Supplementary Material: Group Optimization for Multi-attribute Visual Embedding

In sections below, we represent details on algorithm analysis, comparisons with traditional methods and crowd data collection.

1. Details on Algorithm Analysis

As mentioned in the paper, our key idea is to collect and embed qualitative measures in groups. In a query, we ask a worker to classify $\mathbb{N}$ images into at most $\mathbb{B}$ bins/clusters $\{S_c\}$. We are interested to see how parameters in crowd query affect the multi-embedding accuracy. Besides, it is inevitable to collect noisy answers during a public data collection platform, even though we set several schemes to improve data quality. Therefore, we conduct a stress test to analyze the robustness of our algorithm under noises.

1.1. Stress Test under Noise

To test the robustness of our method to noise and erroneous measures we performed the following test. We collected the 228,000 tuples from the clustering query answers of "A" and "O" embeddings, which form ground truth tuples. Then we add noise to the ground truth tuples by randomly selecting tuples and reversing their similar/dissimilar label. The portion of reversed tuples defines the noise level.

Figure 1 summarizes the results, where we can observe that our method tolerates small amount of query noise, and the performance degrades smoothly. This indicates that the grouped optimization method is robust to reasonable degree of noisy measures.

1.2. Query Sampling, Query Size, and Bin Number

There are three factors controlling how we present each query to users: query sampling strategy, query size $\mathbb{N}$ and bin number $\mathbb{B}$. Query sampling controls how the images clustered are sampled from the entire set in each attribute, either randomly or locally (as proposed in [1]). In the local sampling strategy, the recovery is divided into several phases. Each hase first updates the embeddings, and then sample points to form queries from neighboring regions, rather than randomly, in the following phase. $\mathbb{N}$ is the number of images to be clustered in each query. $\mathbb{B}$ is the number of bins — the maximum number of clusters allowed for each query. We evaluate the performance of our group optimization with synthetic data generated with different sampling strategy, $\mathbb{N}$ and $\mathbb{B}$.

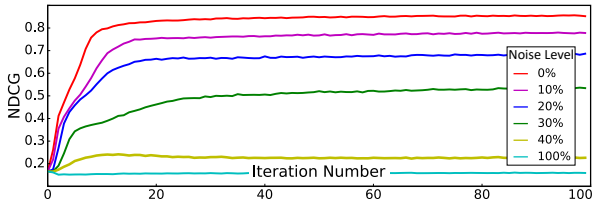

Figure 1: NDCG curve during the recovery of "AO" embeddings, with various levels of noise.

We summarize the results in Figure 2. Note that, to make fair comparisons, we make sure $|Q| \times \mathbb{N} = 3000$, i.e., more queries were used in tests with smaller query size. We can see that larger $\mathbb{N}$ and $\mathbb{B}$ give better recovery results, which is not surprising, since more information can be collected from such queries. However, the NDCG gradually goes down with larger $\mathbb{B}$ when $\mathbb{N} = 10$, since our approach degrades to the non-group case. On the other hand, smaller $\mathbb{N}$ and $\mathbb{B}$ are more friendly to crowd workers. In particular, a small value of $\mathbb{B}$ encourages workers to make decisions based on a single attribute in each query. We also show that the local sampling strategy proposed in [1], when extended to multiple-attribute setting, is more effective with small query sizes and bin numbers. There exists a sweet point which balances these factors and maximize the cost-effectiveness. the optimization progresses.

2. Comparisons with Single Embedding Methods

We also compare our method with several previous single embedding methods, including GNMDS [2], CK-

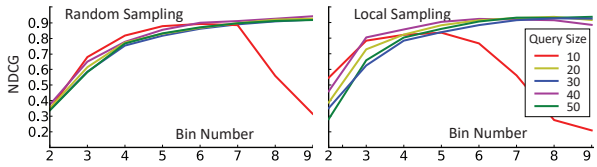

Figure 2: NDCG curve during the recovery of "AO" embeddings, using different query sampling strategy, query size and bin number.

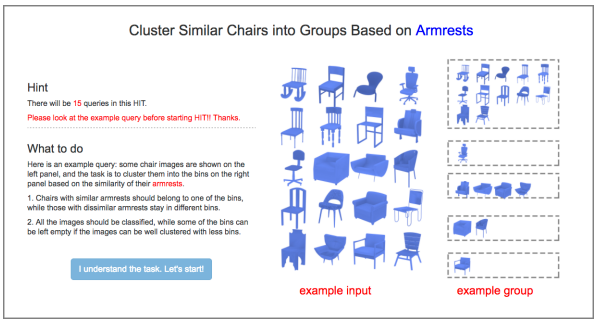

Figure 4: MTurk HIT introduction for predefined-attribute (*e.g.*, arm) experiment on Chair dataset.

L [3], and (t-)STE [4] using datasets from previous work: Food [5], Music [6], Pima [7], and Vogue [8]. Pima dataset contains 768 instances, each having 8 features indicating people's physical conditions. We adopt the method in [9] to produce similarity triplets, which generate 100 triplets for each instance in each attribute. The other three datasets offer individually collected $\{T(i, j, k)\}$ similarity triplets. We further convert each $\{T(i, j, k)\}$ triplet to $\{T(i, j, 1)\}$ and $\{T(i, k, 0)\}$ tuples, which are grouped together.

As in [9], generalization error is used to evaluate the performances, which describes the dissatisfaction ratio of new recovered triplets in the ground truth. For multi-embedding methods, a triplet is considered to be satisfied if its distance relationship in one of the multi-attribute embedding is consistent with that in ground truth. Since single embedding methods cannot recover multi-attribute embedding, we compare with their ability to recover corresponding high dimensional space. Figure 3 shows that multi-attribute embedding methods outperform single embedding methods, and the tendency is more obvious as dimensions/attributes increase.

Note that, for these datasets, which contain only minimal grouped information — each triplet is considered as a group, our algorithm degrades to non-group case. However, as shown in Figure 3, even in this case, our method performs comparable with the other methods. Also note that our method does not make any assumption on the underlying data distribution, as that in (t-)MVTE. In a summary, *our method performs better when group information is present, and comparable when only non-grouped triplets are provided*, while making no assumption of the data distribution.

3. AMT Data Collections

Data Collection on Chair Dataset. We collected semantic similarity data using clustering queries (produced by 2 updating phases of local sampling strategy) instead of the more traditional triplet queries, using a drag-and-drop graphical UI with $\mathbb{N} = 20$ clustering images shown on the left and $\mathbb{B} = 5$ grouping bins on the right. Workers were required to cluster the 20 images into the bins, with similar ones in the same bin. An experimental task in AMT for each worker is considered as a human intelligence task (HIT). Each HIT begins with an examplar introduction with guidelines for crowd workers (see Figure 4), followed by 15 queries. As a quality control, 3 queries in each HIT are ground truth sentinels, answers from a worker with lower than 70% sentinel accuracy will be rejected. Additionally, only crowd workers with higher than 80% approval rate can accept our HIT.

We distribute in total 45,000 queries with ground truth sentinels for the predefined-attributes experiment, producing 3,000 HITs. After quality control, we collected 41,287 valid clustering query answers, which are aggregated to 7,953,827 final tuples. Workers are paid on average \$0.25 to cluster 15 queries and spent an average of 6 minutes per HIT. The total cost was about \$800 spent over roughly a month and a half.

Data Collection on Poster Dataset. Film posters present rich semantic information, which makes it hard for workers to cluster queries without any predefined attributes. To simplify the task, we collect clustering queries (produced by 1 updating phase of local sampling strategy) using a drag-and-drop graphical UI with $\mathbb{N} = 10$ clustering images shown on the left and $\mathbb{B} = 2$ grouping bins on the right. Workers were asked to cluster the 10 images into the bins according to their own preferences. In this experiment, each HIT begins with an examplar introduction with guidelines for crowd workers (see Figure 5), followed by 10 queries. Since there is no ground truth answers for this experiment, we constrain valid crowd workers as those higher than 80% approval rate as a quality control.

In total, we distributed 840 queries, producing in total

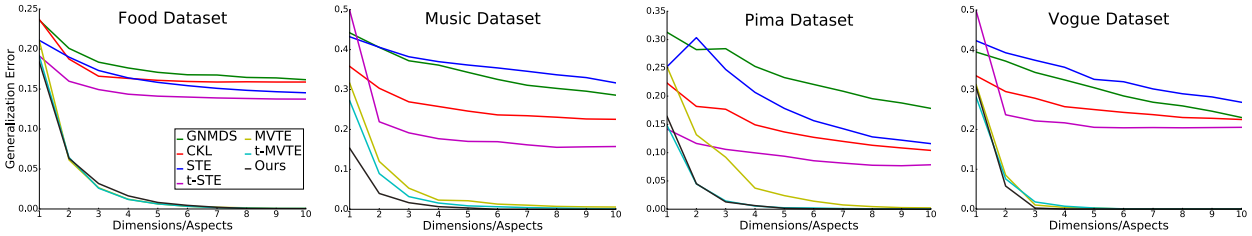

Figure 3: Comparison with state-of-the-art single and multiple embedding methods on Food, Music, Pima and Vogue datasets.

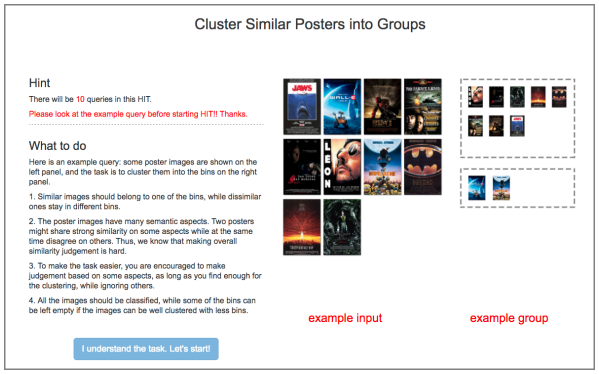

Figure 5: MTurk HIT introduction for unknown-attribute experiment on Poster dataset.

84 HITs. After excluding incomplete answers, we collected 800 valid queries which are aggregated to 36,000 final tuples. Workers are paid on average $0.25 to cluster 10 queries and spent an average of 3 minutes per response. Please note that the reward was higher in this experiment to accelerate the data collection. The total cost was about $24 spent over one day.

References

[1] Y. Kleiman, G. Goldberg, Y. Amsterdamer, D. Cohen-Or, Toward semantic image similarity from crowdsourced clustering, Vol. 32, 2016, pp. 1045–1055.

[2] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, S. Belongie, Generalized non-metric multidimensional scaling, in: Proc. of Int. Conf. on AI and Statistics, San Juan, Puerto Rico, 2007.

[3] O. Tamuz, C. Liu, S. Belongie, O. Shamir, A. Kalai, Adaptively learning the crowd kernel, in: L. Getoor, T. Scheffer (Eds.), Proc. of ICML, ACM, New York, NY, USA, 2011, pp. 673–680.

[4] L. van der Maaten, K. Weinberger, Stochastic triplet embedding, in: IEEE Int. Workshop on Machine Learning for Signal Processing, 2012, pp. 1–6. doi:10.1109/MLSP.2012.6349720.

[5] M. J. Wilber, I. S. Kwak, S. J. Belongie, Cost-effective hits for relative similarity comparisons, in: 2nd AAAI Conference on Human Computation and Crowdsourcing, 2014.

[6] D. P. Ellis, B. Whitman, A. Berenzweig, S. Lawrence, The quest for ground truth in musical artist similarity., in: ISMIR, Paris, France, 2002.

[7] J. W. Smith, J. Everhart, W. Dickson, W. Knowler, R. Johannes, Using the adap learning algorithm to forecast the onset of diabetes mellitus, in: Proc. of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1988, p. 261.

[8] H. Heikinheimo, A. Ukkonen, The crowd-median algorithm, in: Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2013, November 7-9, 2013, Palm Springs, CA, USA, 2013.

[9] E. Amid, A. Ukkonen, Multiview triplet embedding: Learning attributes in multiple maps, in: Proc.of ICML, 2015, pp. 1472C–1480.