Deep Predictive Coding for Multimodal Spatiotemporal Representation Learning

Anonymous Author(s)

Abstract

Common sense reasoning relates to the capacity of *learning representations* that disentangle hidden factors behind spatiotemporal sensory data. In this work, we hypothesize that the predictive coding theory of perception and learning from neuroscience literature may be a promising candidate for implementing such common-sense inductive biases. We build upon the *PredNet* implementation by Lotter, Kreiman, and Cox (2016) and extend its application to the challenging task of inferring abstract, everyday human actions such as *cooking* and *diving*. Our transfer learning experiments demonstrate good generalization of learned representations on the UCF-101 action classification dataset for both visual and auditory modalities.

1 Motivation and Methods

The *PredNet* model by Lotter et al. (2016) was shown to learn representations that disentangle latent variables correlated to the movement of objects in synthetic and natural images. We extend their study to address the following questions: (1) Can unsupervised predictive coding models learn higher-level spatiotemporal concepts, namely quotidian activities such as *driving* or *exercising*? (2) Are predictive coding inductive biases general enough so that these models can also learn from auditory information?

Predictive coding networks Inspired by the predictive coding theory (Friston & Kiebel, 2009), the PredNet model relies on the idea that to predict the next video frame, a model needs to capture latent structure that explains the image sequences. The model architecture consists of recurrent convolutional layers (Xingjian et al., 2015) that propagate bottom-up prediction errors, which are used by the upper-level layers to generate new predictions. For implementation details, please refer to the PredNet architecture description by Lotter et al. (2016).

Unsupervised training We evaluate predictive coding models trained on different quantities of unlabeled videos. The main idea is that the more data we use to train the model, the more "common sense" it should get about how events unfold in the world and, as a consequence, it should be better at disentangling latent explanatory factors. Using as starting point a PredNet pre-trained on the KITTI dataset (Geiger, Lenz, Stiller, & Urtasun, 2013), we further train the model with unlabeled videos (67 hours of visual data and 37 hours of auditory data) from the Moments in Time dataset (Monfort et al., 2018), a large-scale activity recognition dataset.

Supervised action recognition For each sequence of ten frames in the input (video frame or audio spectrogram), the PredNet activations for each layer are spatially pooled to match the higher-level layer dimensions and concatenated to form one tensor representation with dimensions $(16, 20, 339)$ corresponding to a one-second spatiotemporal pattern. Those representations are then flattened and used as input to an action classifier consisting of a Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) layer (64 hidden units) and a fully connected layer followed by a softmax activation that outputs a probability distribution over the UCF-101 action classes.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

2 Results and Discussion

Action recognition using visual data The predictive coding model with random weights gives a poor top-1 accuracy of 1.64%, which is slightly above the random baseline (Table 1). However, when we train the classifier with features generated by the 67-hour predictive coding model, the accuracy increases to 51.9%, which is competitive with results from the unsupervised "tuple verification" by Misra, Zitnick, and Hebert (2016) and an LSTM classifier using the Inception convolutional network (Carreira & Zisserman, 2017). It is worth to note, however, that in both of those approaches, the convolutional models are fine-tuned end-to-end using the UCF-101 labels. In our case, the predictive coding weights were kept fixed, and only the weights from the LSTM classifier were optimized for the specific task.

Predictive coding can also model auditory data We trained the LSTM classification model using predictive coding representations extracted using audio spectrograms from the 51 action classes of the UCF-101 dataset that contain auditory information. The top-1 accuracy results are reported in Table 2. As expected, the audio information is much less useful to distinguish action classes, as many videos have soundtracks and other kinds of audio data that are completely unrelated to the activity. Still, there was a significant improvement from the classifier trained on the features generated by the random-weights model to the classifier based on the 37-hour pre-trained model. For comparison, we also report the results of the Caffenet version by Wang, Yang, and Meinel (2016), which is a convolutional network trained on audio spectrograms. Remarkably, our simple one-layer LSTM classifier is competitive with their complex convolutional model trained end-to-end using action class labels, which demonstrates the generality of the predictive coding inductive bias.

Table 1: Visual action recognition. Accuracies (top-1 percentage) for different pre-trained models on the test set of UCF-101 split 1. We also include results for the CNN tuple verification (Misra et al., 2016) and an LSTM classifier trained on top an Inception convolutional network trained from scratch (Carreira & Zisserman, 2017).

Features + Classifier	Accuracy (%)	Pre-training dataset
PredNet Video random + LSTM	1.64	-
PredNet Video 67h + LSTM	51.9	Moments in Time
CNN tuple verification	50.2	UCF-101
Inception + LSTM	54.2	-

Table 2: Auditory action recognition. Accuracies (top-1 percentage) for different models on the test set of UCF-101 split 1 (only videos from the 51 classes that contain audio).

Features + Classifier	Accuracy (%)	Pre-training dataset
PredNet Audio random + LSTM	22.7	-
PredNet Audio 37h + LSTM	24.8	Moments in Time
Caffenet (Wang et al., 2016)	25.2	-

3 Final Remarks

This work explores unsupervised learning from spatiotemporal data and uses video understanding tasks as a proxy to evaluate the quality of learned representations. We focus on models that can learn from large amounts of unlabeled videos and use this experience to solve downstream tasks involving smaller labeled datasets. Therefore, we *do not* pursue the solution of the action recognition problem itself, for which all the state-of-art approaches depend on a copious amount of labeled data for pre-training (Carreira & Zisserman, 2017). Our results show that predictive coding representations learned using the Moments in Time dataset (Monfort et al., 2018) (without in-domain fine-tuning) are competitive with other unsupervised baselines when evaluated on UCF-101 action recognition task.

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).

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221.

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.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

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

Misra, I., Zitnick, C. L., & Hebert, M. (2016). Shuffle and learn: unsupervised learning using temporal order verification. In *European conference on computer vision* (pp. 527–544).

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*.

Wang, C., Yang, H., & Meinel, C. (2016). Exploring multimodal video representation for action recognition. In *Neural networks (ijcnn), 2016 international joint conference on* (pp. 1924–1931).

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems* (pp. 802–810).