Exchangeable Deep Neural Networks for Set-to-Set Matching and Learning Supplementary Materials

Yuki Saito[1,2][0000−0003−0492−414X], Takuma Nakamura[1][0000−0001−7904−4724], Hirotaka Hachiya[3][0000−0003−3748−4101], and Kenji Fukumizu[4,2][0000−0002−3488−2625]

[1] ZOZO Research, Jingumae, Shibuya, Tokyo, Japan
{yuki.saito,takuma.nakamura}@zozo.com
[2] The Graduate University for Advanced Studies, SOKENDAI, Tachikawa, Tokyo, Japan
[3] Wakayama University, Wakayama-shi, Wakayama, Japan
hhachiya@wakayama-u.ac.jp
[4] The Institute of Statistical Mathematics, Tachikawa, Tokyo, Japan
fukumizu@ism.ac.jp

A More Details of Models

In ablation study, we replace our mCS with max pooling, average pooling, projection metric [3], covariance matrix [9, 1], set kernel [4], and cosine similarity metric [8]. For projection metric, we use the inner product as described in [3]. For covariance matrix, we calculate two covariance matrices and the inner product between the two matrices [10]. Note that we also normalize the calculated similarities as described in [8]. For set kernel, we use Gaussian kernel and multiple kernel learning as described in [5].

B Set-Data Augmentation

In this section, we describe our set-data augmentation (set-aug) method. Algorithm 1 shows the set-aug algorithm. As we described in the paper, given positive person image pairs X and several negative person images Z, we create set pairs randomly on each training iteration. Here, index i is the iteration number in each epoch.

We use the set-aug on Road Group dataset using Algorithm 1. In each training iteration, we choose the number of base-set-size $s \in \{3, 4\}$ randomly, and select $s - 1$ paired images using $randomSelectPairedImage$. Furthermore, we add one noise image to each set randomly with a probability of 85%, and drop an image from each set randomly with a probability of 50%.

C IQON Dataset

IQON (www.iqon.jp) is a user-participating fashion web service sharing outfits for women. IQON Dataset [7] contains images with 480×480 size.

Algorithm 1: Set-Data Augmentation.

Data: paired-image dataset X, noise-image dataset Z, index i
Result: paired sets $(\mathcal{X}, \mathcal{Y})$
begin
  //select an image-pair and create initial paired sets from i-th paired-image in X, where $|\mathcal{X}| = |\mathcal{Y}| = 1$
  $(\mathcal{X}, \mathcal{Y}) \longleftarrow selectPairedImage(X, i)$
  //randomly select multiple paired images
  $(\mathcal{X}', \mathcal{Y}') \longleftarrow randomSelectPairedImage(X)$
  $\mathcal{X} \longleftarrow \mathcal{X} \cup \mathcal{X}'$
  $\mathcal{Y} \longleftarrow \mathcal{Y} \cup \mathcal{Y}'$
  //randomly drop the image(s) and use the remained set
  $\mathcal{X} \longleftarrow randomDrop(\mathcal{X})$
  $\mathcal{Y} \longleftarrow randomDrop(\mathcal{Y})$
  //randomly select the noise image(s) (if possible, select the images captured on the same camera of each target set)
  $\mathcal{X}'' \longleftarrow randomSelectImage(Z)$
  $\mathcal{Y}'' \longleftarrow randomSelectImage(Z)$
  $\mathcal{X} \longleftarrow \mathcal{X} \cup \mathcal{X}''$
  $\mathcal{Y} \longleftarrow \mathcal{Y} \cup \mathcal{Y}''$

To create our training dataset from IQON dataset, we set the maximum and minimum numbers of items for each outfit as eight and four, respectively; if the outfit contains more than eight items, then we randomly select eight items from it. The outfits contain roughly 5.5 items on average. After this operation, we created our training datasets.

D Additional Experiments

D.1 Subset Matching for Fashion Set Recommendation

In this section, we consider a different variation of the fashion set matching task to include item category restrictions and focus on the case of $Q = 1$, where Q is the number of outfits mixed in the set. We call this task subset matching.

In subset matching, for evaluation, K subsets $\{\mathcal{Y}^{(1)}, \cdots, \mathcal{Y}^{(K)}\}$ are provided as a set of matching candidates to the reference subset $\mathcal{X}$, while maintaining the category restrictions for each fashion item. That is, these K candidates only contain same-category fashion items, e.g, tops and bottoms.

For training, we also give the item category restriction. Note that without any category restrictions, the models tend to be trained to select the candidate $\mathcal{Y}^{(k)}$ that contains non-overlapped fashion category items, e.g., shoes, with $\mathcal{X}$ when we train the model in the case of $Q = 1$. To avoid this situation, we introduce the category restrictions to the K candidates in training/testing phases.

To implement the item category constraint described above, we use the triplet loss with softplus function [2] in the subset matching problem. Here, we prepare a negative set $\mathcal{Y}^{(n)}$ by selecting random items under the category restrictions. Then, we train the models using the reference set $\mathcal{X}$ as an anchor set and the subset $\mathcal{Y}^{(p)}$ as a positive set, where $\mathcal{X} \cup \mathcal{Y}^{(p)}$ and $\mathcal{X} \cup \mathcal{Y}^{(n)}$ corresponds to the given complete outfits and unmatched outfits, respectively.

Table 1 shows the comparison results. Our models showed significant improvements compared with baseline models.

Table 1: Accuracy of subset matching (%). Cand indicates the number of candidates to be matched.

Method	Cand:4	Cand:8
Set Transformer	39.2	22.7
$\text{BERT}_{\text{SMALL}}$	50.5	33.8
$\text{BERT}_{\text{BASE}}$	50.5	33.5
$\text{BERT}_{\text{BASE}-\text{AP}}$	50.0	33.5
GNN	30.3	17.3
HAP2S	29.4	16.8
Cross Attention (ours)	58.1	41.9
Cross Affinity (ours)	**60.2**	**43.3**

D.2 Weak Point Analysis

The main weak point of our models is in calculation cost. We consider that our models are promising to match a reference and candidate sets in high accuracy, but impose more substantial calculations. For example, a one-set-input function, i.e., the extension of Set Transformer, can transform a set of features individually for two sets to match. Also, after the feature extractions, it does not require calculations except the inner product in matching two vectors. Comparing with the Set Transformer, our models and the extensions of BERT and GNN models need additional calculation costs in matching two sets; they need paired sets for the feature extraction. Figure 1 shows the calculation time in the testing stage, where $|\mathbf{Y}|$ indicates the number of candidate sets. The calculation time of these models except for the Set Transformer significantly increased when the number of candidate sets increased.

Reducing the calculation costs preserving the interactions is challenging but interesting, and we leave it as future work.

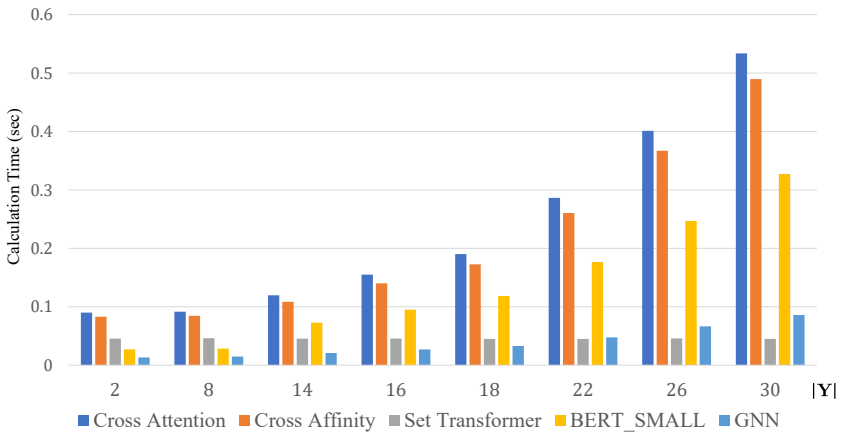

Fig. 1: Inference time for set-to-set matching. Here, we test each model 110 times successively and plot the median in the last 100 records. We randomly generated pseudo data for the calculation, which are sets of vectors on $\mathbb{R}^{512}$. Each set contains eight data. We used GeForce GTX 970 for the calculation.

E Limitations

In this section, we discuss several issues or ideas of our models and consider them as future works.

Imbalanced Samples. Our K-pair-set loss may suffer from the imbalance between positive and negative training samples when K is large, leading to bad performance; thus, a strategy of hard sample mining [6] may be needed.

Simple Interactions. Our simple CSeFT modules only include interactions from inter-sets. Introducing interactions between intra-sets and inter-sets into the CSeFT module may improve set matching results.
Matching Per Paired Sets. As described in Section D.2, one of the limitations of the set-to-set matching model is the computation cost. Currently, it is not easy to apply to larger-scale search/retrieval-like tasks.
Feature Representation. We consider that introducing regularization terms will improve matching results via mapping the same person's feature vectors into similar or identical representations in group re-id tasks.

References

1. Cai, Y., Takala, V., Pietikainen, M.: Matching groups of people by covariance descriptor. In: 2010 20th International Conference on Pattern Recognition. pp. 2744–2747. IEEE (2010)
2. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. CoRR **abs/1703.07737** (2017), http://arxiv.org/abs/1703.07737
3. Huang, Z., Wu, J., Van Gool, L.: Building deep networks on grassmann manifolds. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
4. Kim, J., McCourt, M., You, T., Kim, S., Choi, S.: Practical bayesian optimization over sets (2019)
5. Li, C.L., Chang, W.C., Cheng, Y., Yang, Y., Póczos, B.: Mmd gan: Towards deeper understanding of moment matching network (2017)
6. Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J.: Working hard to know your neighbor's margins: Local descriptor learning loss. In: Advances in Neural Information Processing Systems. pp. 4826–4837 (2017)
7. Nakamura, T., Goto, R.: Outfit generation and style extraction via bidirectional LSTM and autoencoder. CoRR **abs/1807.03133** (2018), http://arxiv.org/abs/1807.03133
8. Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: Asian conference on computer vision. pp. 709–720. Springer (2010)
9. Wang, R., Guo, H., Davis, L.S., Dai, Q.: Covariance discriminative learning: A natural and efficient approach to image set classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2496–2503. IEEE (2012)
10. Zhu, P., Zhang, L., Zuo, W., Zhang, D.: From point to set: Extend the learning of distance metrics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2664–2671 (2013)