Weakly Supervised Face Naming with Symmetry-Enhanced Contrastive Loss: Supplementary Material

Tingyu Qu[1], Tinne Tuytelaars[2], Marie-Francine Moens[1]
[1] Department of Computer Science, KU Leuven
[2] Department of Electrical Engineering, KU Leuven
{tingyu.qu, tinne.tuytelaars, sien.moens}@kuleuven.be

A. Additional Experiment on D_{2name}

We show the changes in similarity scores for Abigail Breslin, who appears in 160 samples of D_{2name}, with the associated faces during training in Figure 1. Although without NONAME added to the name list, the model can differentiate between matched faces and unmatched faces to some extent with a positive difference in similarity scores for matched and unmatched faces, the difference is very small as compared to that in the case with NONAME added. It indicates that the model is more confident with NONAME added. The reason is that having a NONAME that is not aligned to any face with known identity can help the model differentiate between matches when we have high ambiguities, because the computation of our dense similarity score relies on taking the maximum of similarity scores in each image-caption pair. The contrastive nature of the model learns to optimize the similarity scores of matched pairs to high positive values, as those for NONAME stay at the near-zero region ([-0.1, 0.1]).

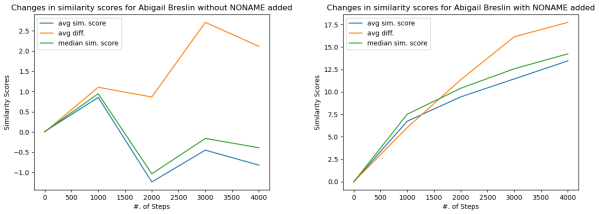

Figure 1. Changes in similarity scores of Abgail Breslin as training progresses. Orange line shows the changes in average difference in similarity scores for matched faces and unmatched faces; Blue line and green line show the changes in mean and median of similarity scores for matched faces, respectively. Left: without NONAME, Right: with NONAME.

B. Face Recognition Baseline

As a simple baseline, we conduct experiment using plain face recognition method combined with text processing, where we directly align faces to names based on face features from face recognition model and name features from the text encoder (denoted as FaceRec). As shown in Table 1, the performance of our SECLA and SECLA-B models are significantly better than the performance of FaceRec. This simple classification-based alignment approach yields dissatisfying results for the proposed task.

Model	LFW Precision	LFW Recall	LFW F1	CelebTo Accuracy
FaceRec	48.27%	53.38%	50.69%	39.55%
SECLA	76.96%	85.11%	80.83%	87.46%
SECLA-B	**77.94%**	**86.19%**	**81.86%**	**88.36%**

Table 1. Performance comparison with plain face recognition method combined with text processing (FaceRec).

C. Hyperparameter Sensitivity

We present experiments regarding hyperparameter sensitivity on LFW dataset in Table 2 and Table 3.

α	Precision	Recall	F1
0.05	76.92%	85.06%	80.79%
0.15	**76.96%**	**85.11%**	**80.83%**
0.25	76.66%	84.78%	80.52%
0.5	76.40%	84.49%	80.25%
1	76.41%	84.50%	80.25%
5	75.97%	84.02%	79.79%

Table 2. Performance of SECLA on LFW dataset with different choices of α.

As shown in Table2, the performance of SECLA remains at the same level for $\alpha <= 1$. The performance decrease a lot when we increase α to 5, indicating that too strong symmetry constraint is not suitable for LFW dataset.

As for batch size, typically contrastive loss works better for larger batch size. While in our case, larger batch size also means more disagreement from face-to-name and

name-to-face directions. In our experiments, batch size=20 works best for LFW dataset.

It can also be seen in Table 3 that the performance of SECLA increases slightly when we enlarge batch size from 32 to 40, which shows the potential of having better performance with large batch size and proper symmetry constraint using SECLA model.

Batch size	Precision	Recall	F1
4	75.64%	83.65%	79.45%
16	76.78%	84.91%	80.64%
20	**76.96%**	**85.11%**	**80.83%**
32	76.74%	84.86%	80.60%
40	76.88%	85.02%	80.74%

Table 3. Performance of SECLA on LFW dataset with different choices of batch size.

D. Replacing Pre-trained Features

We conduct experiments by replacing pre-trained features with other types of features that do not contain domain knowledge. We present the results in Table 4.

Face Feature	Name Feature	Precision	Recall	F1
FaceNet	BERT-base	**76.96%**	**85.11%**	**80.83%**
FaceNet	one-hot encoding	57.95%	64.09%	60.87%
ResNet-34	BERT-base	64.49%	71.32%	67.73%

Table 4. Performance of SECLA on LFW dataset by replacing pre-trained features. ResNet-34 refers to features from ResNet-34 pre-trained on ImageNet, while one-hot encoding refers to embedding directly taken from one-hot encoding.

By representing each name with one-hot encoded embedding, the performance of SECLA decreases significantly (from F1=80.83% to 60.87%). One of the biggest problems with one-hot encoding is that it cannot distinguish the names of the same person. For example, due to the large corpus used for pre-training BERT, "Bush", "George Bush" and "George W. Bush" are likely from the same person judging by BERT-base. While these three names are completely different for one-hot encoding. The domain knowledge of the identity of names is essential for our contrastive learning-based model, for it does not have supervision during training, but relies on maximizing dense similarity scores for corresponding pairs. Giving completely different name features to the same identity causes confusion. We expect our model to work well with one-hot encoding if we can perform entity resolution properly before training, which we leave for future work.

The importance of name features can also be seen in another experiment. In this experiment, unlike our method or the previous SOTA [17], where domain-specific features (for face recognition) are used, we adopt features from ResNet-34 pre-trained on ImageNet (ResNet-34 features), which are from a different domain (general image classification). As shown in Table 4, by replacing features from pre-trained FaceNet with ResNet-34 features, we can still achieve satisfactory performance with over 70% ground-truth links correctly linked.

Surprisingly, even with the bad choices of features, our SECLA model still achieves better performance than other weakly supervised deep learning-based vision-language alignment methods [12, 25] without the usage of context information of the caption or the image. And we also achieve comparable performance in terms of recall with the previous SOTA [17]. It shows the superiority of our method in this task.