DH-Set: Improving Vision-Language Alignment with Diverse and Hybrid Set-Embeddings Learning
Supplementary Material

Kun Zhang[1,2] Jingyu Li[3] Zhe Li[4] S.Kevin Zhou[1,2,5,6] *
[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, USTC
[2] MIRACLE Center, Suzhou Institute for Advance Research, USTC
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[4] School of Cyberspace Science and Technology, USTC
[5] Key Laboratory of Intelligent Information Processing of CAS, ICT, CAS
[6] State Key Laboratory of Precision and Intelligent Chemistry, USTC
{kkzhang@, jingyuli@iai., lizhe777@mail.}ustc.edu.cn, s.kevin.zhou@gmail.com

1. Experimental Investigation of Semantic Representation Tendencies of Dimensions

Here, we describe the details of experiments to investigate the relationship between local dimensions and the specific semantics in the representation space, using our trained state-of-the-art model DH-Set under the Faster R-CNN+BERT backbone.

First of all, let us recall the general definition of semantic similarity of cross-modal features. For arbitrary textual feature $\boldsymbol{u} = \{a_i\}_{i=1}^{d} \in \mathbb{R}^{1\times d}$ and visual feature $\boldsymbol{v} = \{b_i\}_{i=1}^{d} \in \mathbb{R}^{1\times d}$, where a_i and b_i are the i-th local dimension in the d-dimensional space, existing methods typically aggregate all dimensional correspondence to reflect their semantic similarity, *i.e.*, $\sum_{i=1}^{d} s_i$, where s_i can be determined by the product of scalars a_i and b_i in the inner product operation.

Thus, for a textual-visual pair with aligned semantics, its semantic similarity is determined by the sum of all cross-modal dimensional correspondence s_i, where the larger s_i is, the greater the contribution of the i-th dimension to this semantic. That is, the i-th local dimension in the representation space is more inclined to describe this semantic.

Based on the above analysis, we take all word-region pairs with the same semantics, *e.g.*, 'man' (a total of 30,275 pairs), on the Flickr30K training dataset. we first obtain all the cross-modal semantic correspondence vectors, and collect the dimension index sets of the largest top-100 $\{s_i\}$ in each vector, *i.e.*, the dimensions most significantly reflect the semantics of 'man'. Then, we count the co-occurrence probabilities of each dimension and sort the probabilities of all dimensions from large to small, as depicted in Fig. 1(a), where we can find that some dimensions (marked with colors) have obvious tendencies to represent the semantic concept of 'man'. Similarly, as shown in Fig. 1(b) and Fig. 1(c), we show the corresponding results for the semantic concept 'woman' (a total of 15,753 pairs) and 'dog' (a total of 6,104 pairs). From the results, we can see that semantics is indeed mainly represented by some key local dimensions in the representation space.

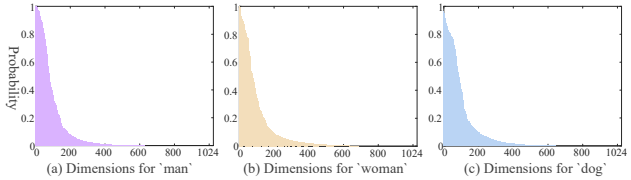

Figure 1. Visualization of the finding that specific semantics is mainly expressed by some key local dimensions (marked in colors) in the representation space.

2. Learnable Truncation Threshold

For each vector $\mathbf{w}_j^i \in \mathbb{R}^{1\times d}$ of the learnable matrix $\mathbf{W}^i$, we design an adaptive sparsity strategy. Specifically, we first learn a truncation threshold based on the value range of the learnable weights:

$$t_j^i = \min(\mathbf{w}_j^i) + (\max(\mathbf{w}_j^i) - \min(\mathbf{w}_j^i)) \cdot \text{sigmoid}(\alpha_i), \quad (1)$$

where α_i is a learnable parameter; $\min(\cdot)$ and $\max(\cdot)$ are operations for obtaining minimum and maximum values respectively.

*Corresponding author.

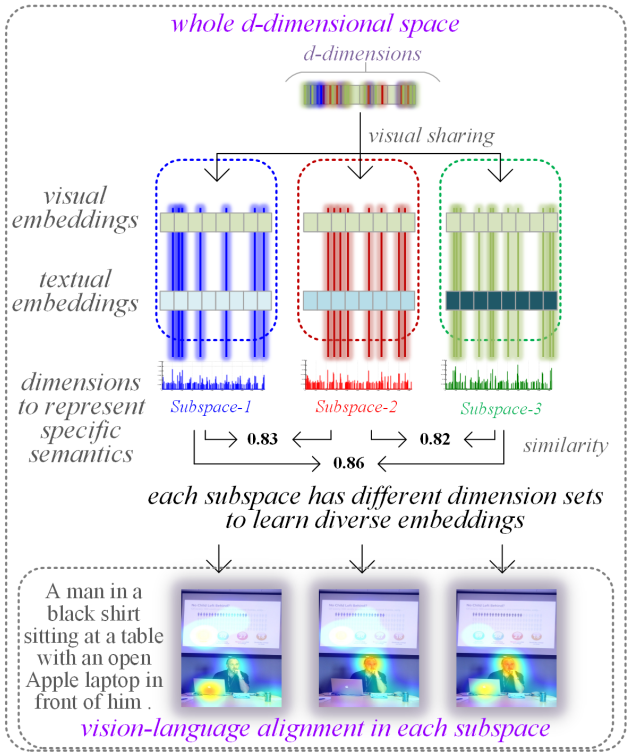

Figure 2. Analysis of why shared visual features can model semantic diversity.

In order to maintain the propagation of gradients in sparse truncation, we propose an adaptive truncation mask based on threshold t_j^i as:

$$mask_j^i = \tanh(e^{\lambda(\mathbf{w}_j^i - t_j^i)}), \tag{2}$$

where λ is a large scaling constant that keeps masks of items larger than the threshold and discards masks of items smaller than the threshold. Finally, we apply mask $mask_j^i$ to the weight $\mathbf{w}_j^i$ to ensure sparsity, *i.e.*, $mask_j^i \odot \mathbf{w}_j^i$.

3. Why shared visual features can model semantic diversity?

In our work DH-Set, we propose to share visual representations that have semantic distribution differences due to input variance, and model semantic diversity under the guidance of a set with different textual embeddings. Therefore, a question will be asked: Why can shared visual features model semantic diversity? As shown in Fig.2, we analyze the reason that in the whole d-dimensional representation space, we constrain the local dimensions with important semantic representation tendencies in each subspace. As shown in the colored indicator bars in Fig.2 (for clarity, we only draw the part with a dimension probability greater than 0.015), different sets of dimensions constitute the representation of specific semantics in the subspace. As a result, even for shared d-dimensional visual features, different subsets of dimensions within it also characterize diverse semantics. Finally, it is possible to capture the diverse alignment in Vision-Language (VL), as the example shown at the bottom of Fig.2, where different subspaces can focus on diverse visual content to resolve semantic ambiguity in VL alignment.

Table 1. Comparison of the overall running time.

DH-Set	Running Time($\times 10^{-6}s$)					Average
w/o $\mathbf{sim}_{hybrid}$	4.64	4.71	4.85	4.67	4.76	4.726±0.0055
w $\mathbf{sim}_{hybrid}$	3.39	3.44	3.32	3.43	3.36	3.338±0.0019

4. More Visualization Cases

Here, we give more visualization of the comparison of VL alignment between our DH-Set and the existing method that serves as a baseline without semantic diversity modeling, as shown in Fig.3. Besides, we also give more visualization of VL alignment in different subspaces, as shown in Fig.4, where we can see that different subspaces show diversity to align different visual-linguistic content.

5. Time-Consuming Comparison

As analyzed in the paper, the computational complexity of our proposed set similarity is reduced by a factor of k compared to the original method (k is the size of the embedding set, which is greater than or equal to 3 in our method). As shown in Tab.1, we compare the consumption of the overall running time per image-text pair with and without hybrid inference. We can find that the proposed method can significantly reduce the computational time, with an average relative reduction of 28.3%, which proves the efficiency of the design of our DH-Set.

6. Details of Feature Extraction

ResNet152+BiGRU. Following the existing method [14], we employ the ResNet-152 [7] pre-trained on ImageNet to encode the input image, where we apply average pooling local features and feed the output to one fully-connected layer to obtain global features. The text is encoded by the BiGRU [2] with the pre-trained GloVe vectors [11].

Faster R-CNN+BiGRU/BERT. The Faster R-CNN model [13] in conjunction with ResNet-101, which is pre-trained on Visual Genome [8], is adopted to detect objects and other salient regions. Top-K (K=36) local regions are selected for each image, and we obtain theirs as the mean-pooled convolutional feature with 2048 dimensions. Then a fully connected layer is used to transform each region into final features. The textual encoder uses more advanced BERT [4], which is a transformer-based model pre-trained

(a) A man sitting on a statue outdoors , near *a woman standing* next to him .

(b) A man talking to a *family unit* holding some sort of *probing device* as they look on and *smile* to be polite .

(c) A man in blue overalls is *sawing wood* .

(d) *Three people* walk past a tourist attraction behind a wrought iron fence .

(e) *Two man sit* and chat on a stone lanai surrounded by plants .

Figure 3. Comparison of VL alignment between our DH-Set and the baseline without semantic diversity modeling. This verifies that the proposed DH-Set can align cross-modal details (marked in purple) more accurately and comprehensively, demonstrating its superiority.

Table 2. Comparison on CUB 200-2011 Dataset.

		Swin-224+BERT							Swin-384+BERT						
CUB 200-2011 test set	Methods	I → T			T → I			rSum	I → T			T → I			rSum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
	LAPS$_{'24}$[12]($SOTA$)	69.8	92.5	96.7	53.7	78.3	87.2	478.2	72.9	94.8	97.8	59.0	82.3	89.8	496.6
	Our DH-Set	**71.1**	**93.4**	**97.4**	**58.4**	**82.2**	**89.7**	**492.2**	**74.6**	**95.5**	**98.3**	**62.2**	**85.3**	**91.7**	**507.8**

on large-scale Wikipedia and Bookcorpus. We add a fully-connected layer for the last layer of pre-trained BERT to obtain word features.

Swin Transformer (Swin)+BERT. The visual encoder is implemented by Swin Transformer [10]. The image is partitioned into non-overlapping patches based on the spatial distribution. Subsequently, we feed these patches as a visual token sequence into the vision transformer, which consists of multiple self-attention layers. We use the image resolutions 224×224 or 384×384. The textual encoder is employed by BERT [4] as the above.

Vision Transformer (ViT)+BERT. The visual encoder exploited Vision Transformer (ViT) [5] directly uses the image patches as inputs. Moreover, built on the pre-trained vision-language models [9, 12], we use the output tokens from the ViT encoder and the BERT encoder as visual and textual features, respectively.

7. Adapt to Existing Holistic VL Alignment or Fragmental VL Alignment

It is worth noting that our proposed DH-Set can be adapted to the existing holistic visual-language (VL) alignment [12], which represents images and texts as global features for similarity calculation, as well as the existing fragment visual-language alignment [1, 6, 14], which calculates similarity based on the fine-grained features of visual regions/patches and textual words.

Specifically, under the holistic VL alignment paradigm, the input visual and textual data are represented as global feature sets $\{v_i^g\}_{i=1}^k \in \mathbb{R}^{k\times d}$ and $\{u_i^g\}_{i=1}^k \in \mathbb{R}^{k\times d}$, respectively. During training, the similarity is calculated according to Eq.10 in the paper, while during testing, it is calculated according to Eq.12 in the paper.

Under the fragmental VL alignment paradigm, the input visual and textual data are represented as local feature sets. For example, each word is represented as $\{u_{i,p}^l\}_{i=1,p=1}^{k,m} \in \mathbb{R}^{k\times d}$ and each region/patch is represented as $\{v_{i,q}^l\}_{i=1,q=1}^{k.n} \in \mathbb{R}^{k\times d}$, where m and n denote the number of words and regions/patches, respectively. First, for each word, we calculate its attention weight with all image regions/patches using Eq.12 and obtain the weighted aggregated visual embeddings set $\{\hat{v}_{i,p}^l\}_{i=1,p=1}^{k,m} \in \mathbb{R}^{k\times d}$. Therefore, for each word's embeddings set $\{u_{i,p}^l\}_{i=1,p=1}^{k,m}$ and the corresponding aggregated visual embeddings set $\{\hat{v}_{i,p}^l\}_{i=1,p=1}^{k,m}$ with aligned semantics, we calculate the similarity of each word through Eq.10 during training and

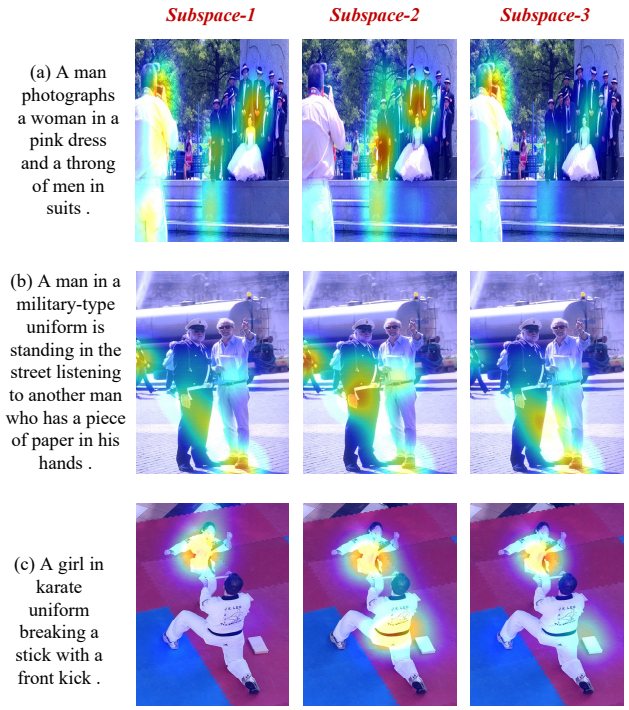

Figure 4. Visualization of VL alignment in different subspaces, where polysemy can be captured in diverse representation spaces.

through Eq.12 during testing. The final image-text alignment is determined by the average of all word similarities.

8. Comparison on CUB 200-2011 Dataset.

Following the earlier set-based method PCME[3] evaluated via CUB 200-2011 (having 11,788 images of 200 fine-grained bird categories, per image with ten captions), a harder and more reliable benchmark [3]. Following the prior train/test split (150/50 classes), we replicate the existing SOTA baseline LAPS [6] and our DH-Set (same training details as in paper, *e.g.*, 30 epochs), as shown in Tab. 2, equipped with our DH-Set, rSum outperforms >10%, verifying its superiority.

References

[1] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proc. Conf. Comput. Vis. and Pattern Recognit.*, pages 15789–15798, 2021. 3

[2] Kyunghyun Cho. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[3] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8415–8424, 2021. 4

[4] 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. 2, 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[6] Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 26307–26316, 2024. 3, 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. Conf. Comput. Vis. and Pattern Recognit.*, pages 770–778, 2016. 2

[8] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *Int. J. Comput. Vis.*, pages 32–73, 2017. 2

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[10] 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. In *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 3

[11] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proc. Conf. Empirical Methods in Natural Language Process.*, pages 1532–1543, 2014. 2

[12] 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. In *Proc. Int. Conf. Mach. Learn.*, pages 8748–8763, 2021. 3

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2016. 2

[14] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proc. Conf. Comput. Vis. and Pattern Recognit.*, pages 1979–1988, 2019. 2, 3