Supplementary Material for Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision

Chao Jia [1] **Yinfei Yang** [1] **Ye Xia** [1] **Yi-Ting Chen** [1] **Zarana Parekh** [1] **Hieu Pham** [1] **Quoc V. Le** [1] **Yunhsuan Sung** [1] **Zhen Li** [1] **Tom Duerig** [1]

1. Remove Near-Duplicate Test Images from Training Data

To detect near-duplicate images, we first train a separate high-quality image embedding model following (Wang et al., 2014) with a large-scale labeled dataset as in (Juan et al., 2020), and then generate 4K clusters via k-means based on all training images of the embedding model. For each query image (from the ALIGN dataset) and index image (from test sets of downstream tasks), we find their top-10 nearest clusters based on the embedding distance. Each image is then assigned to $\binom{10}{3}$ buckets (all possible combinations of 3 clusters out of 10). For any query-index image pair that falls into the same bucket, we mark it as near-duplicated if their embedding cosine similarity is larger than 0.975. This threshold is trained on a large-scale dataset built with human rated data and synthesized data with random augmentation.

2. Evaluation on SimLex-999

The image-text co-training could also help the natural language understanding as shown in Kiros et al. (2018). For instance, with language only, it is very hard to learn antonyms. In order to test this capability of ALIGN model, we also evaluate the word representation from ALIGN model[1] on SimLex-999 (Hill et al., 2015), which is a task to compare word similarity for 999 word pairs. We follow Kiros et al. (2018) to report the results on 9 sub-tasks each contains a subset of word pairs: *all, adjectives, nouns, verbs, concreteness quartiles (1-4)*, and *hard*.

The results are listed in the Table 1 compared to Picturebook (Kiros et al., 2018) and GloVe (Pennington et al., 2014) embeddings. Overall the learned ALIGN perform better than Picturebook but slightly worse than GloVe embeddings. What is interesting is that the ALIGN word embeddings has a similar trend of Picturebook embeddings, with better performance on *nouns* and *most concrete* categories but worse on *adjs* and *less concrete* categories compared to GloVe embeddings. ALIGN word embedding achieves the highest performance on the *hard* category, which similarity is difficult to distinguish from relatedness. This observation confirmed the hypothesis from Kiros et al. (2018) that image-based word embeddings are less likely to confuse similarity with relatedness than text learned distributional-based methods.

Table 1. SimLex-999 results (Spearman's ρ).

	GloVe	Picturebook	**ALIGN**
all	**40.8**	37.3	39.8
adjs	**62.2**	11.7	49.8
nouns	42.8	**48.2**	45.9
verbs	**19.6**	17.3	16.6
conc-q1	**43.3**	14.4	23.9
conc-q2	41.6	27.5	**41.7**
conc-q3	42.3	46.2	**47.6**
conc-q4	40.2	**60.7**	57.8
hard	27.2	28.8	**31.7**

[1] As ALIGN uses the wordpiece tokens, one word can be split into multiple pieces. We feed the wordpieces of a word into ALIGN model and use the [CLS] token representation before the project layers as the word embeddings.

References

Hill, F., Reichart, R., and Korhonen, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2015.

Juan, D.-C., Lu, C.-T., Li, Z., Peng, F., Timofeev, A., Chen, Y.-T., Gao, Y., Duerig, T., Tomkins, A., and Ravi, S. Graph-rise: Graph-regularized image semantic embedding. In *Proceedings of ACM International Conference on Web Search and Data Mining*, 2020.

Kiros, J., Chan, W., and Hinton, G. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the*

2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., and Wu, Y. Learning fine-grained image similarity with deep ranking. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2014.