SELF-SUPERVISED FEATURE LEARNING FOR ONLINE MULTI-OBJECT TRACKING

Brian Weston*, Bradley Collicott†, Mrunal Sarvaiya*

Departments of Computer Science and Aeronautics/Astronautics†, Stanford University, Stanford, CA*

Background and Project Goals

Multi-object tracking (MOT) from video lies at the intersection of multiple core problems in computer vision. The MOT process consists of 4 stages: Detection, Feature Extraction, Motion Prediction, and Data Association. DeepSORT [1], one of the current state-of-the-art methods in real-time tracking, uses deep learning to provide a dense feature representation for use in a data association algorithm.

The data association step (i.e. associating a detection with an existing trajectory) in MOT is widely considered to be the bottleneck for current performance. Therefore, this work seeks to investigate methods for improving feature representation to reduce the likelihood of ID switching (IDSW) and incorrect associations. We attempt to improve the feature representation in 2 ways: (1) self-supervised transfer learning on a pre-trained ResNet-18, and (2) self-supervised training of a convolutional autoencoder network.

Problem Approach and Methodology

Pretext Tasks for Transfer Learning:

- ResNet-18 Pretrained on ImageNet
- Jigsaw Puzzle and Rotation Prediction Tasks

Fig. 1: Puzzle and Rotation Task Input/Label Examples

Fully Self-Supervised Method:

- Autoencoder image reconstruction

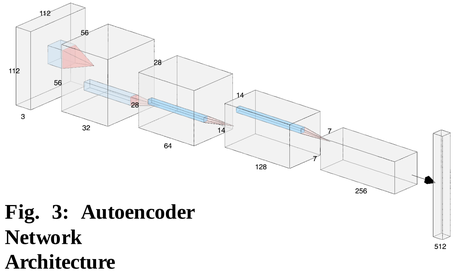

Fig. 3: Autoencoder Network Architecture

Baseline: Cosine Metric Learning Network from original DeepSORT implementation [1].

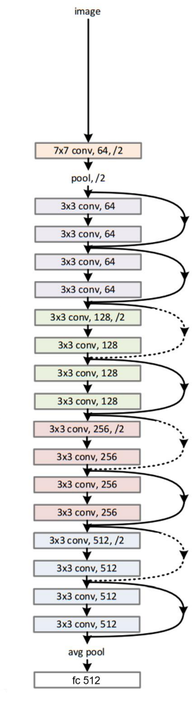

Fig. 2: ResNet-18 Architecture

Dataset

MOT Evaluation:

- MOT 17 Dataset
- Over 1300 unique IDs
- Over 11000 video frames

Self-Supervised Transfer Learning:

- MARS Person Re-ID Dataset
- Multi-view pedestrian tracks

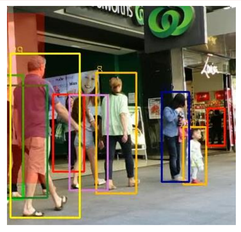

Fig. 4: Annotated MOT17 Video Frame

Fig. 5: Example Pedestrian Track from MARS dataset

Convolutional Autoencoder Reconstruction Task:

- MS COCO Object and Scene dataset

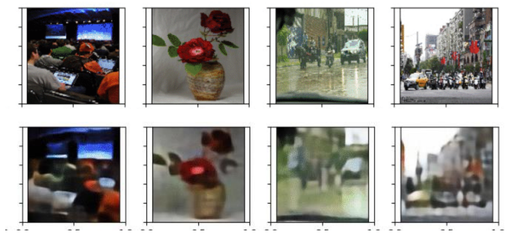

Fig. 6: MS COCO Images (top with Autoencoder Reconstructions (bottom)

Metrics

The CLEAR MOT metrics established in [2] will be used to evaluate tracking performance.

$$MOTA = 1 - \frac{FN + FP + IDSW}{GT} \qquad MOTP = \frac{\sum_{t,i} IOU(d_{t_i}, \hat{d}_{t_i})}{\sum_t c_t}$$

- FN: Ground truth object with no hypothesis
- FP: Hypothesis for which no ground truth associated.
- IDSW: A change to a correct ID association.
- FRAG: Trajectory hypothesis that covers less than 80% of a ground truth track

Additionally, we consider:

- MT: trajectories correctly tracked >= 80% of frames
- ML: Trajectories correctly tracked <= 20% of frames

Results and Analysis

	No. Params	MOTA ↑	MOTP ↑	MT ↑	ML ↓	IDSW ↓	FRAG ↓	FP ↓	FN ↓
Baseline [23]	2.8m	*48.207*	*83.746*	340	573	*1458*	*3680*	161291	*11738*
ResNet-18 [9]	11.2m	47.869	**83.561**	362	546	2176	4175	159140	14308
ResNet-18 + Puzzle	11.2m	47.488	83.541	363	**543**	2412	4231	159183	14301
ResNet-18 + Rot.	11.2m	47.112	83.297	354	545	4100	4603	159582	14494
Autoencoder (COCO)	6.8m	**47.947**	83.557	**367**	546	**2064**	**4172**	**159054**	**14243**
Autoencoder (MOT17)	6.8m	47.970	83.559	370	544	2048	4136	159064	14174

Table 1. MOT17 Evaluation Results – **BOLD**=Best results in this study; ***Italics***=Baseline Obtains best results

MOT performance:

- Autoencoder outperforms pre-trained ResNet-18 in MOTA, MT, IDSW, FRAG, FP, FN
- Baseline network performs best in holistic metrics, but is slightly outperformed in MT, ML
- Self-Supervised transfer learning on rotation and puzzle tasks proved ineffective for the ResNet-18
- Autoencoder performance does not significantly improve when trained on the evaluation set

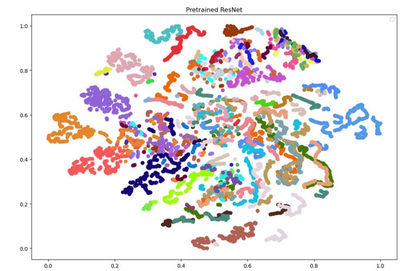

Fig. 7: ResNet-18 tSNE

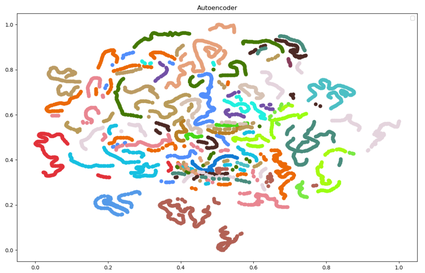

Fig. 8: Autoencoder tSNE

TSNE:

- Both ResNet-18 and Autoencoder networks are capable of partitioning the feature vector space
- Autoencoder clusters have less overlap than ResNet-18, indicating better differentiation between IDs – supported by quantitative results

Conclusion and Future Work

- Autoencoder trained on self-supervised image reconstruction produces suitable feature descriptors for MOT
- Self-supervised transfer learning is ineffective for augmenting a pre-trained network's feature vector
- Future work includes combining self-supervised autoencoder with metric learning for person reidentification

References

[1] N.Wojke,A.Bewley,andD.Paulus.Simpleonlineandreal- time tracking with a deep association metric. *Proceedings - International Conference on Image Processing, ICIP*, 2017- September:3645–3649, 2 2018.

[2] K.BernardinandR.Stiefelhagen.Evaluatingmultipleobject tracking performance: The clear mot metrics. 1 2008.