IEEE RO-MAN 2019

TATA CONSULTANCY SERVICES

SMAK-Net: Self-Supervised Multi-level Spatial Attention Network for Knowledge Representation towards Imitation Learning

Kartik Ramachandruni, Madhu Vankadari, Anima Majumder,
Samrat Dutta, Swagat Kumar
TCS Research & Innovation

October 16, 2019

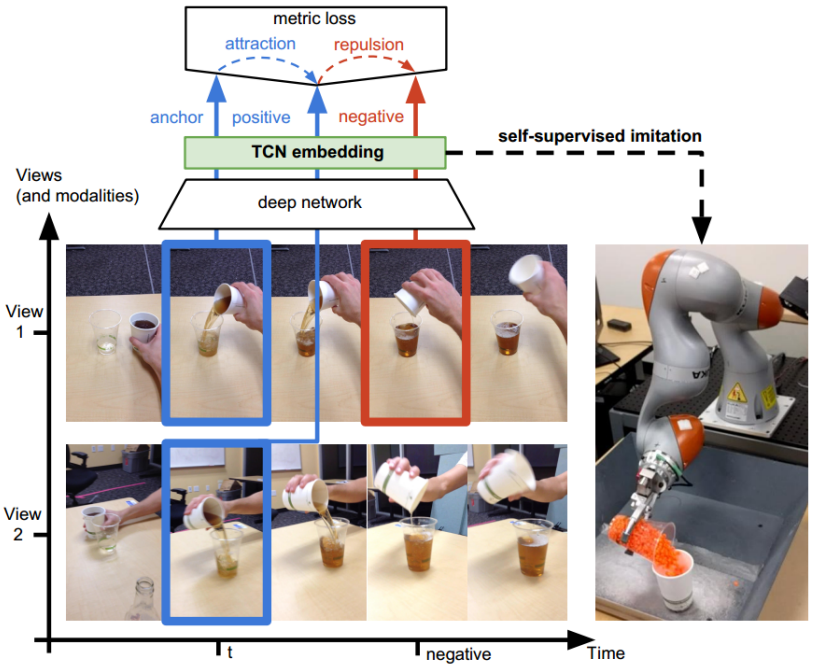
Email ID: {kartik.vr, madhu.vankadari, anima.majumder, d.samrat, swagat.kumar}@tcs.com

Objective of this work

- *Learning from observation*: Generate low dimensional feature representations or embeddings using video demonstrations of tasks
- *Self-supervised approach*: Training data used should not have any explicit labeling of video frames
- *Invariant feature representations*: Embeddings should be invariant to viewpoint, scale and appearance (background variation, nuisance variables, etc.)

State of the Art: Time Contrastive Network (TCN)[1]

- Self-supervised approach to learn low dimensional representations from multi-view video demonstrations
- Metric learning uses concurrent frames as anchor and positive and consequent frame as negative to bring closer in embedding space
- Metric loss function:



$$L = min(D_{an} - D_{ap} - \alpha, 0) \quad (1)$$

where D_{ap} and D_{an} are L2 distances between anchor-positive and anchor negative embeddings, α is distance margin

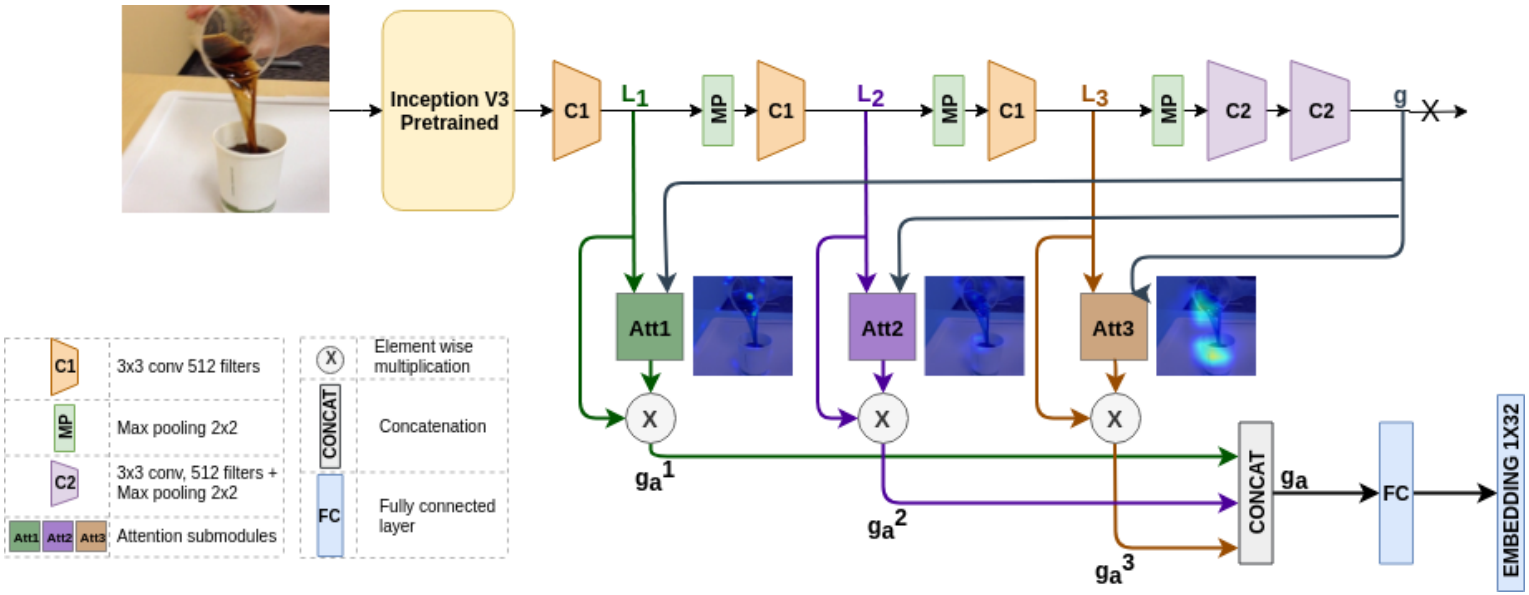
[1] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, Time-contrastive networks: Self-supervisedlearning from video, in 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 11341141, IEEE, 2018.

Shortcomings of TCN

- Bottleneck layer in TCN is a Spatial Softmax layer[2] which outputs the expected position of highest activation within the feature map
- No trainable mechanism is present to ignore redundant information and highlight task-relevant information
- TCN is unable to exploit the variety of contextual and spatial information present at different depths of the CNN pipeline

[2] C. Finn, X. Yu Tan, Y. Duan, Y. Darrell, S. Levine, and P. Abbeel. Learning visual feature spaces for robotic manipulation with deep spatial autoencoders. CoRR, abs/1509.06113, 2015.

SMAK-Net Architecture



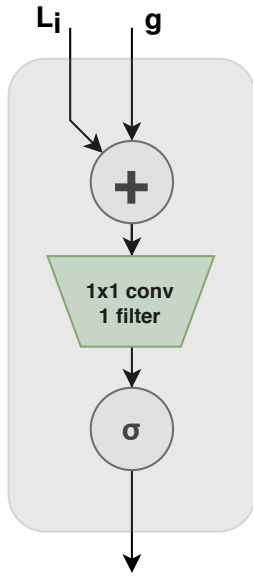
- Multi-level spatial attention module is used to extract spatial and contextual information from different depths of the CNN pipeline
- Attention submodules are trained to highlight the information within feature maps of shallow layers that are relevant to the imitation task

Spatial Attention Sub-module

- Convolutional layer learns compatibility score matrix between local feature map L_i and global feature vector g:

$$c_i =< u, L_i + g >, i \in \{1...n\} \qquad (2)$$

  where u is trainable weight vector (convolution layer), i is the layer from which L_i is extracted
- Element-wise weighted averaging is performed upon local feature map using normalized (sigmoid normalization) compatibility scores as weights



S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, Learn to pay attention, arXiv preprint arXiv:1804.02391, 2018.

Training Strategy

- Multi-view Pouring Dataset: Two-view video demonstrations of pouring task being performed
- Metric learning (N-pairs Loss[3]) is used to bring concurrent frames together in the embedding space
- Performance evaluation: Two validation metrics i.e., temporal alignment error and labeled classification error are used to compare with TCN

[3] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, in Advances in Neural Information Processing Systems, pp. 18571865, 2016.

Quantitative Results

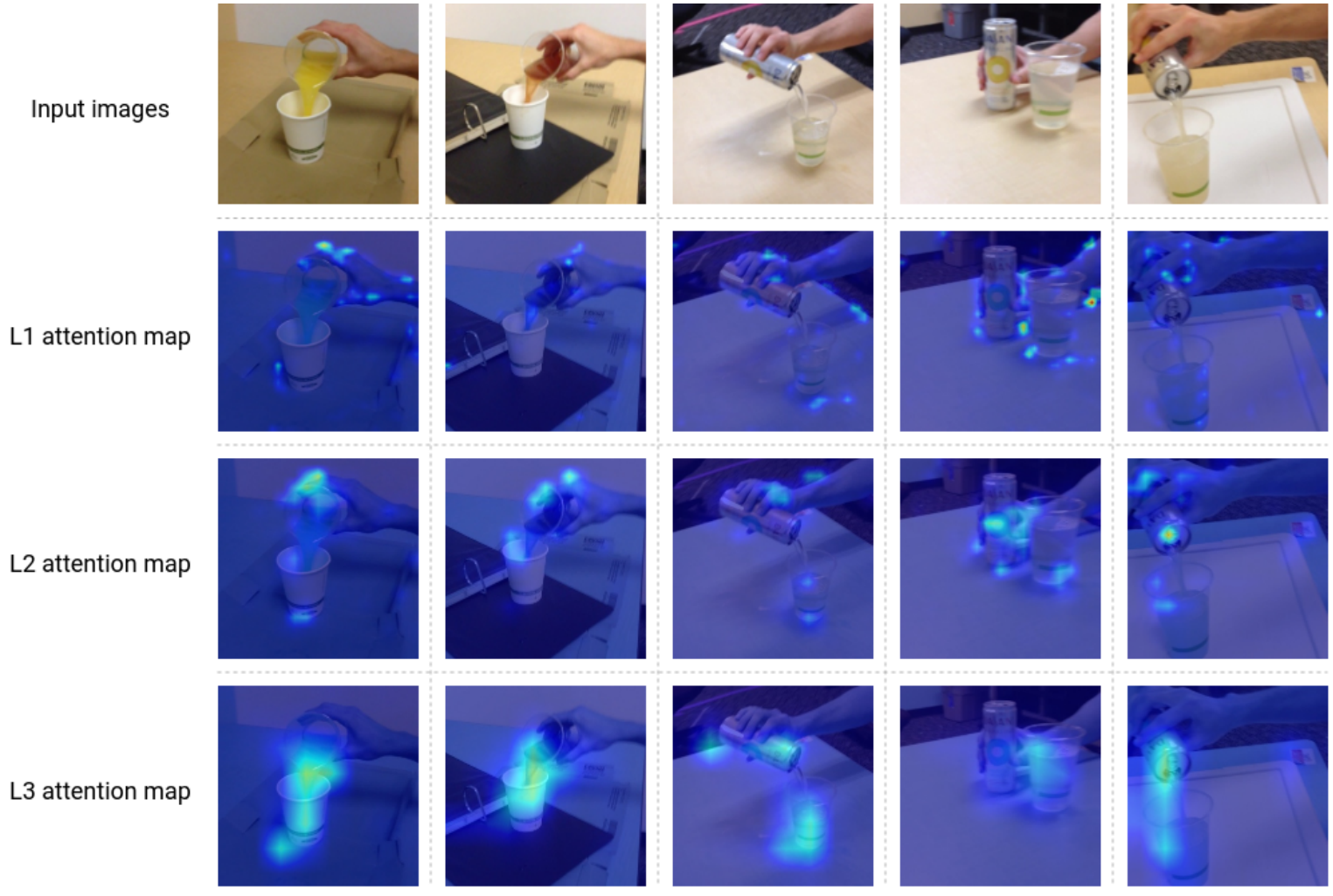
Network Architecture	Training iterations	Alignment error	Classification error
Multi-view TCN (baseline)	224k	17.5%	19.3%
SMAK-Net Att12	69k	14.1%	16.71%
SMAK-Net Att23	69k	13.0%	16.01%
SMAK-Net Att13	69k	12.2%	**15.98%**
SMAK-Net Att123	**69k**	**10.9%**	16.1%

Comparison between SMAK-Net architecture and state-of-the-art TCN. Two validation metrics defined by the latter are used to compare the two methods and a significant improvement is seen in the proposed network. Att12 indicates that only the L1 and L2 attention maps are used to generate the embedding vector. Att23 and Att13 similarly follow this notation and Att123 indicates that all three attention maps are being used.

Parameter	SMAK-Net	SMAK-Net (extended)	SMAK-Net (softmax)	SMAK-Net (Resnet-pretrained)
No. of parameters	1082k	1580k	1082k	1080k
Temporal Alignment error	10.92%	11.9%	13.6%	13.1%
Classification error	16.1%	17.01%	16.64%	18.86%

Ablation study of the network using different network configurations. SMAK-Net denotes the proposed architecture. SMAK-Net (extended) is similar to the proposed architecture but with more convolutional layers. SMAK-Net (softmax) replaces the sigmoid normalization with a softmax normalization. SMAK-Net (Resnet-pretrained) is an architecture similar in structure and number of parameters to the SMAK-Net and is used with a resnet pretrained network.

Qualitative Results



Attention weight maps generated by the multi-level spatial attention module. Top row images are the original input images fed to the network. The other row images are attention maps resized to the input image dimension and superimposed upon the images.

Conclusions

- The proposed CNN based feature representation network called SMAK-Net is trained in a self-supervised manner using multi-level spatial attention module
- The performance of the network demonstrate how spatial attention can aid in generating invariant and information-rich embeddings as well as help in faster generalization (reduction in number of training iterations)
- The accuracy of the proposed SMAK-Net has improved over SOTA by 6.5% and the training time is reduced by 155k iterations
- Ablation studies are also performed to validate the proposed framework configuration

Thank You