Deep Predictive Coding for Multimodal Representation Learning

Informatics Project Proposal

Marcio Fonseca, s1780875

Master of Science
School of Informatics
University of Edinburgh
2018

Abstract

A significant amount of unlabelled video datasets covering many kinds of events and natural phenomena is readily available to enrich machine learning models with common sense knowledge. However, existing approaches lack the necessary inductive biases humans use to learn temporal and multimodal representations from correlated sensory data. Although recent deep learning models inspired by the *predictive coding* theory from neuroscience have shown promising results in unsupervised learning from videos, the application of these techniques to multimodal data is still unexplored in the literature. This project proposal aims to fill this gap by extending the predictive coding approach to the problem of representation learning from multimodal temporal data. In particular, we suggest a methodology for evaluation of learned representations using supervised tasks such as activity recognition and natural language understanding using aligned information of video, audio, and text.

1 Introduction

Deep Predictive Coding networks, inspired by the *predictive coding* literature from neuroscience (Friston & Kiebel, 2009; Rao & Ballard, 1999), frame the unsupervised learning problem as the capacity of predicting future sensory data in a sequence. These networks can predict complex object movements in synthetic and natural videos, resulting in learned representations that are useful for estimating latent variables such as steering angle in an autonomous vehicle setting (Lotter, Kreiman, & Cox, 2016).

Given this successful application on videos, we hypothesise that representations learned using the predictive coding approach could lead to better performance in multimodal tasks involving temporal data, such as cross-modal retrieval (Aytar, Vondrick, & Torralba, 2017), grounded language learning (Hermann et al., 2017), and action/event recognition in videos (Monfort et al., 2018). Many of the models used in these tasks use a naive approach to extract features from video and audio, using the last layer activations of pre-trained classifiers, which indicates that the problem of learning high-quality representations from multimodal temporal data is not sufficiently explored in the literature.

In this proposal, we aim to fill this gap by extending the deep predictive coding network introduced by Lotter et al. (2016) to work with multimodal data. We suggest two experiments to evaluate learned representations in different timescales. The first experiment is based on the activity recognition task introduced by Monfort et al. (2018), which will assess our model capacity to capture patterns from video and audio data spanning three seconds. In the second experiment, we intend to test our model in a more complex task presented by Frermann, Cohen, and Lapata (2017), which involves multiple inferences over several minutes and an additional linguistic modality.

The remainder of this proposal is organized as follows. We start by briefly presenting related research concerning predictive coding theory (Section 2), then we state the motivation and elaborate the planned approach (Section 3) and the evaluation criteria (Section 4). Finally, a detailed work plan is suggested in Section 5.

2 Background

The brain as a *prediction machine* is an emerging view in theoretical neuroscience and cognitive science literature. Rao and Ballard (1999) proposed a model of visual processing consisting of feedback connections carrying predictions for the next

lower-level activity and feedforward connections that send back the error between predictions and actual activations. These prediction errors convey information to guide the refinement of future predictions and, as a result, explain the latent structure of sensory information. Their hierarchical predictive model exhibited some of the extra-classical receptive field inhibition effects observed in previous studies of the visual cortex, and suggested that neurons with those properties can be regarded as ”residual error detectors”. Friston and Kiebel (2009) extend this idea by indicating that perception corresponds to the inversion of *hierarchical dynamical models* to explain sensory information.

The hierarchical prediction machine interpretation is also an active area of research in cognitive science. Clark (2013) argues that the predictive coding theory may offer a ”unifying model of perception and action”, and also account for attentional phenomena, by stating that all these components are in the same ”family business” of minimising ”sensory prediction error from our exchanges with the environment”. This unified framework of mind, brain, and action gives valuable insights on how to integrate computational machine learning elements, Bayesian a priori constraints and implementation details at the neural level.

2.1 Predictive Coding in Machine Learning

The predictive coding formulation inspired machine learning implementations aiming to observe how these models perform on real data empirically. Chalasani and Principe (2013) propose a hierarchical *linear dynamic model* that can adapt the priors according to the context. The central idea of their model is that the objective of each layer is to predict the representation of the layer immediately below, using information from layers above and temporal information from previous states. Lotter et al. (2016) introduce a model inspired on similar principles, but instead of using greedy layer-wise training, they rely on more recent deep learning approaches such as recurrent convolutional networks trained end-to-end with backpropagation. We detail the main components of Lotter et al. (2016) model in Section 3.2.

2.2 Multimodal Machine Learning

The sensory information we receive from the environment is not only correlated in space in time but also across different *modalities* such as visual, auditory, and tactile information. While there is extensive literature covering many multimodal machine

learning models, here we are mainly interested in the techniques used to produce multimodal representations. According to Baltrušaitis, Ahuja, and Morency (2017), multimodal representations can be classified into two categories: joint and coordinated. Joint representations are obtained by applying a function f (e.g., a neural network) that maps unimodal representations to multimodal representations. Alternatively, coordinated representations result from the application of a constraint between the unimodal representations obtained by projection functions f and g:

$$f(u_1) \sim g(u_2),$$

where the coordination constraint $\sim$ enforces some measure of similarity (e.g., cosine similarity) or a specific structure between the unimodal representations.

These concepts, which we intend to study in more detail during the literature review task (Section 5), will be relevant when considering the best strategies to generate a joint representation for the predictive coding errors.

3 Motivation and approach

Our research project concerns the application of machine learning models based on the predictive coding theory to multimodal tasks. The novel contribution of our work will be to determine if a hierarchical predictive coding architecture can learn representations using prediction errors over *joint* visual and auditory sensory data and not only from unimodal data as reported in previous works. We believe that a more principled technique to learn representations from multimodal temporal data is fundamental to improve the performance on tasks that require common sense reasoning such as natural language understanding, conversational AI, and robotics.

3.1 Impact and future research

A successful application of predictive coding model would be relevant to leverage a large amount of cheaply available video datasets conveying information about real-world dynamics that could be transferred to many domains. Furthermore, the model predictions can be used to study how the system behaves over time and compare with results from cognitive science literature such as:

- **Attentional models**: there is a strong interest in modelling attentional models that could lead to more efficient perceptual processing. Judd, Durand, and

Torralba (2012) introduce a new dataset with eye tracking data and point out the majority of visual saliency models only account for bottom-up processing, which is easier o model. As suggested by Clark (2013), attention fits nicely in the interplay between top-down and bottom-up in predictive coding.

- **Decision making**: many studies in neuroscience try to uncover the neural basis of decision making (Yang & Shadlen, 2007). An interesting research direction could relate predictive coding errors with neural activations that correlate with accumulation of evidence over time.

- **Semantic priming**: priming is an effect in which a stimulus (e.g., a picture) influence reaction time to subsequent stimuli (e.g., pictures or words) depending on their semantic relationship (Sperber, McCauley, Ragain, & Weil, 1979). An experiment could investigate how prediction errors are affected by the presentation of subsequent stimuli with different degrees of semantic similarity.

3.2 Methodology

We build on the *PredNet* implementation by Lotter et al. (2016), which was shown to perform well on unsupervised learning tasks using video data. Inspired by the predictive coding theory, their model relies on the idea that to predict the next video frame, a model needs to capture relevant latent structure that explains the image sequences. The PredNet architecture (see Figure 1) consists of recurrent convolutional layers that propagate bottom-up prediction errors which are used by the upper-level layers to generate new predictions. Equations (1) to (4) show how each module is calculated.

$$A_l^t = \begin{cases} x_t & \text{if } l = 0 \\ \textsc{MaxPool}(\textsc{ReLU}(\textsc{Conv}(E_{l-1}^t))) & l > 0 \end{cases} \tag{1}$$

$$\hat{A}_l^t = \textsc{ReLU}(\textsc{Conv}(R_l^t)) \tag{2}$$

$$E_l^t = [\textsc{ReLU}(A_l^t - \hat{A}_l^t); \textsc{ReLU}(\hat{A}_l^t - A_l^t)] \tag{3}$$

$$R_l^t = \textsc{ConvLSTM}(E_l^{t-1}, R_l^{t-1}, \textsc{Upsample}(R_{l+1}^t)) \tag{4}$$

The model loss function is defined by the weighted sum of layer errors as follows:

$$L_{train} = \sum_t \lambda_t \sum_l \frac{\lambda_l}{n_l} \sum_{n_l} E_l^t, \tag{5}$$

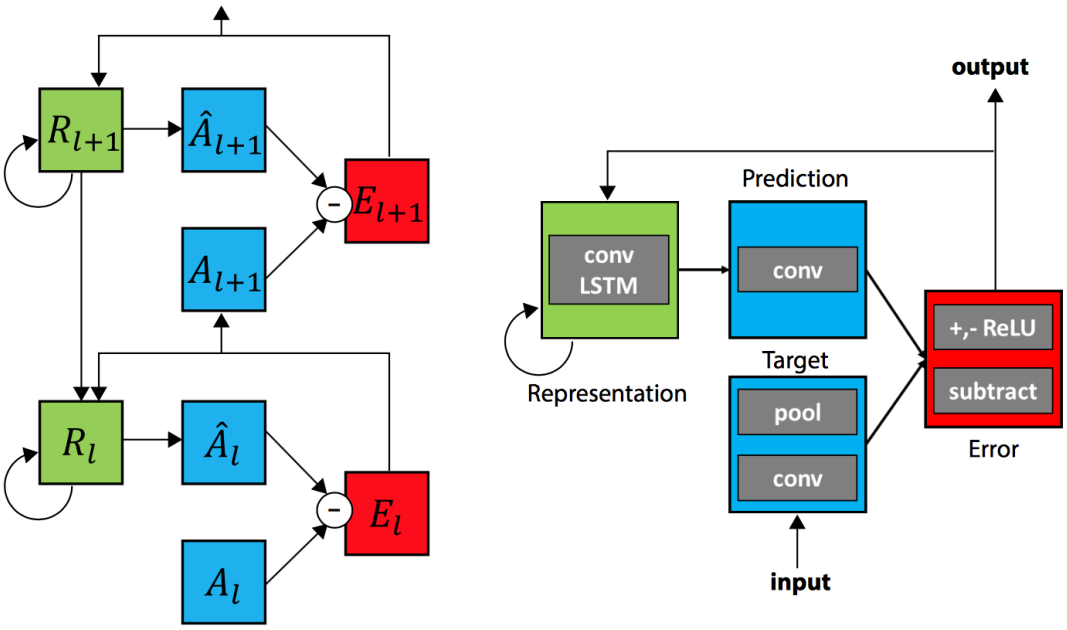

Figure 1: PredNet architecture. Reprinted from Lotter et al. (2016), Figure 1.

where λ_t and λ_l hyperparameters that define the weights for the errors for each timestep and hierarchical level respectively.

Our preliminary approach will consist of one independent predictive coding network for each modality and then create joint representations of the errors E_l by stacking visual and auditory errors in different channels and propagating this information to an upper predictive coding layer with similar architecture. We expect to try several variations of this basic model and select the best according to the performance on the validation set.

To evaluate the quality of learned multimodal representations, we intend to use next-frame-based evaluation metrics proposed by Lotter et al. (2016) (see Section 4) and also, we choose two supervised tasks that require the inference of latent structure of scenes. The selected tasks are based on datasets that are publicly available and have a reasonable size that allows investigation within the time constraints of the project. Moreover, we give preference to datasets that have reference baseline implementations that will be extremely valuable to assess our results. In next sections, we present the two tasks we set out to apply the predictive coding approach.

3.3 Moments in Time Dataset

The Moments in Time Dataset (Monfort et al., 2018) consists of one million human-annotated videos covering a diverse set of activities and events that unfold within three seconds. The dataset was designed to include short videos that convey sufficient spatio-temporal-auditory information to interpret the world dynamics and match the average duration of human working memory. Each video is annotated with one of 339 actions corresponding to the most frequently used verbs in the English language, and the average number of videos per class is 1,757.

In addition to the original dataset, Monfort et al. (2018) provide a smaller version specifically designed for the student track of the *Moments in Time Challenge 2018* hosted at the Conference on Computer Vision and Pattern Recognition 2018[1]. This dataset has only 200 classes and 100,000 training videos and seems to have an ideal size to be explored in the context of this project.

One interesting characteristic of the Moments in Time dataset is that a significant proportion of classes reflect sound-dependant events such as ”clapping in the background”. We believe these cases will provide valuable insights on the contribution of each modality to model performance.

3.4 *Whodunnit*

The short videos in the Moments in Time dataset will be useful to evaluate the capacity of our model to identify basic activities that span three seconds. However, we are interested in investigating if the model can also capture longer-term phenomena that potentially involve the composition of one or more basic activities. To this end, we propose a second experiment, in which we test our predictive coding model on a harder language understanding problem.

Frermann et al. (2017) formulate a supervised learning task in which a machine learning model is trained to infer who are the crime perpetrators on episodes of a crime drama television show. They create dataset consisting of aligned video, audio, and screenplay text from 39 episodes of *CSI: Crime Scene Investigation* and annotate with labels indicating if the crime perpetrator was mentioned in each scene.

Apart from the paired linguistic data, the *whodunnit* task has some significant differences from the Moments in Time task. While the problem involves only a binary labelling task, each *CSI* episode is approximately 43 minutes long, which means the

[1] Available at http://moments.csail.mit.edu/challenge.html

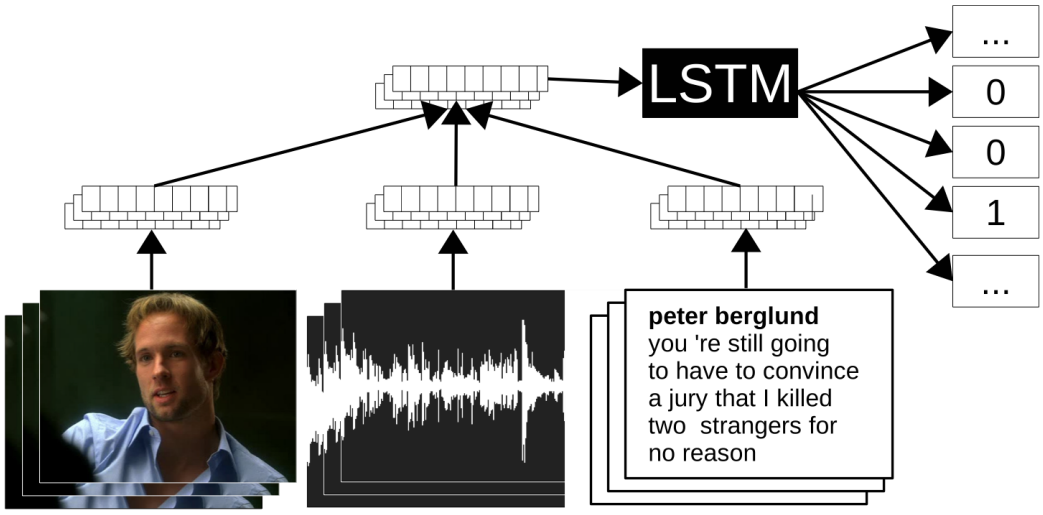

Figure 2: The *whodunnit* problem modelled as a multimodal sequence labelling task. We propose a predictive coding model to learn better representations for the video and audio modalities (first two on the left). Reprinted from Frermann et al. (2017), Figure 4.

model has to account for much longer-term information to derive beliefs about the perpetrator. Also, it is possible to track how these beliefs change over time and draw comparisons with human behaviour.

As in the Moments in Time case, we have access to the LSTM model implementation designed by Frermann et al. (2017) (see Figure 2), which use as video features the last layer activations of an Inception-v4 convolutional network pre-trained on static object classification task. We believe that predictive coding models could perform better in this task because they are specifically designed to handle temporal data and, most importantly, they can be pre-trained in an unsupervised way on cheaply available unlabelled video datasets.

4 Evaluation

We propose evaluation metrics for unsupervised and supervised tasks. Regarding the unsupervised case, we note the PredNet model is also a generative model that predicts new frame images in each step. To evaluate the quality of frame predictions, we will use the mean square error (MSE) and the Structural Similarity Index Measure (SSIM) (Wang, Bovik, Sheikh, & Simoncelli, 2004), which suitable to capture the *perceptual* difference between predicted and actual frames. As in Lotter et al. (2016), we will use

these metrics to compare different model architectures against a naive baseline such as one that copies the last frame.

For the action classification task, we will apply the same top-1, and top-5 accuracies used to evaluate the Moments in Time baselines. In particular, Monfort et al. (2018) recommend the use of top-5 accuracy because more than one action could take place during the three seconds of a video. The top-k accuracy is defined by the percentage of samples for which the correct label is among the k best-ranked predictions.

In the *whodunnit* task, we will use *precision*, *recall*, and F_1 metrics, which are useful for the proposed unbalanced binary classification problem. In particular, *precision* is defined by the fraction of inferred perpetrator mentions that are true positive mentions and *recall* is the fraction of true perpetrator mentions the model correctly identified as a perpetrator mention. The F_1 score is the harmonic mean of *precision* and *recall*:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{6}$$

If we are successful in training the models for both tasks, we will also have the opportunity to conduct transfer learning experiments. Specifically, we expect that some of the ordinary activities found in the Moments in Time dataset also appear in parts of the *CSI* episodes and, as a consequence, a model pre-trained on the activity recognition task should result in better performance or, at least, faster convergence in the *whodunnit* task. Thus, we plan to include in our results performance metrics for models that derive video and audio representations using three approaches: (1) directly from a predictive coding model pre-trained on the Moments in Time dataset only, (2) fine-tuning the same pre-trained model on the *CSI* episodes, (3) training a predictive coding model from scratch on the *CSI* episodes only.

5 Work plan

A work plan over approximately fifteen weeks is shown in Figure 3. We briefly describe each task as follows:

- **Literature review**: in this task, we focus on the details of the model we plan to implement, especially the one described by Lotter et al. (2016).
- **Implement baseline model**: setup the environment and adapt the reference code provided by Lotter et al. (2016) to perform a small-scale experiment.

- **Dataset preparation**: preprocessing routines such as extracting video frames and audio spectrograms, and splitting data into training, validation and test sets.
- **Model implementation**: model implementation for the specific task, including data input pipelines, loss functions and evaluation metrics.
- **Model tuning**: perform hyperparameter search, investigate model variations, and collect experimental data for reporting. This task allows overlapping with other tasks since we expect each full experiment to last several hours.
- **Initial draft writing**: write the first draft of the dissertation.
- **Supervisor review**: supervisor suggests corrections and improvements for the final version.
- **Final version writing**: improvements are made based on supervisor feedback.

5.1 Risk analysis

Since our proposal deals with approaches that are relatively unexplored in the literature, in this section, we anticipate some hindrances that might emerge in the course of our project, including the corresponding mitigation measures.

5.1.1 Predictive coding model is too complex

Probability: low — **Impact:** high.
One primary concern is that translating the predictive coding theory into a machine learning model could be hard, especially the free energy formulation by Friston and Kiebel (2009). Fortunately, the model proposed by Lotter et al. (2016) is straightforward both conceptually (see Section 3.2) and regarding implementation. In fact, Lotter et al. (2016) provide a model implementation using the Keras framework[2], which offers high-level APIs that facilitates extension and fast experimentation.

5.1.2 Models take too long to train

Probability: medium — **Impact:** medium.
Recurrent models trained on video and audio datasets usually consume a large amount of computing resources, which may impact the planned deadlines (see Figure 3). To

[2]Available at https://coxlab.github.io/prednet

counter this issue, we chose a smaller version of the Moments in Time dataset, which has about 10% of the original data. Also, we can reduce the scope of experiments for the *whodunnit* task, by using only a model pre-trained on the Moments in Time to extract features for the *CSI* episodes. Therefore, our experiments allow some margin of adaptation to unforeseen obstacles during the project.

6 Conclusion

In this proposal, we present a new potential application of predictive coding models for unsupervised learning using multimodal datasets. We select two datasets/tasks that are useful for evaluation of multimodal representations and suitable for the project resource and time constraints. We believe that the successful implementation of such unsupervised techniques could result in a better transfer of knowledge and a significant impact on tasks such as language understanding and robotics.

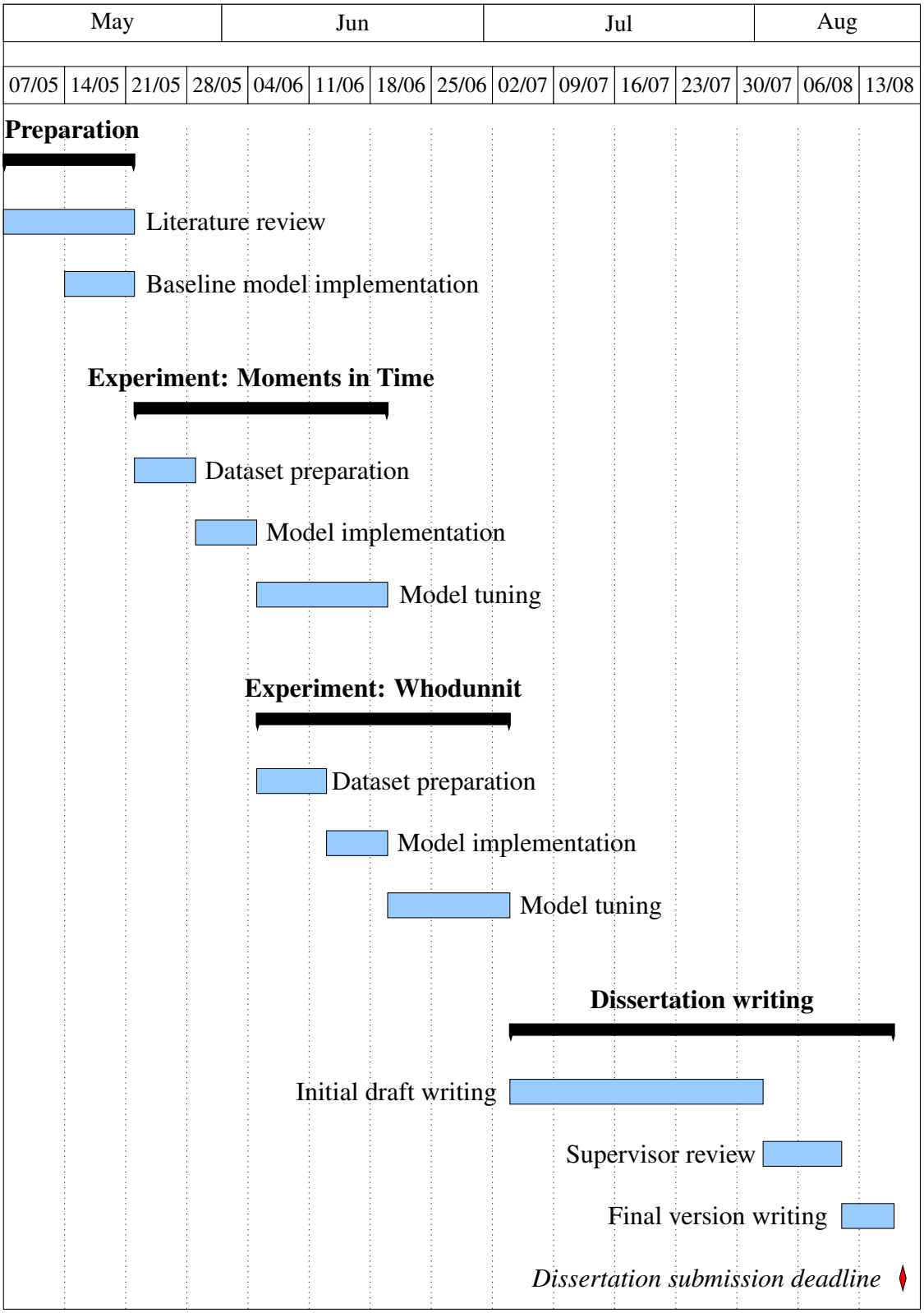

Figure 3: Project plan.

References

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

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2017). Multimodal machine learning: A survey and taxonomy. *arXiv preprint arXiv:1705.09406*.

Chalasani, R., & Principe, J. C. (2013). Deep predictive coding networks. *arXiv preprint arXiv:1301.3541*.

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

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

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.

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., ... others (2017). Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*.

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations.

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*.

Rao, R. P., & 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.

Sperber, R. D., McCauley, C., Ragain, R. D., & Weil, C. M. (1979). Semantic priming

effects on picture and word processing. *Memory & Cognition*, *7*(5), 339–345.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, *447*(7540), 10751080. doi: 10.1038/nature05852