Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation

Pierre Sermanet* Corey Lynch*† Jasmine Hsu Sergey Levine
sermanet,coreylynch,hellojas,slevine@google.com
Google Brain

We propose a self-supervised approach for learning representations of relationships between humans and their environment, including object interactions, attributes, and body pose, entirely from unlabeled videos recorded from multiple viewpoints (Fig. 2). We train an embedding with a triplet loss that contrasts a pair of simultaneous frames from different viewpoints with temporally adjacent and visually similar frames (Fig. 1). We call this model Time-Contrastive Networks (TCN). The contrastive signal encourages the model to discover meaningful dimensions and attributes that can explain the changing state of objects and the world from visually similar frames while learning invariance to viewpoint, occlusions, motion blur, lighting, background. The experimental evaluation of our multi-viewpoint embedding technique examines its application to reasoning about object interactions, as well as human pose imitation with a real robot. We demonstrate that our model can correctly identify corresponding steps in complex object interactions, such as pouring (Table 1), between different videos and with different instances. We also show what is, to the best of our knowledge, the first self-supervised results for end-to-end imitation learning of human motions with a real robot (Table 2). Results are best visualized in videos available at [1] and the full paper is available at [2].

Unsupervised Object Interactions

We compare our multi-view TCN model against the Shuffle & Learn[1] approach, using the exact same architecture and only changing the loss and the last layer. Both models are initialized with ImageNet classification weights, then trained in a self-supervised manner using 15 minutes of multi-view pouring videos (no labels). We test on 5 minutes of unseen pouring videos. An off-the-shelf ImageNet-pretrained Inception model is also used as a baseline (no training on pouring data). Finally, we also propose a single-view TCN to compare with. We find that TCN outperforms all baselines on different quantitative metrics (Table 1), and that multi-view outperforms the single-view model. Both metrics use the nearest neighbor of a reference frame in the embedding of each method. The alignment metric measures how well different sequences of a same demonstrations can be semantically aligned using different embeddings. The attributes classification metric measures how well different attributes that are useful to perform a pouring task are modeled by different embeddings.

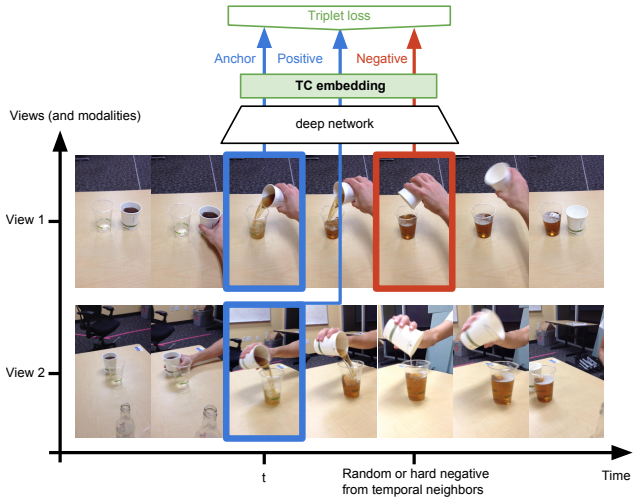

Figure 1: **Time-Contrastive Networks (TCN):** 2 frames taken at the same time from different viewpoints are trained to contrast with a random or hard-negative frame from the same temporal neighborhood using a triplet loss.

End-to-end Self-Supervised Pose Imitation

We apply TCN to the problem of human pose imitation by a robot. With an additional self-supervision signal (Fig. 4), we are able to produce end-to-end imitation without using any labels (Fig. 5). The model is able to learn a complex human to robot mapping entirely self-supervised and is quantitatively better than a human-supervised imitation (Table 2). The combination of all signals (cheap and exact TC and self-supervision, and expensive and noisy human supervision) performs best.

*equal contribution
†Google Brain Residency program (g.co/brainresidency)
[1] sermanet.github.io/tcn
[2] arxiv.org/abs/1704.06888

Figure 2: **Multi-view capture** with two operators equipped with smartphones. Moving the cameras around freely introduces a rich variety of scale, viewpoint, motion-blur and background correspondences between the two cameras.

Method	alignment err.	classification err.
Random	28.1% ± 3.0	54.2%
Inception-ImageNet	27.1% ± 7.7	48.4%
Shuffle & Learn[1]	21.6% ± 6.0	31.0%
single-view TCN (ours)	20.2% ± 6.6	26.8%
multi-view TCN (ours)	**16.3**% ± 5.6	**20.2**%

Table 1: **Pouring alignment and classification errors:** multi-view TCN outperforms all baselines on both metrics. The classification error considers 5 classes related to pouring such as "hand contact with recipient", "container angle", "liquid is flowing", etc.

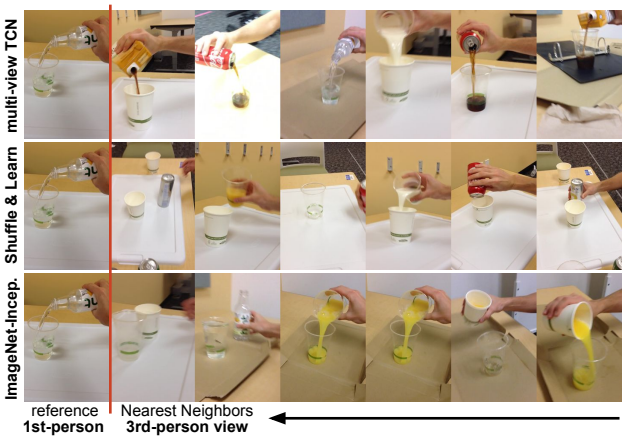

Figure 3: **Nearest Neighbor qualitative comparison** between TCN (top), Shuffle & Learn (middle) and ImageNet-Inception models. On the right, we show the 5 nearest neighbors (from 1st-person perspective test videos) to the reference frame on the left (3rd-person perspective test videos). The TCN nearest neighbors are consistently semantically closer to the reference than the baselines.

References

[1] I. Misra, C. L. Zitnick, and M. Hebert. Unsupervised learning using sequential verification for action recognition. *CoRR*, abs/1603.08561, 2016.

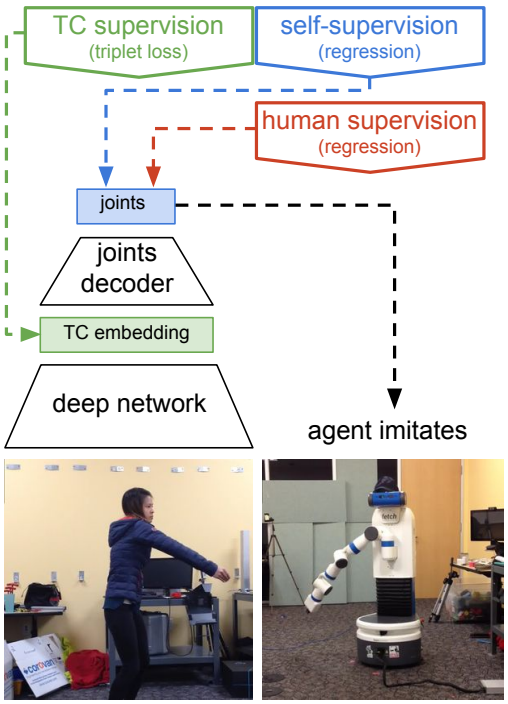

Figure 4: **TCN for end-to-end imitation:** architecture, training and imitation.

Supervision	L2 robot joints error %
Random (possible) joints	42.4 ± 0.1
Self	38.8 ± 0.1
Human	33.4 ± 0.4
Human + Self	33.0 ± 0.5
TC + Self	32.1 ± 0.3
TC + Human	29.7 ± 0.1
TC + Human + Self	**29.5** ± 0.2

Table 2: **Pose imitation error for different combinations of supervision signals.** The error reported is the L2 robot joints distance between prediction and groundtruth, as a percentage error normalized by the possible range of each joint.

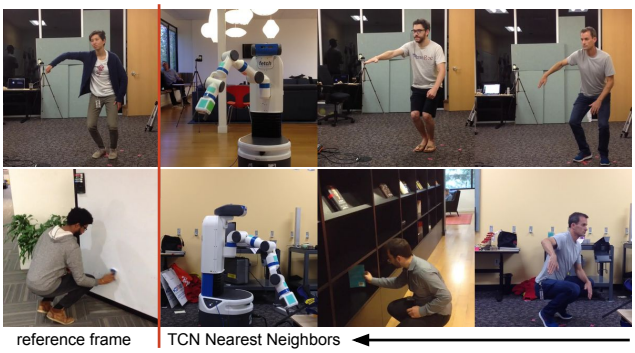

Figure 5: **Label-free Time-Contrastive (TC) embedding**: nearest neighbors in held-out set.