Contrastive Personalization Approach to Suspect Identification (Student Abstract)

Devansh Gupta[1*], **Drishti Bhasin**[1*], **Sarthak Bhagat**[1] , **Shagun Uppal**[1] , **Ponnurangam Kumaraguru**[2] , **Rajiv Ratn Shah**[1]

[1] MIDAS Lab, Indraprastha Institute of Information Technology, New Delhi, India
[2] International Institute of Information Technology, Hyderabad, India
{devansh19160, sarthak16189, shagun16088, rajivratn}@iiitd.ac.in, drishti_b@me.iitr.ac.in, pk.guru@iiit.ac.in

Abstract

Targeted image retrieval has long been a challenging problem since each person has a different perception of different features leading to inconsistency among users in describing the details of a particular image. Due to this, each user needs a system personalized according to the way they have structured the image in their mind. One important application of this task is suspect identification in forensic investigations where a witness needs to identify the suspect from an existing criminal database. Existing methods require the attributes for each image or suffer from poor latency during training and inference. We propose a new approach to tackle this problem through explicit relevance feedback by introducing a novel loss function and a corresponding scoring function. For this, we leverage contrastive learning on the user feedback to generate the next set of suggested images while improving the level of personalization with each user feedback iteration.

Introduction

Explicit relevance feedback is a method used in targeted information retrieval, where a system poses certain queries and utilizes user feedback to iteratively improve the queries, thus efficiently reducing the search space to retrieve the target information. Adapting to the user's preferences with each feedback iteration becomes crucial for an efficient system.

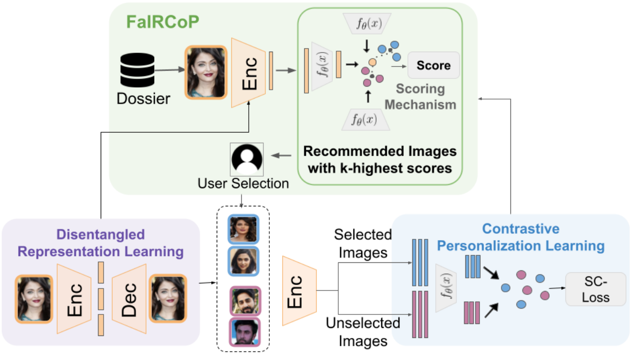


Figure 1: An illustration of our proposed approach.

In this paper, we tackle the problem of personalized recommendations through relevance feedback where the system presents the user with a set of query images and the user provides feedback by selecting similar/dissimilar images. A key challenge in this task is modelling user preferences with limited contextual data from previous queries in an unbiased fashion. Deep metric learning has proved to be an effective learning paradigm where an abstract notion of similarity can be projected to a space with quantifiable standard metrics. Since such models are trained using pairwise similarities on input data, it is an effective technique to apply in sparse data settings. We use this concept to model user preferences during relevance feedback. We focus on the problem of *suspect identification*, which plays an important role in forensic investigations. The witness is suggested some images from a criminal data dossier and is expected to identify similar/dissimilar images. The aim is to use this information to personalize the system based on user's perceptual preferences, and reach the suspect's image in minimum iterations.

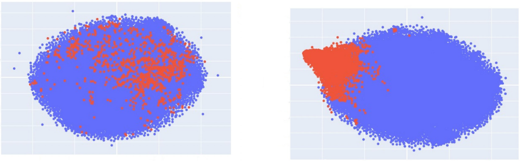
Methodology

We use an extensive criminal database (Jain et al. 2021) for testing our method. We now explain different stages of our approach in detail.

Learning Latent Representations: In our dataset, we encountered various images with missing or noisy labels. To avoid relying on these labels, we use unsupervised learning to represent our images using disentangled representations because they tend to be interpretable, work well on downstream tasks and do not depend on large amounts of data. We used the method proposed in (Hu et al. 2018) to learn disentangled representations for each image in order to curate a set of representations for all images in the database.

Personalized Relevance Feedback: For personalized feedback, we use a fully connected neural network called a projection network which projects the original representations on to a space. The projected space acts as a representative to the user preferences by attempting to project all the images selected as similar by the user in one cluster and all the images selected as dissimilar in another cluster. In order to see how our approach gets adapted to the user preferences in a robust manner, we explain the training stage and the inference stage of our approach. In the training stage, we

[*] These authors contributed equally.



(a) Initial Projections

(b) Final Projections

Figure 2: Latent visualizations of selected (blue) and non-selected (red) images before and after SCLoss optimization.

define a new loss function called the SCLoss which stands for Separating Cluster Loss, given by Equation (1), closely related to the NT-Xent loss defined in (Chen et al. 2020). We used cosine similarity to compute similarity.

$$
\begin{aligned}
L_s(S,D) &= \sum_{x \in S} \sum_{y \in S-\{x\}} -log \frac{e^{sim(x,y)/\tau}}{\sum_{k \in D} e^{sim(x,k)/\tau}} \\
L_d(S,D) &= \sum_{x \in D} \sum_{y \in D-\{x\}} -log \frac{e^{sim(x,y)/\tau}}{\sum_{k \in S} e^{sim(x,k)/\tau}} \qquad (1) \\
SCLoss(S,D) &= \frac{L_s(S,D)}{2|S|(|S|-1)} + \frac{L_d(S,D)}{2|D|(|D|-1)}
\end{aligned}
$$

In Equation (1), set S represents the projections of a batch of similar images chosen while set D represents a projected batch of images not selected when suggested, τ is a scaling constant, and sim is the cosine similarity function. The inference stage consists of extracting a scoring from the framework in Figure 1. Since the projected space has two separate clusters for similar and dissimilar images, we derive a score of an image based on the similarities of its projected representation with both the cluster centers mentioned in Equation (2). In the given score function, S_a represents the set of projected representations of all the images which were chosen as similar by the user while D_a represents all the projected representations of images which were shown to the user but not chosen, over all previous iterations.

$$
score(u) = sim\left(\frac{1}{|S_a|}\sum_{x \in S_a} x\right) - sim\left(\frac{1}{|D_a|}\sum_{x \in D_a} x\right) \qquad (2)
$$

We combine the above two stages to create a relevance feedback mechanism where the projection network can be trained on similar and the dissimilar entities selected by the user for any number of epochs under any specified conditions. It must be ensured that each set of images for training at any iteration must have a substantial number of images from the previous set which had already been used for training. This step ensures that only two clusters are being formed (as shown in Figure 2) in the projected space since the previously trained images act as anchors to the respective clusters, thus, ensuring that the projection network is trained to project the new input image representations into the previously formed respective clusters and not create their own clusters in the projected space.

Results

We compare our algorithm to Rocchio algorithm (Siradjuddin, Triyanto, and S. 2019) to perform explicit relevance feedback on the criminal database (Jain et al. 2021). We observe that our algorithm retrieves the target image in a fewer number of iterations and with a higher magnitude of average relevance than Rocchio in different settings. We compare and select images using a combination of different image embeddings and used the Euclidean distance metric to select similar/dissimilar images in each iteration to roughly emulate a human's feedback. We use FaceNet (Schroff, Kalenichenko, and Philbin 2015), HOG embeddings (Dalal and Triggs 2005), and MIX (Hu et al. 2018) to compare the algorithms (see Table 1).

	No. of Iterations		Avg. Relevance	
Simulator	Rocchio	**Ours**	Rocchio	**Ours**
FaceNet+MIX	891.1	**103.1**	0.75	**0.90**
HOG+MIX	528.0	**86.9**	0.47	**0.80**
FaceNet+HOG+MIX	631.4	**73.4**	0.58	**0.77**
MIX	247.4	**31.5**	0.57	**0.78**

Table 1: Evaluation metrics of 10 simulation runs on Rocchio (Siradjuddin, Triyanto, and S. 2019) and our algorithm.

Conclusion

In this paper, we propose SCLoss and a scoring mechanism for targeted image retrieval through relevance feedback. We believe this algorithm gives new perspectives to personalization of suggestions in online settings where very less data is available to derive context.

Acknowledgements

Rajiv Ratn Shah is partly supported by the Infosys Center for AI and the Center of Design and New Media at IIIT Delhi.

References

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 1597–1607. PMLR.

Dalal, N.; and Triggs, B. 2005. Histograms of Oriented Gradients for Human Detection. *CVPR 2005*, 886–893.

Hu, Q.; Szabó, A.; Portenier, T.; Favaro, P.; and Zwicker, M. 2018. Disentangling Factors of Variation by Mixing Them. In *CVPR*.

Jain, A.; Shah, M.; Pandey, S.; Agarwal, M.; Shah, R.; and Yin, Y. 2021. *SeekSuspect: Retrieving Suspects from Criminal Datasets Using Visual Memory*. ACM.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Siradjuddin, I. A.; Triyanto, A.; and S., M. K. 2019. Content Based Image Retrieval with Rocchio Algorithm for Relevance Feedback Using 2D Image Feature Representation. In *MLMI*, 16–20. ACM.