Deep Predictive Coding for Spatiotemporal Representation Learning

Marcio Fonseca

Master of Science

Cognitive Science and Natural Language Processing

School of Informatics

University of Edinburgh

2018

Abstract

TODO

Acknowledgements

Many thanks to my mummy for the numerous packed lunches; and of course to Igor, my faithful lab assistant.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Marcio Fonseca*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Learning common sense from the world with a limited amount of data and supervision is a remarkable capability of human cognition. Notably, infants in their second year of life experience an impressive increase in the vocabulary acquisition known as vocabulary explosion (McMurray, 2007). This cognitive phenomenon builds upon several previous cognitive milestones, such as understanding object permanence (Baillargeon et al., 1985), spatiotemporal continuity (Spelke et al., 1994), inertia, and gravity, and other knowledge about world dynamics that require extensive observation and interaction with the environment, and are essential building blocks of human-like intelligence.

In contrast, recent success in Artificial Intelligence is mostly circumscribed to probabilistic models that capture patterns from massive amounts of static, human-curated datasets. Most of the time, these models learn to infer from a probabilistic distribution over symbolic entities (e.g., language sentences) or raw sensory data (e.g., images and sounds) to a set of high-level symbolic categories relevant to a particular problem. As a result, the performance of these models is bounded by the size and quality of the datasets and, crucially, they exhibit insufficient capacity of generalising what they learn to solve novel tasks. In particular, distinguishing everyday events such as walking, running, and exercising is an open problem in computer vision research (Carreira and Zisserman, 2017; Monfort et al., 2018).

In fact, current AI approaches lack the innate machinery infants use to make sense of environmental dynamics. Such inductive biases should exploit known regularities of the world such as the constancy of the laws of physics, as proposed by Srivastava et al. (2015). Also, an intelligent agent should be able to reason about the temporal order of events, which can serve as an unsupervised learning signal for spatiotemporal data (Misra et al., 2016). In this work, we are interested in a more general, neuroscience-

inspired inductive bias in which the brain is portrayed as a hierarchical machine that improves its internal model of the world by processing the error between predicted and actual sensory stimuli. This *predictive coding* model of the human brain has been applied in theoretical neuroscience to explain processing in the visual cortex (Rao and Ballard, 1999), perceptual categorisation (Friston and Kiebel, 2009), and in cognitive science as a unified account of perception and action (Clark, 2013).

To investigate the design of machines that acquire common sense by observing the world, we capitalise on a deep learning implementation of the predictive coding model published by Lotter et al. (2016). Their deep predictive coding network was shown to learn representations that disentangle latent variables correlated to the movement of objects in synthetic and natural images. We extend the study of such models to address the following questions:

- Can unsupervised predictive coding models learn higher-level spatiotemporal concepts, namely quotidian activities such as *driving* or *exercising*?
- Are predictive coding inductive biases general enough so that these models can also learn from auditory information?

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 unlabelled videos and use this experience to solve downstream tasks involving smaller labelled 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 labelled data for pre-training and often the combination of handcrafted features to encode temporal patterns (Carreira and Zisserman, 2017). Our main contributions are summarised as follows:

- We extend the work of Lotter et al. (2016) by using predictive coding representations to decode higher-level concepts that require understanding of the world dynamics. The learned representations are evaluated on small-scale tasks and on UCF-101 (Soomro et al., 2012), a popular action recognition benchmark.
- We train the predictive coding model on a much larger dataset, about 60 times larger than previous work (Lotter et al., 2016) and show that model continues to improve future frame predictions, even when the training dataset includes a large number of unrelated classes.

- Inspired by sensory substitution literature from neuroscience (Stiles and Shimojo, 2015), a novel application of the predictive coding model is proposed for unsupervised representation learning from audio data. Our multimodal predictive coding model performs better than the video and audio-only versions, suggesting that the different modalities provide complementary information that is useful for the action classification task.

The rest of this work is structured as follows. In chapter 2 we succinctly introduce the predictive coding model of the brain and its general formulation, free-energy principle. This review allows to appreciate the elegance and ambition of the free-energy theory and also understand some limitations of our neural network model in approximating the complex dynamical hierarchical processing proposed by Friston and Kiebel (2009). In chapter 3, we describe the methods we use to train predictive coding models, extract spatiotemporal representations, and use them to train downstream action recognition models. Chapter 4 details the experiments used to quantitatively and qualitatively evaluate the usefulness of predictive coding representations. In chapter 5, we discuss the primary experimental results and propose future research directions. Finally, chapter 6 presents a recapitulation of our findings and final remarks.

Chapter 2

Background

2.1 The free-energy principle (FEP)

2.2 From FEP to predictive coding

2.3 Deep Learning meets predictive coding

2.4 Action recognition models

Chapter 3

Methods

Our goal is to empirically determine if predictive coding networks provide effective useful biases to capture patterns from spatiotemporal data. In particular, learned representations should disentangle explanatory variables for the observed inputs and also be useful for downstream supervised models (Bengio et al., 2013). In this chapter, we detail the methods used to test if predictive coding representations exhibit those properties. First, we describe the unsupervised pre-training step, including the datasets used to generate each predictive coding model that is used in all other experiments. Then, we explain how representations are extracted and used by the action recognition classifiers.

3.1 Unsupervised pre-training

The first part of the experiments consists in training the predictive coding architecture using an unlabelled action recognition dataset. The main idea is that the more data we use to train the model, the more ”common sense” it should get about the world and, as a consequence, it should be better at solving other tasks.

As stated in our project proposal, the Moments in Time dataset (Monfort et al., 2018), which is a large-scale activity recognition dataset, is our dataset of choice for obtaining unlabelled videos. This dataset is particularly suited to this experiment because it has a broad semantic coverage of actions (339 actions/events) and also a fixed duration of three seconds for each video, which allows us to curate two balanced subsets with 3 hours (10 actions) and 67 hours (200 actions) of video data. Each of these datasets is used to train different versions of the unsupervised model, which are compared to a model with random weights (no training) and a version trained on the KITTI

dataset (Geiger et al., 2013), kindly provided by Lotter et al. (2016). A summary of the different pre-trained models is shown in Table 3.1.

Model name	Dataset	Frames	Hours	Action classes
PredNet random	-	0	0	-
PredNet KITTI	KITTI	$\approx 41K$	≈ 1	-
PredNet Moments 3h	Moments in Time	$\approx 120K$	≈ 3.3	10
PredNet Moments 67h	Moments in Time	$\approx 2.4M$	≈ 66.6	200

Table 3.1: Versions of the predictive coding models trained used across the experiments.

As explained in Section 2.3, the predictive coding 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. We follow the same training configuration used in the original PredNet implementation proposed by Lotter et al. (2016), with four modules consisting of 3x3 convolutional layers with 3, 48, 96, and 192 filters. The videos are subsampled at ten frames per second, and the each network input is a sequence of ten frames for which the model generates ten frame predictions. The model is trained via backpropagation to minimise the weighted sum of the activity of the error layers as defined as follows:

$$L_{train} = \sum_t \lambda_t \sum_l \frac{\lambda_l}{n_l} \sum_{n_l} E_l^t, \tag{3.1}$$

where E_l^t represents the errors units in layer l at timestep t. The layerwise weights λ_l impose a smaller penalty for higher-level layers ($\lambda_0 = 1, \lambda_{>0} = .1$), which according to Lotter et al. (2016) results in better predictions. All timestep weights λ_t were set to zero, except for the first timestep.

3.2 Extracting spatiotemporal representations

We follow the same approach described by Lotter et al. (2016) to extract representations from each layer of the predictive coding model. For each sequence of ten frames in the input, the R_l activations are read for each layer l, which are then spatial 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. The idea of using "deep" representations that include activations from all layers is similar to recent unsupervised learning approaches in natural language processing such as Embeddings from Language Models (ELMo) (Peters et al., 2018). We assume that each layer might learn representations that reflect different timescales and the downstream classification model can learn to weight each layer according to the specific task. We leave the study of how each layer contributes to the action recognition task for future work.

3.3 Supervised action recognition

Once the one-second predictive coding representations were extracted, each moment representation corresponding to one second was flattened and used as input to an action classifier (Figure 3.1). As we are focused on the relative quality of representations and not state-of-the-art performance, we chose a simple linear support vector machine (SVM) to compare different predictive coding models. Since the linear SVM is a non-probabilistic model, video-level predictions are made by majority voting of predictions for each one-second feature.

Additionally, a recurrent neural network classifier was used to assess the importance of modelling longer timespans. In this case, a Long Short-Term Memory (LSTM) layer (Hochreiter and Schmidhuber, 1997) with 64 hidden units received a sequence of five overlapping one-second moment representations totalling three seconds, corresponding the duration of each video in the Moments in Time dataset. For videos longer than three seconds (from the UCF-101 dataset), the video-level predictions were calculated by averaging the predictions for each three-second clip. Table 3.2 shows the architecture of the neural network classifier.

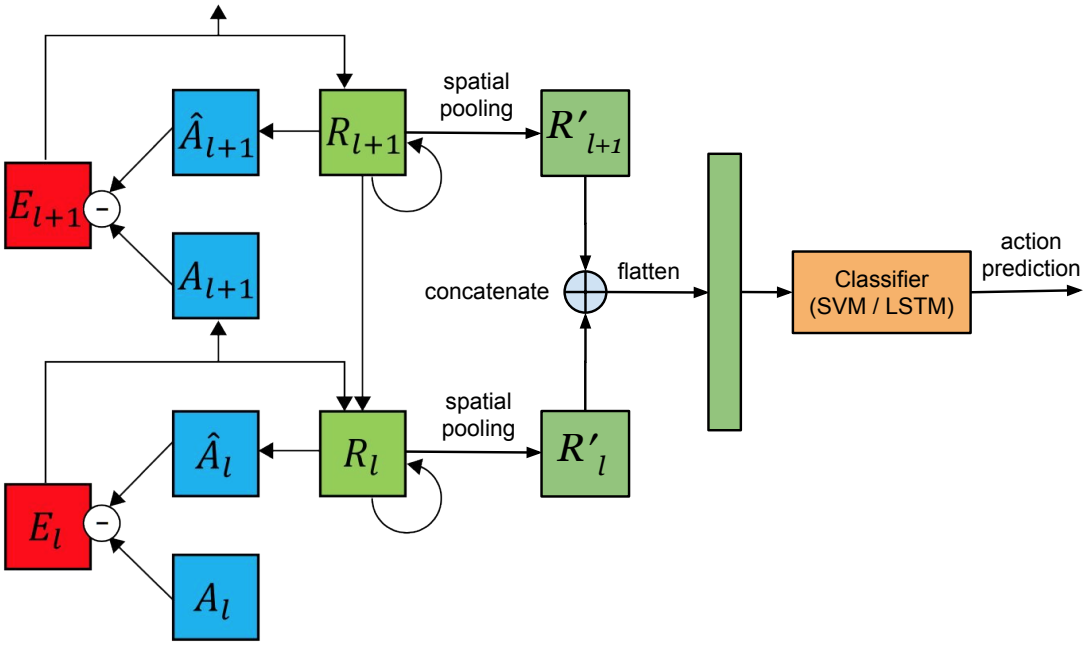

Figure 3.1: Action classification architecture showing only two predictive coding layers. Errors E_l are calculated between the layer input A_l and the prediction $\hat{A}_l$. Representations R_l for each layer are concatenated after a spatial pooling operation. The resulting tensor is flattened and passed as input to an action classifier. Adapted from Lotter et al. (2016) Figure 1.

Layer name	Output shape
Input	(5,16,20,339)
Flatten	(5,108480)
LSTM	(64)
Dense	(number of classes)
Softmax	(number of classes)

Table 3.2: Recurrent neural network classifier layers.

Chapter 4

Experiments

In this chapter, we detail the experiments we use to assess the quality of predictive coding representations. We first pre-train the predictive coding network on varying sizes of unlabelled video datasets and evaluate the learned representations on small-scale action recognition tasks. Then we investigate a novel application of the same predictive coding architecture to learn representations from audio spectrograms. Finally, we perform a transfer learning experiment and evaluate the audio and video representations on a widely used action recognition benchmark, the UCF-101 dataset (Soomro et al., 2012).

4.1 Next-frame predictions

According to Lotter et al. (2016), to predict the next frames a model needs to build an internal model that explains the movements of objects present in a given scene. Therefore, the most straightforward method to evaluate learned representations is to measure the quality of the generated predictions. Since the evaluation of generative models is a complex subject by itself (Theis et al., 2015), we follow Lotter et al. (2016) and use simply the mean squared error (MSE) between the predicted and actual frames as a quantitative measurement of prediction fidelity.

After training the predictive coding models as described in Section 3.1, we calculate the mean square error of predictions on a held-out dataset of 1000 videos spanning 10 action classes, as well as the results for a naive baseline that merely copies the last frame. The results in Table 4.4 show that as we add more data, the generative model yields better frame predictions, reducing the MSE error by 29.8% relative to the baseline. Also, it is clear that the improvement starts to plateau, giving only about 3% of

improvement when increasing the dataset size from three hours to 67 hours of video. However, it is worth to note that the Moments 67h dataset has videos from 200 different classes, including spatiotemporal information that differ significantly from the 10-class evaluation set. Therefore, we observe that the model continues to learn even when exposed to out-of-domain spatiotemporal data.

Model name	MSE	Relative change
Copy last frame	0.00795	0
PredNet random	0.14422	+1711.3%
PredNet KITTI	0.00816	+2.6%
PredNet Moments 3h	0.00581	-26.9%
PredNet Moments 67h	**0.00558**	**-29.8%**

Table 4.1: Evaluation of different pre-trained models on a held-out set of 1000 videos spanning 10 action classes. Relative changes are computed in relation to the "copy last frame" baseline.

Inspecting sample frame predictions is also useful to qualitatively analyse how the model generalises spatiotemporal concepts. Figure 4.1 shows the final five frame predictions from a ten-frame sequence sampled from the *exercising* class. Interestingly, the predictions given by a predictive coding network with random weights (second row) are not random and seem to copy some of the most salient features of the previous frame without any colour information preserved. This observation illustrates the power of the predictive coding inductive bias and also explain why the randomly initialised model performs surprisingly well on action recognition tasks (see sections 4.2, 4.3, and 4.4) despite the poor performance regarding the MSE metric reported in Table 4.4.

The frames generated by the model trained on the KITTI dataset (third row) accurately predict the colour information but exhibit a significant amount of blurriness in the moving portions of the image (e.g., the woman's legs). Most importantly, there is still a strong bias towards copying the contents of the previous frame. On the other hand, the predictions from the model trained on 67 hours of the Moments in Time dataset (fourth row) are less blurry and show a remarkable capacity of generalisation of scene dynamics. If we observe the position of the woman's knees relative to the workout top in the last two frame predictions, we verify that the KITTI model prediction is very similar to the previous frame, while the Moments model extrapolates and

Figure 4.1: Last five frame predictions from a 10-frame timestep sequence sampled from the *exercising* class. First row shows ground truth frames. Second and third rows show predictions for predictive coding model trained on KITTI dataset, and trained on 67h hours of videos from the Moments in Time dataset respectively. KITTI model predictions are blurrier and tend to be more similar to the previous frame.

”imagines” that legs should be at a higher position, much closer to the actual frame.

4.2 Small-scale action recognition

While the next-frame prediction analysis from Section 4.1 indicates that the predictive coding model can generalise spatiotemporal patterns, there is still no evidence that the learned representations are transferable to other domains. To address this question, we investigate the usefulness of predictive coding representations to decode high-level action concepts such as *walking*, *exercising*, and *cooking*. In particular, we are interested in scenarios in which there is a limited amount of labelled data that are expensive to obtain, and a large number of unlabelled videos cheaply available for the unsupervised predictive coding model.

Intuitively, we know that some of the actions are readily determined by the objects that appear in the scene while other events require fine-grained distinction of object dynamics. For instance, if objects such as food and cookware are identified in a given frame, one could readily infer that the activity is *cooking*. In contrast, to differentiate between *walking* and *running*, the details of the temporal dynamics of the entities present in the scene are crucial, making the classification task very challeng-

ing for current video understanding models. In fact, most of the state-of-the-art action classification approaches rely on hand-crafted features such as optical flow to encode temporal information, as discussed in Section **??**.

Our small-scale experiment uses 100, 50, and 100 videos per class for training, validation, and testing respectively. We create two binary classification tasks, which are intentionally designed to explore different challenges of action perception discussed above. The binary *spatial* task consists in classifying videos from *cooking* and *walking* classes, which should be effortlessly distinguishable just by identifying the entities involved in the action. Similarly, the binary *temporal* task uses the target actions *running* and *walking*, which should require a subtler perception of scene dynamics. We also report results for classification tasks involving ten classes. The unsupervised predictive coding model is allowed to use the training split but never touches the validation and test sets.

In the linear SVM classifier, we use the squared hinge loss, a penalty parameter $C = 1.0$, stopping tolerance $1e{-}4$, and L2 regularisation. The neural network classification models used an LSTM layer with 512 hidden units followed by a fully connected layer. cross-entropy loss for training and the parameters were optimised using the Adam gradient-based algorithm (Kingma and Ba, 2014). Since we are dealing with small datasets and high-dimensional moment representations, an aggressive dropout regularisation of 0.9 was applied (Srivastava et al., 2014). Early stopping was used to stop training after the validation loss stopped improving for ten epochs. The SVM and LSTM classifiers were implemented using the scikit-learn (Pedregosa et al., 2011) and Keras (Chollet et al., 2015) open-source libraries respectively.

Classification accuracies are listed in Table 4.2, which also include baseline models using features extracted from a VGG16 model (Simonyan and Zisserman, 2014b) pre-trained on the ImageNet dataset (Krizhevsky et al., 2012). In the following paragraphs, we discuss the most relevant results.

When models pre-trained on Imagenet fail Confirming the findings of previous work (Carreira and Zisserman, 2017), pre-training models on ImageNet gives a substantial improvement in classification performance, especially on the binary spatial and 10-class tasks, for which the identification of objects is determinant. However, there is a significant drop in accuracy on the binary temporal task, which indicates the model falls short of capturing fine-grained temporal patterns needed to distinguish between *running* and *walking* actions. For this reason, many of the state-of-the-art approaches

Features + Classifier	2-class spatial	2-class temporal	10-class
VGG random + SVM	67.0	56.0	18.7
VGG ImageNet + SVM	85.5	67.0	**52.8**
VGG ImageNet + LSTM	**87.4**	58.4	43.2
PredNet random + SVM	67.6	62.6	30.1
PredNet KITTI + SVM	73.2	**70.7**	39.8
PredNet Moments 3h + SVM	73.2	66.1	39.5
PredNet Moments 67h + SVM	74.2	65.1	41.4
PredNet Moments 67h + LSTM	81.6	55.8	42.9

Table 4.2: Evaluation of different pre-trained models on a held-out set of 100 videos per class. First set of rows list results for a baseline model using features extracted from the VGG16 convolutional image classifier (Simonyan and Zisserman, 2014b).

for action recognition are based on ”two-stream” models (Simonyan and Zisserman, 2014a; Carreira and Zisserman, 2017), which integrate hand-crafted temporal features such as optical flow and raw RGB frames.

Predictive coding temporal inductive bias Representations generated by predictive coding models with random weights outperform random VGG features in all tasks, which indicates that predictive coding incorporates inductive biases that are better suited to spatiotemporal perception. Remarkably, the features extracted from the predictive coding model trained on only 41K unlabelled frames from the KITTI dataset gives the best performance on the binary temporal task, outperforming by a significant margin the VGG model pre-trained on more than one million labelled images (Simonyan and Zisserman, 2014b).

The more data, the better As we add more data to the predictive coding model, the performance on the binary spatial task improves while the accuracy on the binary temporal task falls consistently. We believe this performance drop is due to the introduction of a large number of unrelated classes in the 3h (10 classes) and 67h (200 classes) versions, as the performance on the 10-class seems to improve nicely with larger datasets. Nevertheless, a more careful experiment would be required to confirm this hypothesis.

Modelling longer time spans In our experiments, the LSTM models use sequences of five VGG/moment representations spanning three seconds of spatiotemporal data, which results in better performance on the binary spatial task compared to a simple average of SVM predictions for each representation. On the other hand, performance is severely affected on the binary temporal task for both types of representations. Again, further investigation is needed to understand the reason behind these results.

4.3 Exploring the audio modality

The human brain possesses an extraordinary capacity to adapt its anatomical and functional structure in response to environmental changes, which is a phenomenon known as neuroplasticity in the neuroscience literature (Draganski et al., 2004; Maguire et al., 2000). One of the most exciting practical applications that leverage this property is the possibility of compensating sensory loss by encoding sensory data in a format that can be read by another peripheral sensory organ. For instance, recent work has shown the applicability of encoding visual information into sounds and tactile information to help blind people to see objects (Bach-y Rita and Kercel, 2003; Stiles and Shimojo, 2015).

Inspired by these sensory substitution experiments, we propose the use of the same predictive coding network designed for action recognition from visual stimuli to process audio data. The idea is to encode the audio information in a format that takes advantage of the capacity of predictive coding networks to capture spatiotemporal patterns, as demonstrated by experiments in sections 4.1 and 4.2. We build upon previous work that uses audio spectrograms images and convolutional neural networks for speech processing (Abdel-Hamid et al., 2014; Zhang et al., 2017) and impose a spatiotemporal coherence between subsequent audio frames. The resulting audio representation is a video showing a scrolling spectrogram with the same duration as the original video. Thus, we recast the audio classification problem as a spatiotemporal pattern recognition which is suitable for the predictive coding architecture.

To generate the scrolling spectrograms we used `ffmpeg` tool (Bellard et al., 2000) `showspectrum` filter with Hanning window function (Harris, 1978) and overlap equal to zero. The generated videos had the same dimension and were sampled at the sample frame rate as the visual data so that we could use precisely the same neural network architecture, feature extraction and training procedure applied to the visual modality in the experiments described above. An important difference, however, is that the audio datasets are significantly smaller, containing around 60 videos per class compared to

100 videos per class for the visual modality.

After training the unsupervised predictive coding model on 2 hours and 37 hours of auditory data, we found the performance pattern in the frame-prediction evaluation was very similar to the video modality results (see Table 4.3). In fact, the relative reduction of the best predictive coding model with respect to the naive baseline was 92.4% (versus 29.8 for the video modality), which suggests that the prediction of audio spectrograms is easier than predicting frames in natural images. This idea is also supported by samples predicted frames (Figure 4.2), which clearly show a pronounced difference in quality between prediction from the models trained on the KITTI and Moments datasets. We believe the low resolution of the spectrogram frames and the simple dynamics of the spatiotemporal pattern (merely scrolling from right to left) are decisive factors for the excellent quality of predictive coding predictions.

Regarding the small-scale action recognition experiment (Section 4.2), the results in Table 4.3 show that adding more training data to the unsupervised training step leads to improvement on both binary tasks. On the 10-class task, there was no improvement in accuracy over the model with random weights, which gave a surprisingly strong baseline score. Notably, all predictive coding models followed by an SVM classifier performed better than the classifiers based on features from a VGG model pre-trained on ImageNet. In this case, the weights learned from an image recognition task do not generalise well to the spectrograms domain, and the inductive biases provided by the predictive model proved to be useful. Due to the small amount of data and extreme overfitting, LSTM classifiers were not used in this experiment.

Model name	MSE	Relative change
Copy last frame	0.011193	0
PredNet random	0.079011	+605.9%
PredNet KITTI	0.011392	+1.8%
PredNet Moments 2h	0.000930	-91.7%
PredNet Moments 37h	**0.000856**	**-92.4%**

Table 4.3: Evaluation of different pre-trained models on a held-out set of 622 videos spanning 10 action classes. Relative changes are computed in relation to the ”copy last frame” baseline.

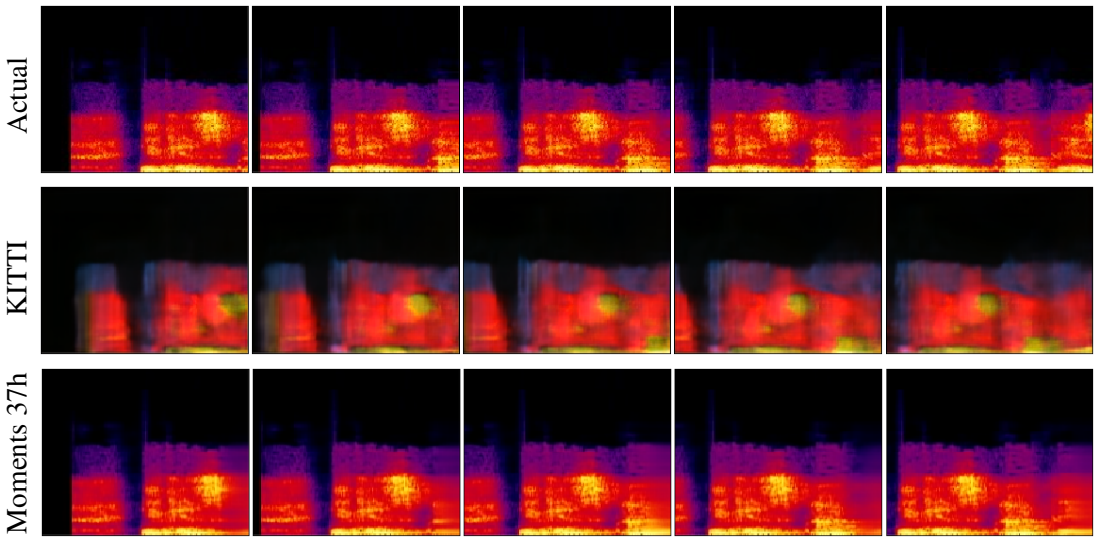

Figure 4.2: Last five spectrogram frame predictions from a 10-frame timestep sequence sampled from the *speaking* class. First row shows ground truth frames. Second and third rows show predictions for predictive coding model trained on KITTI dataset and trained on 37h hours of videos from the Moments in Time dataset respectively. Predictions for the model with random weights are omitted as they are mostly black frames.

4.4 Transfer learning on UCF-101

All the experiments so far were based on a small subset of the Moments in Time dataset specifically designed to test our hypothesis concerning the effects of different amounts of pre-training data on various tasks. One question that remains to be explored is how predictive coding representations compare to other approaches in action recognition literature. To address this question, we chose the UCF-101 dataset (Soomro et al., 2012), a widely used action recognition benchmark with 13,320 video clips spanning 101 action classes. Besides the vast amount of published results, this dataset is suitable for our experiment because it contains only 27 hours of video, which is still manageable for training without expensive distributed computing across several GPUs.

In this experiment, we used our best predictive coding models pre-trained on 67 hours of visual and 37 hours of auditory information from Moments in Time videos with no further fine-tuning. Thus, the task requires the model to generalise patterns learned from the Moments in Time dataset to a new dataset with unseen spatiotemporal concepts.

Features + Classifier	2-class spatial	2-class temporal	10-class
VGG ImageNet + SVM	57.9	53.0	24.7
PredNet Audio random + SVM	63.6	56.8	**30.3**
PredNet Audio KITTI + SVM	62.0	50.8	29.4
PredNet Audio Moments 2h + SVM	66.9	56.8	29.1
PredNet Audio Moments 37h + SVM	**67.8**	**58.3**	30.0

Table 4.4: Evaluation of different pre-trained models on a held-out set of 60 videos per class. First set of rows list results for a baseline model using features extracted from the VGG16 convolutional image classifier (Simonyan and Zisserman, 2014b).

4.4.1 Next-frame predictions

Again, we resort to next-frame predictions to get a qualitative assessment of the model understanding of action dynamics, namely for the *CliffDiving* action from UCF-101. As shown in Figure 4.3, while the model can predict the overall changes such as camera movement that causes the occlusion of the platform, guessing the detailed pattern of the diver's body is difficult, and the prediction degenerates to a blurry blob. Besides the multiple hidden causes of movement, this example introduces extra challenges such as the complex and swift movements in small portions of the image (the diver's acrobatics). To isolate these factors, we chose another sample that shows only the body movements occupying a larger portion of the image and no camera shifts. The predictions in Figure 4.4 show that the model still falls short of predicting the body movements, suggesting that high-speed image changes are particularly difficult for the predictive coding model. Future work to address this issue might include more training data containing complex body movements and working with frame rates higher than we used in these experiments (10 frames per second).

4.4.2 Action recognition

Spatiotemporal representations were extracted as described in Section 3.2 and a recurrent neural network with one LSTM layer (64 hidden units) followed by a fully connected layer was used as an action classifier. The classifier received as input sequences of five predictive coding representations corresponding to 3 seconds of video. For videos with more than 3 seconds of duration, the final video-level predictions were

Figure 4.3: Last five spectrogram frame predictions from a 10-frame timestep sequence sampled from the *CliffDiving* class. First row shows ground truth frames. Second and third rows show predictions for predictive coding model trained on KITTI dataset and trained on 67h hours of videos from the Moments in Time dataset respectively. The model captures the overall camera movement and the position of the diver but falls short of figuring out the finer details.

Figure 4.4: Last five spectrogram frame predictions from a 10-frame timestep sequence sampled from the *CliffDiving* class. First row shows ground truth frames. Second and third rows show predictions for predictive coding model trained on KITTI dataset and trained on 67h hours of videos from the Moments in Time dataset respectively. Even without camera movements, the model fails to predict the body movements and tends to copy the previous frames.

the average of all 3-second predictions (for more details refer to Section 3.3). Table 4.5 lists the top-1 accuracies on the test set of UCF-101 first split, which contains 9537 clips for training and 3783 clips for testing.

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 4.5: Evaluation of different pre-trained models on test set of UCF-101 split 1.

As opposed to the small-scale classification experiments in Sections 4.2 and 4.3, the predictive coding model with random weights gives a poor performance of 1.64%, which is slightly above the random baseline. 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 et al. (2016) and an LSTM classifier using the Inception convolutional network (Carreira and Zisserman, 2017). It is worth to note however that in both of these 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.

We also trained the auditory predictive coding model on the 51 action classes of the UCF-101 dataset that contains audio information. The top-1 accuracy results are reported in Table 4.6. 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 37h pre-trained model. For comparison, we also report the results of the Caffenet version by Wang et al. (2016), which is a convolutional network trained on spectrograms that have no spatial coherence across frames. 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 predictive coding representations.

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	-

Table 4.6: Evaluation of different pre-trained models on test set of UCF-101 split 1 (only videos from the 51 classes that contain audio).

Chapter 5

Discussion and Future work

Our empirical analysis supports the hypothesis of predictive coding as a powerful inductive bias for learning common sense by observing how the world dynamics unfold. The simple predictive coding architecture proposed by Lotter et al. (2016) can predict future frames in a way that suggests a good level of generalisation across different actions, surpassing by a significant margin a baseline predictor that copies the previous frame. Additionally, next-frame analysis proved to be a useful tool to inform the investigation of improvements to the model. For instance, we found predictions are degraded when fast-moving entities appear in the scene, suggesting that temporal resolution used in training was not adequate.

From the action recognition experiments, we learned that for some activities such as *running* and *walking*, the perception of fine-grained dynamics is fundamental, and in these cases, the predictive coding model has an edge on the traditional convolutional approaches. However, there are many cases in which the identification of the entities present is informative enough to determine the action. For instance, the *cliff diving* activity shown in Figures 4.3 and 4.4 could be readily identified as diving by a human even by inspecting a single frame. For this reason, the state-of-the-art action recognition approaches will continue to rely on pre-trained image classifiers. However, there is an opportunity to replace the handcrafted temporal streams (Carreira and Zisserman, 2017) with data-driven approaches such as the predictive coding model.

In the out-of-domain action classification using the UCF-101 dataset, a simple one-layer LSTM classifier on top of predictive coding representations performed on par with deep convolutional models fine-tuned to the visual (Misra et al., 2016; Carreira and Zisserman, 2017) and auditory information (Wang et al., 2016). While these results are far from state-of-the-art action recognition models, our approach learns represen-

tations that are potentially more general and transferable to other tasks for which annotated data is expensive to obtain, since we do not fine-tune the unsupervised model using task-specific labels. Nevertheless, the experimental data revealed limits in the predictive coding model's capacity to learn fine-grained spatiotemporal patterns. In the next sections, we discuss further research direction that may help to address these issues.

5.1 Scaling predictive coding training

One of the most important conclusions of our experiments is that the quality of predictive coding representations improve as we add more data in the pre-training step. In fact, our larger pre-training dataset of 67 hours of video (2.4 million frames) despite being around sixty times larger than the original implementation by Lotter et al. (2016), it is small compared to many recent action recognition datasets such as the full Moments in Time dataset with one million videos (over 800 hours) (Monfort et al., 2018) or the Sports-1M dataset (Karpathy et al., 2014).

Even for lower dimensional natural language processing data, successful unsupervised learning approaches consume a vast amount of computing resources, with pre-training steps lasting over one month on eight GPUs (Radford et al., 2018). Unfortunately, our current Keras-based implementation is not optimised for parallelisation and to further improve the performance given by predictive coding pre-training, a full re-engineering is required to leverage more recent Tensorflow (Abadi et al., 2015) APIs that deliver more efficient parallelisation and input pipelines.

5.2 Multimodal predictive coding

In this work, the visual and auditory models were trained separately and their predictions combined straightforwardly. However, experimental results in neuroscience suggest that the interaction between modalities is much more complex and can occur even at early stages of visual and auditory cortical processing (Ghazanfar and Schroeder, 2006; Falchier et al., 2002). The predictive coding architecture can be extended to explore such forms of early fusion (Snoek et al., 2005). Since the layer predictions already integrate representations from errors from the current layer and representation from upper layers, would be straightforward to add representations from other modalities at each level. This extension would allow empirical priors (Friston, 2010) learned

from each modality condition next-frame predictions, resulting in potentially more efficient multisensory integration.

As already suggested in our project proposal, we hypothesise that multimodal tasks involving videos such as cross-modal retrieval (Aytar et al., 2017) and grounded language learning (Hermann et al., 2017) can benefit from a more sophisticated sensory integration model. In future work, we intend to explore the application of predictive coding representations in such problems.

5.3 Integrating action

One of the most important consequences of the free-energy principle is that action is naturally integrated with perception and learning. To minimise the variational free-energy, the agent not only has to improve its perceptual capabilities by also act to improve its model of the world (Friston and Kiebel, 2009; Clark, 2013). Furthermore, under this framework, the agent has to sample the data, or more generally, act to realise predictions ”that are biased toward preferred outcomes” (Friston et al., 2017), a concept known as active inference. If the multimodal experiments suggested in Section 5.2 are successful, the next step would be to investigate the feasibility of modelling actions and even attentional models (Koelewijn et al., 2010) as additional modalities.

Chapter 6

Conclusion

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545.

Aytar, Y., Vondrick, C., and Torralba, A. (2017). See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*.

Bach-y Rita, P. and Kercel, S. W. (2003). Sensory substitution and the human–machine interface. *Trends in cognitive sciences*, 7(12):541–546.

Baillargeon, R., Spelke, E. S., and Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, 20(3):191–208.

Bellard, F. et al. (2000). Ffmpeg. `https://www.ffmpeg.org`. Accessed Aug 01 2018.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Carreira, J. and 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*, pages 4724–4733. IEEE.

Chollet, F. et al. (2015). Keras. `https://keras.io`. Accessed Aug 01 2018.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., and May, A. (2004). Neuroplasticity: changes in grey matter induced by training. *Nature*, 427(6972):311.

Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience*, 22(13):5749–5759.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, 29(1):1–49.

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

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

Ghazanfar, A. A. and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in cognitive sciences*, 10(6):278–285.

Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83.

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., et al. (2017). Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*.

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

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta psychologica*, 134(3):372–384.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

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

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., and Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8):4398–4403.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, 317(5838):631–631.

Misra, I., Zitnick, C. L., and Hebert, M. (2016). Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer.

Monfort, M., Zhou, B., Bargal, S. A., Andonian, A., Yan, T., Ramakrishnan, K., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., et al. (2018). Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79.

Simonyan, K. and Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576.

Simonyan, K. and Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM.

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

Spelke, E. S., Katz, G., Purcell, S. E., Ehrlich, S. M., and Breinlinger, K. (1994). Early knowledge of object motion: Continuity and inertia. *Cognition*, 51(2):131–176.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Srivastava, N., Mansimov, E., and Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852.

Stiles, N. R. and Shimojo, S. (2015). Auditory sensory substitution is intuitive and automatic with texture stimuli. *Scientific reports*, 5:15628.

Theis, L., Oord, A. v. d., and Bethge, M. (2015). A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.

Wang, C., Yang, H., and Meinel, C. (2016). Exploring multimodal video representation for action recognition. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 1924–1931. IEEE.

Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C. L. Y., and Courville, A. (2017). Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720*.