GeoMM: On Geodesic Perspective for Multi-modal Learning Supplemental Materials

1. Network Structure

Our model structure mainly follows [8], [13] and [2]. Taking [8] as an example, the network includes an image encoder, a text encoder, and a multimodal fusion encoder. The image encoder is a 12-layer transformer with a VIT structure, with initialized weights derived from pre-training on the ImageNet-1k [4] dataset. The text encoder and fusion encoder use pre-trained BERT-base [5] networks, which consist of 12 layers of transformers, with each encoder using 6 layers. To obtain more robust training with noisy web datasets, we maintain the momentum version for each encoder, i.e., $\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t$, where θ_t and θ'_t are the parameters of the main model E_I and the momentum modal E'_I, and α is the momentum parameter ranging between $[0, 1]$. Given a pair of images and text (I, T), following [6], we first perform two data augmentations on the image to obtain two different views of the image. Then, we encode the two augmented images with the original model and momentum model separately as positive pairs. The image tokens are obtained after the image patches are linearly mapped and position encoded. To capture global features, we also concatenate CLS tokens before visual tokens. We encode I and I' with E_I and E'_I respectively to obtain image embeddings, $V = \{V_{cls}, V_1, V_2, ..., V_m\}$ and $V' = \{V'_{cls}, V'_1, V'_2, ..., V'_m\}$. For the text, we tokenize and embed text following BERT[5], and we can similarly obtain text embeddings $W = \{W_{cls}, W_1, W_2, ..., W_n\}$ and $W' = \{W'_{cls}, W'_1, W'_2, ..., W'_n\}$. Subsequently, we concatenate visual and textual embeddings and feed them together into the fusion encoder for feature fusion, thereby learning the joint modal representations.

2. Pre-training Datasets

The details of the pre-training datasets about image-text pairs are shown in below Tab.1.

	COCO	VG	SBU	CC3M
#image	113K	100K	860K	2.95M
#text	567K	769K	860K	2.95M

Table 1. Statistics of the pre-training datasets.

Method	TR	IR
ALBEF	73.1	56.8
ALBEF+$\mathcal{OM}$	73.3	56.4
ALBEF+Geo	**76.2**	**59.2**
MAFA	78.0	61.2
MAFA+$\mathcal{OM}$	77.8	61.1
MAFA+Geo	**79.3**	**62.5**

Table 2. Comparison with Oblique manifold.

3. Pre-training Tasks

Multimodal learning requires elaborate pre-training tasks, and commonly used pre-training tasks include Masked Language Modeling(MLM) [5], Image-Text Matching (ITM) [8], Image-Text Contrastive (ITC) [11], Word Patch Alignment (WPA) [3], and so on. We leverage MLM, ITM, and ITC for multimodal pre-training following [3, 7, 8].

4. Extra Experiments

We also compare our method with the Oblique manifold ($\mathcal{OM}$) [1]. We conduct the experiments on the image-text retrieval task with the COCO dataset and fine-tune setting. We display the R@1 accuracy for text retrieval (TR) and image retrieval (IR), as shown in Tab.2.

5. Proofs

5.1. Proof for Theorem 1

For the graph where the cluster centers represent the vertices of the graph and the adjacent relationship between these cluster centers represents the edges between vertices, we can know this graph possesses N vertices with minimum degree κ (cluster neighbors).

We define all the vertices set as V, and then we construct a random subset X of V ($X \subset V$). Each sample in X is taken from V with a probability of p. Then the expectation scale of X is,

$$\mathbb{E}(|X|) = Np \tag{1}$$

We regard the subset X as the candidate for connected components $\mathcal{S}$. We can thus define the random set Y_X, which

represents the samples in $V - X$ that do not have an adjacent sample in X, that is, for sample $v \in Y_X$, we can not find a sample $x \in X$ that v is subordinate to x. This can also be interpreted as for $v \in Y_X$, any adjacent samples of v not in X, so

$$\begin{aligned} P(v \in Y_X) &= P(\text{v and its adjacent samples not in X}) \\ &= (1-p)^{1+d(v)} \\ &\leq (1-p)^{1+\kappa} \end{aligned} \tag{2}$$

Then we can obtain,

$$\mathbb{E}(|Y_X|) \leq N(1-p)^{1+\kappa} \tag{3}$$

It is apparent that $X \cup Y_X$ can be served as a connected component, and the number of connected components can be represented as,

$$\begin{aligned} \mathbb{E}(|X \cup Y_X|) &\leq \mathbb{E}(|X| + |Y_X|) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y_X|) \\ &= Np + N(1-p)^{1+\kappa} \leq Np + Ne^{-p(1+\kappa)} \end{aligned} \tag{4}$$

Since we want to find the minimal number of connected components, which means we want to find the minimal value of $Np + Ne^{-p(1+\kappa)}$. We can then obtain that when $p = \frac{\ln(\kappa+1)}{\kappa+1}$, the expectation get the minimum value,

$$\frac{N[1 + \ln(\kappa+1)]}{\kappa+1} \tag{5}$$

Therefore we can get

$$|\mathcal{S}| \leq \frac{N[1 + \ln(\kappa+1)]}{\kappa+1}, \quad \text{where} \quad \kappa = \mathcal{F}(\xi). \tag{6}$$

5.2. Proof for Theorem 2

The lower bound is obvious. Here we only prove the upper bound.

We can model the adjacency relationship between cluster centers as a bipartite graph with N points on both sides. The problem can be transformed into a $0-1$ matrix of $\mathcal{M} \in \mathbb{R}^{N \times N}$, where there is no all 1 sub-matrix of $\mathcal{M}_0 \in \mathbb{R}^{A \times A}$, and at this time, how many element 1 can there be in the matrix at most.

We count the following structures. We define that the left and right point sets of the bipartite graph are V_1 and V_2, and then assume that the structure p is selecting a point u from V_1 with a adjacent samples in V_2. Let's start with point u in V_1, and the selection methods of a samples in V_2 is $C^a_{d(u)}$, then the total selections are $\sum_{u \in V_1} C^a_{d(u)} = |S|$. We can also start with a samples in V_2. And once we determine a point in V_2, there are at most $a-1$ u in V_1, otherwise there will be an all 1 sub-matrix of $\mathcal{M}_0 \in \mathbb{R}^{A \times A}$. We can thus obtain,

$$\sum_{u \in V_1} C^a_{d(u)} \leq C^a_N (a-1) \tag{7}$$

Following Jensen Inequality [9], $f(x) = C^a_x$ is a convex function, then,

$$\sum_{u \in V_1} \frac{1}{N} C^a_{d(u)} \geq C^a_{\frac{1}{N}\sum_{u \in V_1} d(u)} = C^a_{\frac{|E|}{a}} \tag{8}$$

So,

$$\begin{aligned} N C^a_{\frac{|E|}{a}} &\leq \sum_{u \in V_1} C^a_{d(u)} \leq C^a_N (a-1) \\ &= \frac{N(N-1)...(N-a+1)}{a!}(a-1) \end{aligned} \tag{9}$$

Retraction is conducted at both sides and then,

$$N\frac{(\frac{|E|}{N} - a + 1)^a}{a!} < N C^a_{\frac{|E|}{a}} < \frac{N^a}{a!}(a-1) \tag{10}$$

After simplification, we can obtain,

$$E(N) \leq (a-1)^{\frac{1}{a}} N^{2-\frac{1}{a}} + (a-1)N \tag{11}$$

The proof of another upper bound is presented as follows,

Let the number of cluster centers be N and for every point x_i, the number of neighboring points is $d(x_i)$. Suppose a initial set $C_\pi = \varnothing$, for all points, we introduce a random permutation $\mathcal{O} : x_1, x_2, x_3, ..., x_n$. For a certain permutation, if all points in front of x_i are x_i's neighboring points, we put x_i into C_π. Finally, all point pairs in C_π are neighboring points.

The probability of a certain point in C_π is $p = \frac{1}{N - d(x_i)}$, then the mathematical expectation of the size of C is,

$$|C_\pi| = \sum_{x_i} \frac{1}{N - d(x_i)} \tag{12}$$

Suppose the size of maximal cluster is $\omega(D)$, we apply Pigeonhole Principle [12] and get:

$$\omega(\mathcal{D}) \geq \sum_{x_i} \frac{1}{N - d(x_i)} \tag{13}$$

What we need to satisfy is,

$$a \geq \omega(G) \geq \sum_{v_i} \frac{1}{N - d(v_i)} \tag{14}$$

According to Cauchy Inequality [10],

$$a \sum_{v_i} (N - d(v_i)) \geq \sum_{v_i} \frac{1}{N - d(v_i)} \sum_{v_i} (N - d(v_i)) \geq N^2 \tag{15}$$

So,

$$a(N^2 - 2|E|) \geq N^2 \tag{16}$$

We can thus obtain,

$$|E| \leq \frac{N^2}{2}(1 - \frac{1}{a-1}) \tag{17}$$

References

[1] E Andruchow, Gustavo Corach, and D Stojanoff. Geometry of oblique projections. *arXiv preprint math/9911133*, 1999. 1

[2] Jaeseok Byun, Dohoon Kim, and Taesup Moon. Mafa: Managing false negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27314–27324, 2024. 1

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1

[7] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1

[8] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021. 1

[9] Edward James McShane. Jensen's inequality. 1937. 2

[10] Dragoslav S Mitrinovic and Petar M Vasic. *Analytic inequalities*. Springer, 1970. 2

[11] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[12] Wojciech A Trybulec. Pigeon hole principle. *Journal of Formalized Mathematics*, 2(199):0, 1990. 2

[13] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 1