One-shot multiple-object tracking using contrastive learning

5AUA0 - Advanced Sensing using Deep Learning - Group 12 - Team 1

Kevin Delnoije - 0942300
Eindhoven University of Technology
Department of Electrical Engineering
k.r.e.delnoije@student.tue.nl

Nadine Nijssen - 1012656
Eindhoven University of Technology
Department of Electrical Engineering
n.a.a.nijssen@student.tue.nl

Abstract

In multi-object tracking the leading paradigm is tracking-by-detection which is often a two step approach. Recent one-shot approaches have shown promising results that are able to run in real-time. One-shot models learn detections and appearance embeddings jointly. We built upon the one-shot method by changing the way the embeddings are trained. Softmax based features are trained by classifying the embedding feature map to the correspond track-IDs. We propose a pairwise loss that is able to learn embeddings without track-ID labels. On the MOT17 dataset we obtain competitive results with respect to the softmax based method. Code is available here.

1. Introduction

Many problems in computer vision are nowadays tackled with deep learning approaches. One of these tasks is multi-object tracking (MOT) where multiple objects are given a unique ID across a series of video frames, these are called tracklets. This problem finds its application in many different fields and can be used for tasks such as surveillance and autonomous driving, where real-time MOT is essential. The task is often solved with the tracking-by-detection paradigm [9], which consists of two steps: detection and association. In the detection step, targets are localized and indicated with bounding boxes. The association step is responsible for assigning the detections to new or existing tracklets. Extracting appearance embeddings of the instances can help during the association step when location and motion information alone is insufficient due to occlusion or overlapping instances. When an instance is lost for several frames the embeddings can be used to re-assign the instance to the original track.

In this project we implement a MOT system based on FairMOT by Zhang et al. [13]. In their work they use an encoder-decoder network to extract a high resolution feature map. They follow it up with two parallel branches: one detection branch to predict the bounding boxes, and one re-ID branch to extract identity embedding features.

For creating a tracking dataset, all instance locations need to be annotated which is usually done using bounding boxes and each instance in a video sequence needs to be given a unique track-ID. This research investigates the possibility to learn identity embeddings using a contrastive loss by comparing person detections in a single frame without using their track-ID.

2. Related work

2.1. Two-step

The dominant paradigm in multiple object tracking is based on a two-step approach where seperate models are used for the detection and appearance features. Detections can be either public (given by the dataset) or a private detector can be used. The detection outputs are used to crop and resize instances from the original image to be used by the CNN feature extractor which creates the appearance embeddings. Yu et al. (POI) [10] use a GoogLeNet based feature extractor. They train on multiple person re-identification datasets with softmax and triplet loss on the cosine distance of the appearance features.

Another approach to extract appearance features is to train CNNs with a loss function that learns features that distinguish better between different objects. This is often done with a Siamese network [4] [15] or a CNN pre-trained on person re-identification datasets [12] [2], trained using contrastive [4] or triplet [15] [12] [2] loss function on the Euclidean [4] [12] [2] or cosine [15] distance between the appearance features.

2.2. One-shot

Recent research has shown that jointly detecting and learning embeddings (JDE) gives competitive results [9]. This makes it possible to use a single model for the tracking

task allowing for real-time inference speed. The object detection network is augmented by adding an extra head that is responsible for learning appearence features. This method suffered from an high number of identity switches compared to existing two-step approaches. Follow up work [13] shows that due to the coarse nature of anchor boxes and the relatively large output stride of the network ambiguities exist when learning embeddings. By using an detection network based on keypoints and a smaller output stride [14], the number of identity switches was comparable to two-step methods.

2.3. Deep metric learning

Most work in deep metric learning is focused on classification tasks like CUB-2011, Cars196 and Stanford online products. Each image here consists of a label, for example the car type. Usually half of the classes are used for training, the other half is used to test whether the embeddings can cluster similar classes together. It has been shown that softmax classification loss is a strong baseline for learning embeddings [3] [11]. Therefore, it is no surprise that the FairMOT baseline obtains good results with embeddings. Pairwise metric learning has also been applied successfully to these datasets [6].

Extending the losses to generate embeddings for our problem is non-trivial because of two reasons. First of all, unlike classification tasks where each image is represented by a single embedding, the annotated images in the MOT domain contain multiple objects so the output feature map captures features of multiple instances, and second, it is a multi-task learning problem where both detection and appearance embeddings need to be learned.

3. Method

The FairMOT network in [13] is used as baseline, see Figure 1. The re-id head of this network learns an feature map $\mathbf{E} \in R^{128 \times W \times H}$, where W and H are the width and height of the output feature map. To get the identity embedding of a person at point (x, y), the 128 dimensional embedding vector is extracted from the feature map at that point.

3.1. Identity embedding classification loss

In the FairMOT network, a fully connected layer maps the embedding vectors to the corresponding track-ID of the unique instance. They use the cross-entropy loss function to learn to predict these labels, treating identity embedding as a classification task. During inference the fully connected layer is ignored and embedding vectors are extracted from the location that matches the detection.

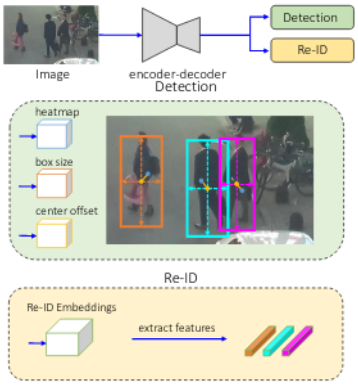


Figure 1: Diagram of the baseline network from the original paper of FairMOT[13].

3.2. Identity embedding pairwise ranking loss

We propose to use metric learning with the pairwise ranking loss. The pairwise ranking loss uses positive and negative pairs of samples. The aim is to minimize the distance between positive pairs and have a distance larger than margin m between the negative pairs. The loss is calculated for each pair in the batch and can be written as:

$$L_{pair} = \begin{cases} d(x_a, x_p) & \text{if positive pair} \\ \max(0, m - d(x_a, x_n)) & \text{if negative pair} \end{cases} \quad (1)$$

where $d(\cdot)$ is either the cosine (cos.) or Euclidean (Euc.) distance. The total loss is the mean of the loss per pair in the batch.

3.3. Sampling methods

Different sampling or mining methods can be used to extract positives and negatives belonging to an anchor embedding. We extract the anchors, positives and negatives from the same frame. The anchors are the embeddings at the location of the ground truth bounding box centers. All anchors in a single image belong to a unique person. Negatives to an anchor are then the other anchors in an image. We use hardest negative sampling, where the negative with the smallest distance to the anchor is selected as the hardest negative.

For positive sampling we take an embedding that is slightly shifted from the center point on the embeddings map, but still belongs to the same person. We calculate a Gaussian radius r based on the width and height of the bounding box. The radius is also used to generate the heatmap for the detection ground truth, see Figure 2. Then, using the bounding box center width and height on the embeddings map, we randomly step a half a radius up, down, left or right, see (2). The embedding at that point is taken as

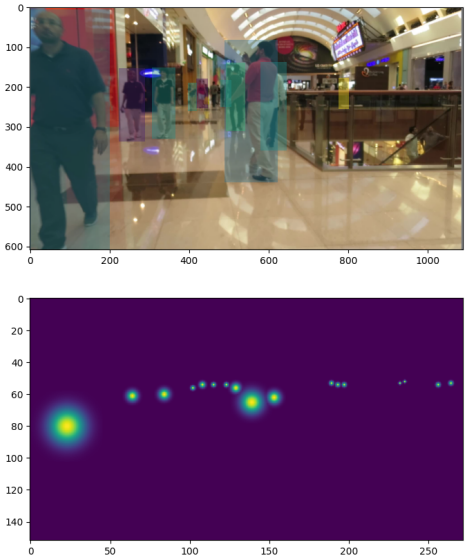


Figure 2: The first frame of sequence 11 with ground truth bounding boxes on the largest 10 persons (top) and the heatmap showing all persons (bottom).

the positive sample.

$$(w_p, h_p) = \begin{cases} \text{up/down} & (w_a, h_a \pm \frac{r}{2}) \\ \text{or left/right} & (w_a \pm \frac{r}{2}, h_a) \end{cases} \quad (2)$$

4. Experiments

4.1. Data

The MOT17 [5] dataset with SDP (scale-dependent pooling) detections is used for evaluating the methods. The dataset focuses on multiple people tracking. The set consists of 7 train and 7 test video sequences. We use 5 sequences for training (sequences 4, 5, 9, 10 and 13) and 2 for validation (sequences 2 and 11). The videos are filmed in unconstrained environments with both static and moving cameras. The average pedestrian density is between 8.3-69.8 per frame. More information on the dataset can be found on the MOT Challenge website.[1]

4.2. Baseline

To perform experiments with different loss functions, we first establish a baseline. The backbone network that we use is High-Resolution Network V2 (HRNetV2) [8] with 18 layers. To use this as a baseline, the backbone network is trained from ImageNet initialized weights using the same method as FairMOT. This baseline model is used for retraining different methods. When doing this, either the backbone is frozen (fr.) or both the backbone and detection head are frozen (fr.det.). Freezing both backbone and detection ensures that the detection quality stays as in the baseline, however, it is less realistic in simulating training from scratch than only freezing the backbone.

[1] https://motchallenge.net/

4.3. Evaluation

4.3.1 MOT metrics

The metrics used to evaluate the tracking accuracy are the ones that are mostly used and are shown in the MOTChallenge leaderboards. These are MOTA (Multiple Object Tracking Accuracy), identity switches and the IDF1 (identification F1) score. MOTA belongs to the CLEAR MOT metrics [1] and combines a few simpler metrics: the number of false negatives, of false positives and of ID switches, thereby acting as a summary metric for MOT. MOTA is given as a percentage, where it is better when it is higher. The number of identity switches (IDs) is used to asses the matching accuracy which is affected by the detection quality. The IDF1 score [7] balances identification precision and recall, and it therefore useful to evaluate the performance of the re-ID head. The IDF1 score is defined as the ratio of correctly identified detections over the average number of ground-truth and computed detections.

The tracking scores for models that train with the pairwise loss function using the Euclidean distance also use Euclidean distance calculation to do association based on the embeddings, while the other models use cosine distance.

4.3.2 Embeddings evaluation

To evaluate the performance of the appearance embeddings as in [9] using the True Positive Rate at False Accept Rate of 0.1 (TPR@FAR=0.1), where the embeddings are extracted from the ground truth bounding boxes.

The embeddings are visualized using the dimensionality reduction technique t-SNE. We use one validation video sequence (sequence 11) to visualize and evaluate the embeddings. To see better how the embeddings of a single person are clustered together and spread from other persons, we visualize and evaluate only the 10 largest persons (with largest bounding box) from the first frame over the first 100 frames, see Figure 2.

5. Results and discussion

5.1. Baseline results

The baseline model is trained with batch size 8 for 45 epochs on the training set. The tracking performance in terms of IDF1, ID switches and MOTA on the validation set is given in Table 1. This network is used as the our baseline for the experiments.

Table 1: IDF1, ID switches and MOTA on the validation set for the baseline model and the freeze trained models with classification loss.

Model	IDF1	IDs	MOTA
Baseline	39.5%	245	29.5%
Class. loss (fr.)	37.6%	372	28.5%
Class. loss (fr.det.)	40.8%	216	29.9%

5.2. Classification loss results

Two models are trained with the classification loss function. For one model, the backbone of the baseline model is frozen and a learning rate of 1e-4 is used. For the other model, the backbone and detection branch of the baseline model are frozen and the learning rate is 1e-5. Both networks are trained for 10 epochs.

The tracking performance of both models using classification loss is given in Table 1. The TPR score of the embeddings generated with the models is given in Table 3. The biggest differences between the two models is seen in the identity switches and the TPR of the embeddings: if the detection branch is frozen as well, the number of identity switches is lower and TRP is higher. There is not a big difference between the other tracking scores.

5.3. Pairwise loss results

The tracking performances of the models trained using pairwise loss function are given in Table 2. The TPR scores of the embeddings generated with these models are given in Table 3. For the models that use the cosine distance function, the same differences in performance are observed as for the models that use classification loss. However, for the models with Euclidean distance, the ID loss (pairwise loss) is observed to be quite high. Therefore, while training with the frozen backbone, the network has more focus on optimizing the identity embeddings, and focuses less on optimizing the detection branch. This means the detections are less accurate, hence the heatmap centers are not aligned with the actual object centers. The learned re-ID features will then be sub-optimal, which in turn leads to lower tracking scores, due to many false negatives, false positives and identity switches and low identification precision and recall. Whereas freezing the detection branch as well, results in higher tracking scores.

In Figure 3 a visualization of the embeddings for the model with frozen backbone and detection that uses the cosine distance is shown. It can be seen that the embeddings of the same person are clustered together. It is interesting to see that the network is able to cluster individual object in a high dimensional space without having an idea about clusters over the entire video sequence during training, as is the case when using tracking annotations.

Table 2: IDF1, ID switches and MOTA on the validation set for the freeze trained models with pairwise loss, for different distance functions and hyperparameters (lr. = learning rate, m. = margin.)

Model	IDF1	IDs	MOTA
Euc. (fr.) lr.=1e-4, m.=10	14.2%	817	-104.3%
cos. (fr.) lr.=1e-4, m.=0.5	36.5%	303	29.1%
Euc. (fr.det.) lr.=1e-5, m.=10	38.4%	302	29.7%
cos. (fr.det.) lr.=1e-5, m.=0.5	38.4%	258	29.5%

Table 3: Embedding evaluation for different models, calculated for the 10 largest persons from the first frame over the first 100 frames.

Model	TPR@FAR=0.1
baseline	0.8284
class. loss (fr.)	0.6962
class. loss (fr.det.)	0.8586
pair. loss Euc. (fr.)	0.5764
pair. loss cos. (fr.)	0.6702
pair. loss Euc. (fr.det.)	0.7489
pair. loss cos. (fr.det.)	0.7933

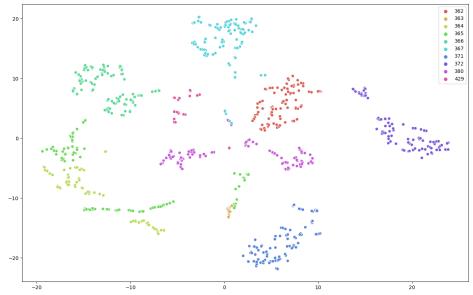


Figure 3: Visualization of the embeddings (using t-SNE) for model: pairwise loss with cosine distance (fr.det.). Shown are the 10 largest persons from the first frame over the first 100 frames. Each color represents a different person.

6. Conclusion

In this research we have shown that one-shot MOT models can learn identity embeddings from single frames without using tracking annotations. We used a pairwise ranking loss function that is able to learn an embedding space using a contrastive method between persons. We obtain competitive results compared to softmax based features that use track-ID for supervision. We believe that these results are promising and encourage more self-supervised methods on

the task of multi object tracking.

References

[1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Eurasip Journal on Image and Video Processing*, 2008.

[2] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification. *Proceedings - IEEE International Conference on Multimedia and Expo*, 2018-July, 2018.

[3] Shota Horiguchi, Daiki Ikami, and Kiyoharu Aizawa. Significance of Softmax-Based Features in Comparison to Distance Metric Learning-Based Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1279–1285, 2020.

[4] Minyoung Kim, Stefano Alletto, and Luca Rigazio. Similarity Mapping with Enhanced Siamese Network for Multi-Object Tracking. *arXiv preprint arXiv:1609.09156*, 2016.

[5] Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A Benchmark for Multi-Object Tracking. *arXiv preprint arXiv:1603.00831*, pages 1–12, 2016.

[6] Qi Qi, Yan Yan, Zixuan Wu, Xiaoyu Wang, and Tianbao Yang. A Simple and Effective Framework for Pairwise Deep Metric Learning. *arXiv preprint arXiv:1912.11194*, pages 1–16, 2020.

[7] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 17–35, 2016.

[8] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8828(AUGUST 2019):1–1, 2020.

[9] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards Real-Time Multi-Object Tracking. *arXiv preprint arXiv:1909.12605*, 2019.

[10] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. POI: Multiple object tracking with high performance detection and appearance feature. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9914 LNCS:36–42, 2016.

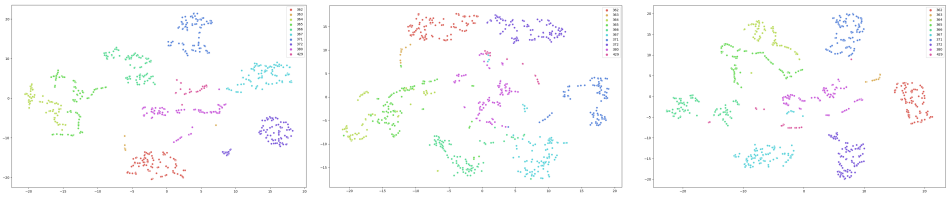
[11] Andrew Zhai and Hao-Yu Wu. Classification is a Strong Baseline for Deep Metric Learning. *arXiv preprint arXiv:1811.12649*, 2019.

[12] Shun Zhang, Yihong Gong, Jia Bin Huang, Jongwoo Lim, Jinjun Wang, Narendra Ahuja, and Ming Hsuan Yang. Tracking persons-of-interest via adaptive discriminative features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9909 LNCS, pages 415–433. Springer, Cham, 2016.

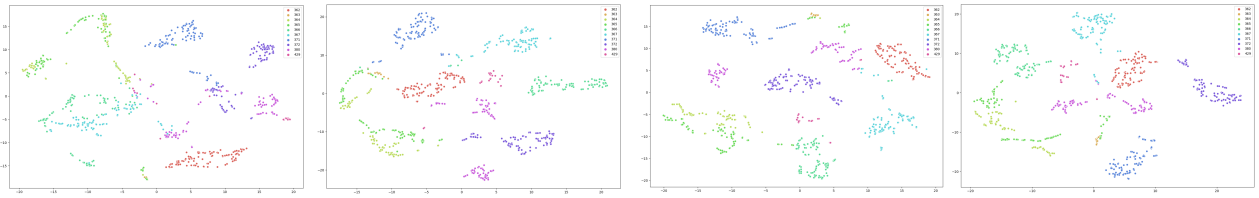
[13] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. A Simple Baseline for Multi-Object Tracking. *arXiv preprint arXiv:2004.01888*, 4 2020.

[14] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as Points. *arXiv preprint arXiv:1904.07850*, 2019.

[15] Zongwei Zhou, Junliang Xing, Mengdan Zhang, and Weiming Hu. Online Multi-Target Tracking with Tensor-Based High-Order Graph Matching. *Proceedings - International Conference on Pattern Recognition*, 2018-Augus:1809–1814, 2018.



(a) Baseline model (b) Classification loss (fr.) (c) Classification loss (fr.det.)



(d) Pairwise loss with Euclidean distance (fr.) (e) Pairwise loss with cosine distance (fr.) (f) Pairwise loss with Euclidean distance (fr.det.) (g) Pairwise loss with cosine distance (fr.det.)

Figure 4: Visualization of the embeddings (using t-SNE) for different models. Shown are the 10 largest persons from the first frame over the first 100 frames. Each color represents a different person.