DATA DE-DUPLICATION AND SEMANTIC ENHANCEMENT FOR CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING-SUPPLEMENTARY

Anonymous authors
Paper under double-blind review

This appendix is organized as follows.

- In Section A, we describe the processing of our text augmentation.
- In Section B, we introduce implementation details of the our experiment.
- In Section C, we list the prompt templates used for zero-shot classification experiments.
- In Section D, we provide more experiments to explore components of DS-CLIP.
- In Section E, we offer more visualization results of augmented text for DS-CLIP.
- In Section F, we show the distribution of visual embedding space with CLIP and DS-CLIP on 10 classification datasets.

A DETAILS OF TEXT AUGMENTATION

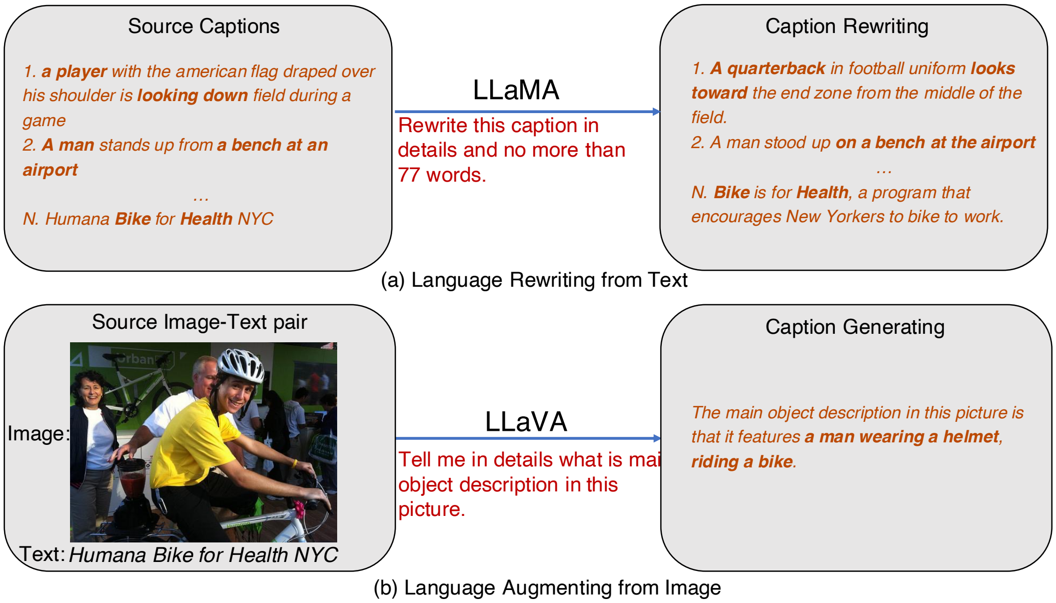
Figure 1 shows processing of text augmentation from LLaMA (Touvron et al. (2023)) and LLaVA (Liu et al. (2023)).

For LLaMA-based Text2Text rewritten, firstly, following LaCLIP (Fan et al. (2023)), we randomly select some texts from pre-training image-text datasets and then generate modified text through ChatGPT (Sardana et al.) as follows. A prompt “Rewrite this caption in detail and no more than 77 words.” is defined. The selected texts are then inputted into ChatGPT with the prompt to generate a refined caption. Secondly, we randomly sampled five origin-rewrite text pairs as instances. Given a text sample to be processed, the instances, the prompt “Rewrite this caption in details and no more than 77 words.” and an especial symbol “==>” are together inputted into LLaMA to generate refined text.

For LLaVA-based Image2Text generating, the processing is shown as follows. Given an image sample to be generated text, firstly, we design a sentence prompt that informs LLaVA about the task of generating image descriptions, such as “Tell me in detail what is main object description in this picture.”. This serves as an initial contextual clue for the LLaVa to understand the objective at hand. Then, the prompt combined with the corresponding image further enables the LLaVA model to generate more semantic texts.

B IMPLEMENTATION DETAILS

Model Architecture. The proposed framework is a general dual-encoder architecture, which consists of an image encoder and a text encoder similar to the origin CLIP (Radford et al. (2021)). By default, the image encoder is a random initialized ViT-B/32 (Dosovitskiy et al. (2020)), which has 12 layers of transformer blocks. Each transformer block has 12 attention heads and the hidden size is set to 768. The text encoder is BERT-base (Devlin et al. (2019)), which shares a similar scale with the image encoder, *e.g.* 12 layers of transformer blocks, 12 attention heads in each transformer block, and 768 hidden sizes. The output of the text encoder and image encoder is projected to 384-dim by linear projection. We adopt the unsupervised model DINO-S/8 (Caron et al. (2021)) as the image encoder to pre-compute image features for clustering. We truncate the input text tokens so that they have a maximum length of 77. The input image is resized to 224×224, and results in 7×7 image patches when the patch size is 32×32 for ViT-B/32.


Source Captions
1. a player with the american flag draped over his shoulder is looking down field during a game
2. A man stands up from a bench at an airport
...
N. Humana Bike for Health NYC
LLaMA
Rewrite this caption in details and no more than 77 words.
Caption Rewriting
1. A quarterback in football uniform looks toward the end zone from the middle of the field.
2. A man stood up on a bench at the airport
...
N. Bike is for Health, a program that encourages New Yorkers to bike to work.
Source Image-Text pair
Image:
Text: Humana Bike for Health NYC
LLaVA
Tell me in details what is mai object description in this picture.
Caption Generating
The main object description in this picture is that it features a man wearing a helmet, riding a bike.

(a) Language Rewriting from Text

(b) Language Augmenting from Image

Figure 1: Text Augmentation through LLM and VLLM.

Dataset and Downstream Tasks. We have 13 widely used downstream datasets: ImageNet (Deng et al. (2009)), ImageNetV2 (Recht et al. (2019)), CIFAR10 (Krizhevsky et al. (2009)), CIFAR100 (Krizhevsky (2009)), Caltech101 (Fei-Fei et al. (2005)), Oxford Pets (Parkhi et al.), SUN397 (Xiao et al. (2014)), Food101 (Bossard et al. (2014)), DTD (Cimpoi et al. (2013)), Stanford Dogs (Khosla et al. (2011)), COCO (Chen et al. (2015)), ADE20K (Zhou et al. (2019)), Flickr30K (Young et al. (2014)) and MSRVTT (Xu et al. (2016)). Table 1 summarizes the details of these datasets. For ImageNetV2, we use the same training data of ImageNet for the linear probe classification experiments. For COCO, Flicker30K, and MSRVTT for the image-text and video-text retrieval task, we only evaluate the zero-shot top-K retrieval items, thus we don't need training data. The pre-trained model is used to extract embeddings from images, videos, and texts separately. Similarity scores between image/video embeddings and text embeddings are used for ranking. We use the R@K to report the recall of top-K retrieval items. There are 30,000 images and each image has corresponding five text on Flickr30K. COCO and ADE20K are used for detection and segmentation evaluation. The classification datasets use classification accuracy as an evaluation metric, except for Caltech101 and Oxford Pets, which use averaged per-class accuracy. The detection datasets use mean average precision as the evaluation metric. The segmentation datasets use mean Intersection over Union(mIoU) as an evaluation metric.

Table 1: Details of downstream datasets.

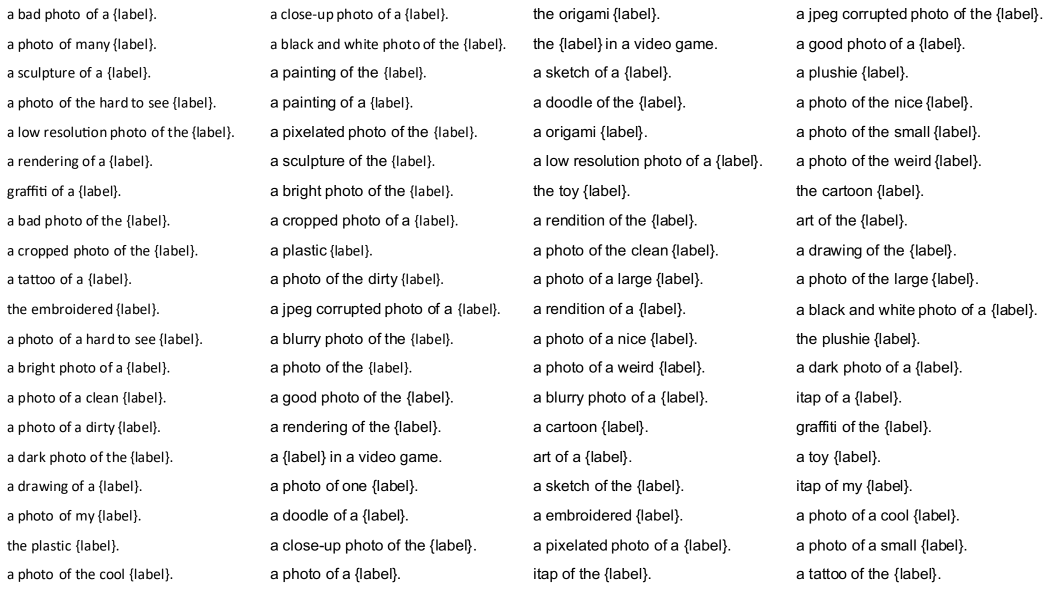
Dataset	**#Classes**	**#Train**	**#Test**	**Metric**
ImageNet	1,000	1,281,167	50,000	accuracy
ImageNetV2	1,000	–	50,000	accuracy
CIFAR10	10	50,000	10,000	accuracy
CIFAR100	100	50,000	10,000	accuracy
Caltech101	102	3,060	6,085	mean-per-class
Oxford Pets	37	3,680	3,669	mean-per-class
SUN397	397	19,850	19,850	accuracy
Food 101	102	75,750	25,250	accuracy
DTD	47	3,760	1,880	accuracy
Stanford Dogs	120	12,000	8,580	accuracy
COCO	81	118287	5,000	mean average precision
ADE20K	150	25574	2000	mean Intersection over Union
Filcker30K	–	–	30000	recall of top-K retrieval item
MSRVTT	–	–	1000	recall of top-K retrieval item

Training Settings. We implement our framework with Pytorch (Paszke et al. (2019)). All pre-training experiments with ViT-B are conducted on 8 NVIDIA Tesla A100 GPUs, and experiments with ViT-L are conducted on 96 NVIDIA Tesla A100 GPUs. The batch size is set to 4096. The framework is trained for 32 epochs with LAMB optimizer (You et al. (2020)) and an initial learning rate of 2.5e-3. The learning rate follows a cosine decay schedule with 5 epochs of linear warm-up. Weight decay is set to 0.2. For data augmentation, we random crop a 224×224 patch from the input image, then conduct random horizontal flip, random color distortions, random Gaussian blur, and RandAugment (Cubuk et al. (2020)), following previous works (Mu et al. (2022); You et al. (2022); Radford et al. (2021)). α and β are equal to 1 in training loss.

Training Details of Linear Probe. We freeze the pre-trained image encoder and append a linear classifier after it for linear probe classification. During training, we apply data augmentation to the input image. Concretely, we randomly crop a 224×224 patch from the input image, then conduct a random horizontal flip. During testing, we resize the shorter size to 224 then center crop a 224×224 patch as input image. We train the classifier for 90 epochs except for the ImageNet dataset, for which we train 10 epochs in total due to the large data volume. The learning rate follows a cosine decay schedule with an initial learning rate equal to 0.1. We use SGD with momentum for optimization. Weight decay is not used in our experiments. The batch size is set to 128.

C Prompt Engineering

Following previous works (Radford et al. (2021); Mu et al. (2022)), we extend the category names into sentences with prompts such as "a photo of {label}." before feeding them into the text encoders. For a fair comparison, we adopt the same prompts used in CLIP (Radford et al. (2021)). Specifically, for Oxford Pets, we use "a photo of a {label}, a type of pet.", while for the Food101 dataset, we use "a photo of a {label}, a type of food.". For the other datasets, we use 80 prompt templates as shown in Figure 2. For a given category name, we average the embeddings of different prompted sentences and conduct L2 normalization to obtain the final category embedding.



a bad photo of a {label}.
a photo of many {label}.
a sculpture of a {label}.
a photo of the hard to see {label}.
a low resolution photo of the {label}.
a rendering of a {label}.
graffiti of a {label}.
a bad photo of the {label}.
a cropped photo of the {label}.
a tattoo of a {label}.
the embroidered {label}.
a photo of a hard to see {label}.
a bright photo of a {label}.
a photo of a clean {label}.
a photo of a dirty {label}.
a dark photo of the {label}.
a drawing of a {label}.
a photo of my {label}.
the plastic {label}.
a photo of the cool {label}.

a close-up photo of a {label}.
a black and white photo of the {label}.
a painting of the {label}.
a painting of a {label}.
a pixelated photo of the {label}.
a sculpture of the {label}.
a bright photo of the {label}.
a cropped photo of a {label}.
a plastic {label}.
a photo of the dirty {label}.
a jpeg corrupted photo of a {label}.
a blurry photo of the {label}.
a photo of the {label}.
a good photo of the {label}.
a rendering of the {label}.
a {label} in a video game.
a photo of one {label}.
a doodle of a {label}.
a close-up photo of the {label}.
a photo of a {label}.

the origami {label}.
the {label} in a video game.
a sketch of a {label}.
a doodle of the {label}.
a origami {label}.
a low resolution photo of a {label}.
the toy {label}.
a rendition of the {label}.
a photo of the clean {label}.
a photo of a large {label}.
a rendition of a {label}.
a photo of a nice {label}.
a photo of a weird {label}.
a blurry photo of a {label}.
a cartoon {label}.
art of a {label}.
a sketch of the {label}.
a embroidered {label}.
a pixelated photo of a {label}.
itap of the {label}.

a jpeg corrupted photo of the {label}.
a good photo of a {label}.
a plushie {label}.
a photo of the nice {label}.
a photo of the small {label}.
a photo of the weird {label}.
the cartoon {label}.
art of the {label}.
a drawing of the {label}.
a photo of the large {label}.
a black and white photo of a {label}.
the plushie {label}.
a dark photo of a {label}.
itap of a {label}.
graffiti of the {label}.
a toy {label}.
itap of my {label}.
a photo of a cool {label}.
a photo of a small {label}.
a tattoo of the {label}.

Figure 2: The prompt templates used for zero-shot classification.

D More Experiments

In this section, we further explore the impact of different hyper-parameters and design choices of DS-CLIP with more experiments.

Table 2: Different design choices of D^3 and SE. All models are trained on YFCC15M and evaluated on zero-shot ImageNet classification.

(a) Clustered features.

Clustered Feature Type	Top-1
Image	**34.6**
Text	32.8

(b) Ablating sample type.

Sample Type	Top-1
Weighted	34.3
Sorted	31.6
Uniform	**34.6**

(c) Ablating SE on multi-scale sample data.

Number	SE	Top-1	Number	SE	Top-1
10%	✗	18.3	50%	✗	34.6
	✓	**22.5**		✓	**37.9**
30%	✗	28.6	70%	✗	37.6
	✓	**30.4**		✓	**40.2**

(d) Ablating sample number.

Sample Number	Top-1
10%	18.3
30%	28.6
50%	34.6
70%	**37.6**

(e) Ablating clustered number.

Clustered Prototypes Number	Top-1
1000	34.2
10000	34.6
100000	**34.7**

(f) Ablating text augmentation.

LLaVA	LLaMA	Top-1
✗	✗	34.6
✗	✓	36.5
✓	✗	37.6
✓	✓	**37.9**

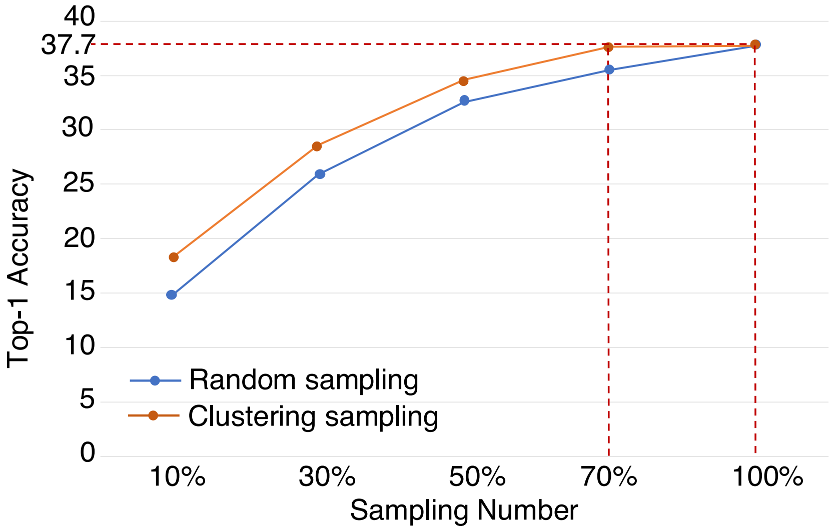
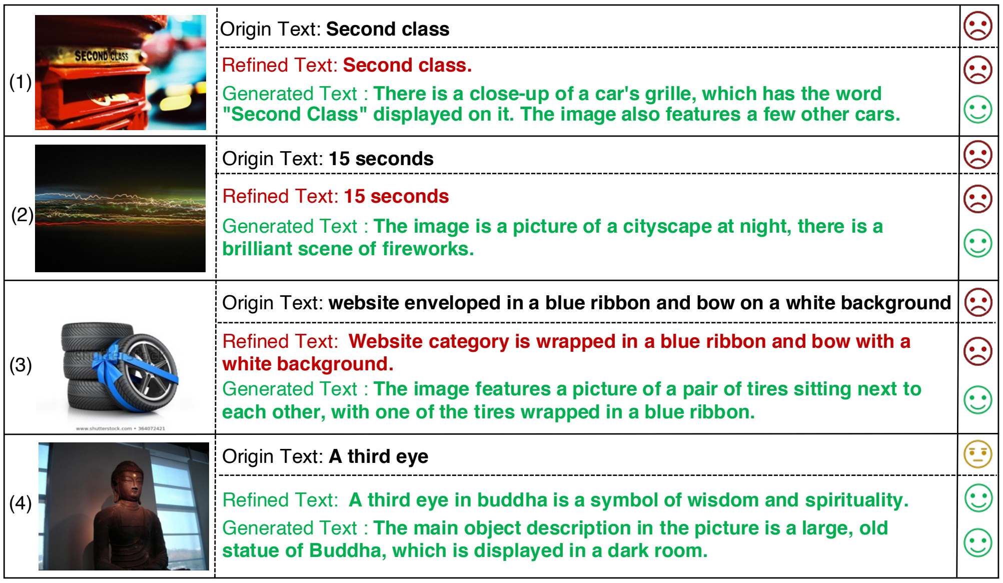


Figure 3: Zero-shot classification performance on ImageNet with different sampling strategies.

Different clustered features and numbers in pre-training. We cluster image-text pairs by different features (image feature and text feature) as references. Specifically, we utilize unsupervised model DINO-S/8 (Caron et al. (2021)) to extract image features for clustering, and adopt sentence transformer (Reimers & Gurevych (2019)) as the text encoder to encode text respectively. Experimental results are shown in Table 2a, CAM provides stable performance with different features for clustering, but the model with the clustering feature of the image performs better than the other.

Clustering number is another important factor, we explore three experiments of different clustering numbers. We observe a small improvement in classification accuracy as the number of clusters increased(34.2%, 34.6%, 34.7%), which had little impact as in Table 2e. We choose the number of clustering (10000) as default.

Different sampling type and number in pre-training. We further test different designs of sampling strategy as shown in Table 2b and 2d. For Weighted, we first calculate the number of samples for each clustered category, then weighted sampling regulation by the open square proportion of the number of samples is executed (Have (2003)). For Sorted, a perspective is that image-text pairs with higher similar scores are more robust than those with lower similar scores. So we first compute the similar score of each image-text pair by extracting features from the CLIP and then choose the sample that has higher scores from the center of each cluster with corresponding proportion. We also report our default setting Uniform, we uniformly select the corresponding proportion of samples



(1)	Origin Text: **Second class**	☹
	Refined Text: **Second class.**	☹
	Generated Text : **There is a close-up of a car's grille, which has the word "Second Class" displayed on it. The image also features a few other cars.**	☺
(2)	Origin Text: **15 seconds**	☹
	Refined Text: **15 seconds**	☹
	Generated Text : **The image is a picture of a cityscape at night, there is a brilliant scene of fireworks.**	☺
(3)	Origin Text: **website enveloped in a blue ribbon and bow on a white background**	☹
	Refined Text: **Website category is wrapped in a blue ribbon and bow with a white background.**	☹
	Generated Text : **The image features a picture of a pair of tires sitting next to each other, with one of the tires wrapped in a blue ribbon.**	☺
(4)	Origin Text: **A third eye**	😑
	Refined Text: **A third eye in buddha is a symbol of wisdom and spirituality.**	☺
	Generated Text : **The main object description in the picture is a large, old statue of Buddha, which is displayed in a dark room.**	☺

Figure 4: Samples of image-text pair with misalignment problem and generated texts from LLaMA and LLaVA.

for one epoch, and then for the next epoch, the corresponding proportion of samples are sampled from the rest in turn. We observe that all these variants produce similar results except SORTED as in Table 2b. The SORTED method selects the same samples with higher scores at each epoch, but samples of other methods are different at each epoch. Diversity of data is fully guaranteed except SORTED method.

We also test different sampling numbers during pre-training. As the number of samples increases, the performance of the model increases as in Table 2d.

Clustering with different sample ratios. Moreover, by utilizing different sampling ratios, the model trained according to the clustered sampling strategy outperforms the model of random sampling as shown in Figure 3. We can attribute this better performance to the diversity of samples maintained in the training epoch with the clustered dataset.

E MORE VISUALIZATION

In this section, we provide more visualization results of augmented texts. As shown in Figure 4, most of the origin samples are inconsistent between image and text. The origin text of sample (1) is the “word” on the image, the origin texts of sample (2) and sample (3) do not describe the main content of the image. The origin text of sample (4) describes the image as incomplete. The refined text based on LLaMA from samples (1) to (3) is inaccurate but the refined text in the sample (4) is a complete supplement. However, all generated texts from LLaVA can accurately describe the content of the image.

Although the augmented text from Text2Text rewriter sometimes may not return texts that provide descriptions about the related-image contents, the proposed DS-CLIP is also robust to feature representation learning. Since the proposed DS-CLIP not only depends on the origin text and augmented texts from LLaMA but also takes generated texts from LLaVA into consideration, it is more robust than origin CLIP (Radford et al. (2021)) trained with only one image-text pair.

F DISTRIBUTOIN OF VISUAL EMBEDDING SPACE BY T-SNE

To gain a deeper understanding of the distinctions between the features learned from DS-CLIP and vanilla CLIP, we provide distribution of visual embedding space with CLIP and DS-CLIP on 10 classification datasets using t-SNE (Maaten & Hinton (2008)) in Figure 5. We visualize the

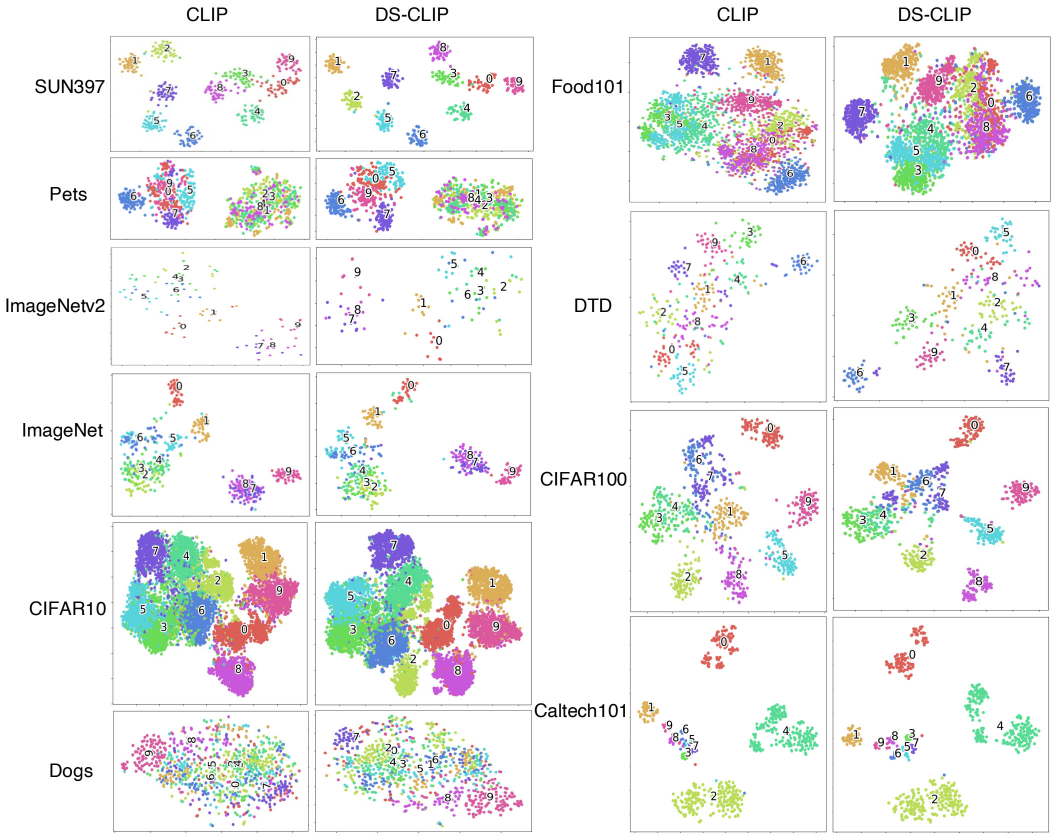


Figure 5: Distribution of visual embedding space with CLIP and DS-CLIP on 10 classification datasets using t-SNE.

features from the first 10 classes for each dataset. From Figure 5, we observe that DS-CLIP trained with augmented texts has more distinguished boundaries and is more compact for each class. This observation suggests that augmented text helps learning an effective image embedding space which is also well-suited for downstream tasks.

References

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. *Food-101 – Mining Discriminative Components with Random Forests*, pp. 446–461. Jan 2014. doi: 10.1007/978-3-319-10599-4_29. URL `http://dx.doi.org/10.1007/978-3-319-10599-4_29`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021. doi: 10.1109/iccv48922.2021.00951. URL `http://dx.doi.org/10.1109/iccv48922.2021.00951`.

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

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. Jan 2013.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. pp. 702–703, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2009. doi: 10.1109/cvpr.2009.5206848. URL `http://dx.doi.org/10.1109/cvpr.2009.5206848`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019. doi: 10.18653/v1/n19-1423. URL `http://dx.doi.org/10.18653/v1/n19-1423`.

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. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Oct 2020.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023.

Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, Apr 2005. doi: 10.1109/cvpr.2004.383. URL `http://dx.doi.org/10.1109/cvpr.2004.383`.

How Much Confidence Do We Have. The square root of n plus one sampling rule. *Pharmaceutical Technology*, 2003.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Jan 2009.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Laurensvander Maaten and GeoffreyE. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research,Journal of Machine Learning Research*, Jan 2008.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. *SLIP: Self-supervision meets Language-Image Pre-training*, pp. 529–544. Jan 2022. doi: 10.1007/978-3-031-19809-0_30. URL `http://dx.doi.org/10.1007/978-3-031-19809-0_30`.

OmkarM Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. pp. 8024–8035, 2019.

Alec Radford, JongWook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Cornell University - arXiv,Cornell University - arXiv*, Feb 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? volume 97, pp. 5389–5400, 2019.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Jan 2019. doi: 10.18653/v1/d19-1410. URL `http://dx.doi.org/10.18653/v1/d19-1410`.

Divesh Sardana, TimothyR Fagan, JohnTimothy Wright, Sardana Sardana, and JamesW Bawden. Chatgpt.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jianxiong Xiao, James Hays, Antonio Torralba, KristaA. Ehinger, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *Springer US,Springer US*, Aug 2014.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *European Conference on Computer Vision*, pp. 69–87. Springer, 2022.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. 2020.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, pp. 67–78, Dec 2014. doi: 10.1162/tacl_a_00166. URL `http://dx.doi.org/10.1162/tacl_a_00166`.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.