the number of images per concept by dividing the total number of images by total number of concepts, which are 1300 and 650, respectively

close neighbor. Similarly, the left model mix "goblet" and "coverlet", probably because they share the same suffix. Our model on right side finds one of the most matched concepts "liqueur glass" which is semantically and visually similar to the query concept. Similar trend is also observed in Fig. 4. All these visualizations demonstrate that our model trained with both type of data has learned the visually-grounded semantic meanings for various concepts.

References

[1] 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. 1

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 3

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, 2021. 1, 2, 3

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3, 5

[6] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. 1

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[8] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[9] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 1

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1

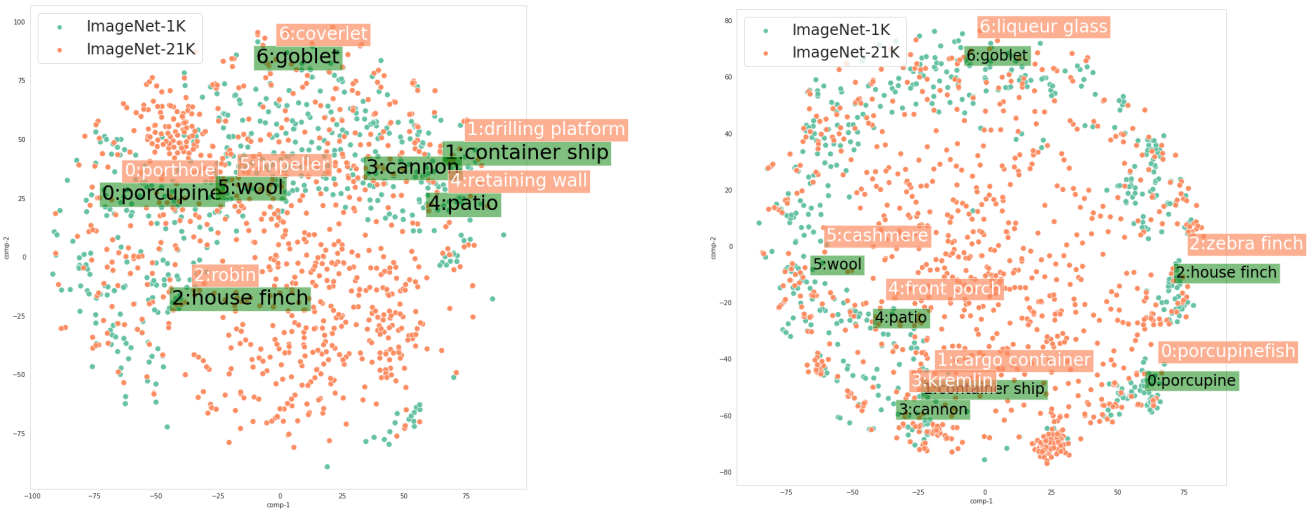

Figure 3. Similar to Fig. 4 in our main submission, we further visualize the t-SNE embedding for visual concepts with models trained with ImageNet-1K (left) and ImageNet-1K+GCC-15M (right).

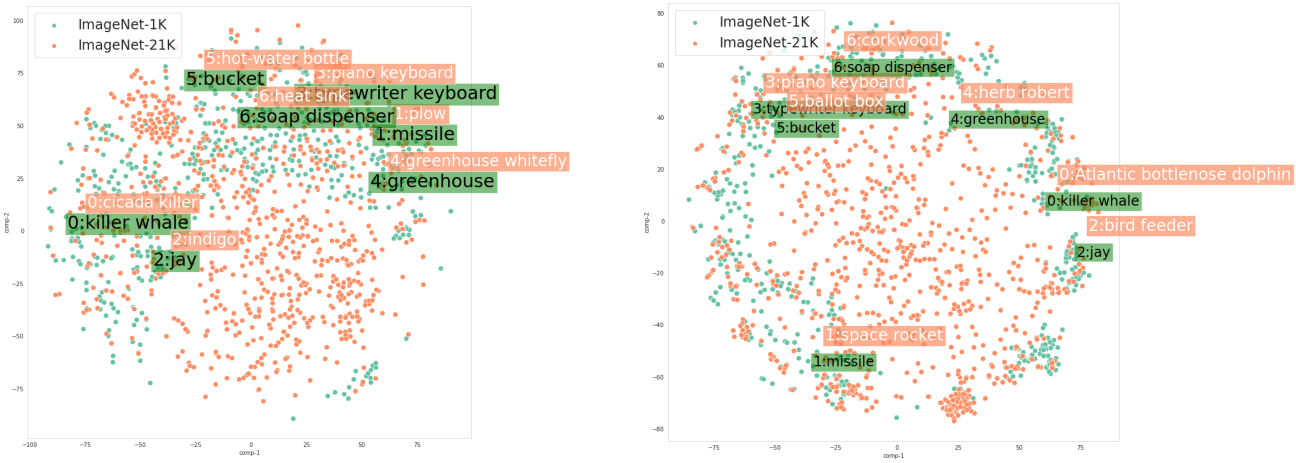

Figure 4. Similar to Fig. 3, we visualize the t-SNE embedding for another random set of visual concepts with models trained with ImageNet-1K (left) and ImageNet-1K+GCC-15M (right). Clearly, our model learned from the combination of image-label and image-text pairs can understand more number of visual concepts.