Deep Predictive Coding for Multimodal Representation Learning

MSc Project Progress Report

Marcio Fonseca, s1780875

Master of Science
School of Informatics
University of Edinburgh
2018

1 Goals

We set out to investigate the application of Predictive Coding networks (PredNets) to unsupervised learning from videos. Lotter, Kreiman, and Cox (2016) showed that PredNets can learn representations that capture latent variables and explain aspects of scene dynamics such as angular velocities. We hypothesise that PredNets representations can also disentangle and generalise higher-level concepts present in spatiotemporal data. In particular, we believe that PredNets can be a powerful unsupervised learning alternative for tasks that require fine-grained perception of temporal patterns such as video classification and activity recognition.

2 Methods

To test our hypothesis we use the Moments in Time dataset (Monfort et al., 2018), which is a large-scale activity recognition dataset consisting of short videos spanning 339 action/event classes such as ”opening”, ”dancing”, and ”smiling”. The broad semantic coverage and the short length of the videos clips (three seconds) make this dataset appropriate to benchmark the performance of PredNet when transferring the learned representations to different domains.

Our first experiment consists of two small-scale binary classification tasks, which were intentionally designed to assess different aspects of action perception. The first task consists in classifying videos from two classes of Moments in Time that should be easily distinguishable just by identifying the entities involved (e.g., ”cooking” vs ”walking”). The second task requires a more subtle perception of scene dynamics (e.g., distinguishing between ”running” vs ”walking”) and pose more difficulties for models that rely on transfer learning from pre-trained image classifiers.

We use a subset of the Moments in Time dataset consisting of 400 unlabelled videos (sub-sampled at ten frames per second) per class. The PredNet model is trained in an unsupervised way to predict the next frame using a top-down generative model. The errors between predictions and the actual frames are propagated bottom-up to update the prior for new predictions (Lotter et al., 2016). An example sequence of frame predictions is shown in Figure 1.

After the unsupervised training step, we feed frame sequences from a separate dataset to PredNet and extract its internal activations as ”moment embeddings” representing the temporal data, which are used to train a Long Short-Term Memory (LSTM)

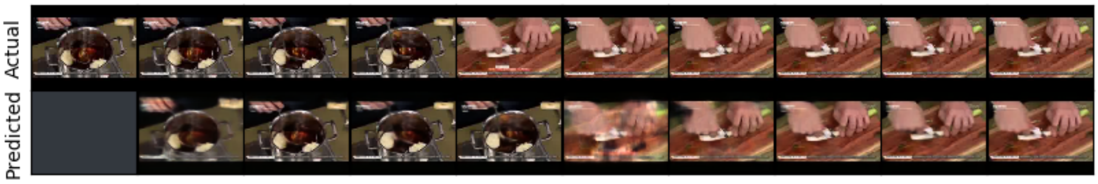

Figure 1: Frame predictions (second row) given by a PredNet pre-trained on the KITTI dataset.

recurrent neural network activity classifier. This supervised step is performed on a labelled dataset consisting of 100, 50, 100 images per class for training, validation, and test splits respectively. We use as a baseline, LSTM classifier architecture trained on the last convolutional features of a pre-trained VGG16 network, an approach that is widely used in the literature (Frermann, Cohen, & Lapata, 2017; Kay et al., 2017).

3 Preliminary results

After training the models according to the description above, we evaluate the classifiers on held-out data. Results are shown in Table 1.

Features	**Easy task (%)**	**Hard task (%)**
VGG (ImageNet)	85.3	48.9
PredNet random weights	52.1	50.0
PredNet KITTI	68.9	52.1
PredNet KITTI + 1200 videos	65.3	52.1
PredNet KITTI + 4000 videos	(*currently running*)	

Table 1: Activity recognition accuracies on test set. The *easy* and *hard* tasks are the binary classification tasks ”cooking” vs ”walking” and ”running” vs ”walking” respectively. PredNet KITTI refers to a model pre-trained on the KITTI dataset (Geiger et al., 2013).

Our main findings are summarised as follows:

- **ConvNet image classifiers do not capture dynamic patterns**: in our experiment, the classifier trained on VGG features perform worse than chance on the ”running” vs ”walking” task, which indicates image classifier features fail to capture information necessary for subtle temporal pattern matching. As expected, the VGG features result in excellent results when entity identification gives a strong hint about the activity (e.g., if food is present in the scene, then a have a strong indicator for the ”cooking” activity).

- **PredNet pre-trained on KITTI dataset performs surprisingly well**: the KITTI dataset contains videos captured from a camera mounted on a car (Geiger et al., 2013), which are unrelated to the action recognition task. Remarkably, the PredNet features give a performance significantly superior to the random baseline on the easy task and consistently better than VGG features on the hard task.

4 Next steps

- **Training PredNet on a larger dataset**: we are currently training the PredNet on a larger set of 4,000 videos to observe how this can improve the classification results reported in Table 1. This dataset has three times the number of frames used by Lotter et al. (2016) and requires around 35 hours of computing time for training on two GPUs.

- **Compare performance with other approaches in the literature**: we plan to test PredNet features on the UCF-101 action recognition dataset (Soomro, Zamir, & Shah, 2012) and compare with baselines reported by Carreira and Zisserman (2017).

- **Explore the audio modality**: our last experiment will apply the same methods described in Section 2 to spectrogram frames extracted from the Moments in Time videos. We expect to observe performance improvements by averaging the results of video and audio modalities.

References

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer vision and pattern recognition (cvpr), 2017 ieee conference on* (pp. 4724–4733).

Frermann, L., Cohen, S. B., & Lapata, M. (2017). Whodunnit? crime drama as a case for natural language understanding. *arXiv preprint arXiv:1710.11601*.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, *32*(11), 1231–1237.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., … others (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.

Monfort, M., Zhou, B., Bargal, S. A., Andonian, A., Yan, T., Ramakrishnan, K., … others (2018). Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*.

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.