REPRESENTING DYNAMICALLY: AN ACTIVE PROCESS FOR DESCRIBING SEQUENTIAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose an unsupervised method for building dynamic representations of sequential data, particularly of observed interactions. The method simultaneously acquires representations of input data and its dynamics. It is based on a hierarchical generative model composed of two levels. In the first level, a model learns representations to generate observed data. In the second level, representational states encode the dynamics of the lower one. The model is designed as a Bayesian network with switching variables represented in the higher level, and which generates transition models. The method actively explores the latent space guided by its knowledge and the uncertainty about it. That is achieved by updating the latent variables from prediction error signals backpropagated to the latent space. So, no encoder or inference models are used since the generators also serve as their inverse transformations. The method is evaluated in two scenarios, with static images and with videos. The results show that the adaptation over time leads to better performance than with similar architectures without temporal dependencies, e.g., variational autoencoders. With videos, it is shown that the system extracts the dynamics of the data in states that highly correlate with the ground truth of the actions observed.

1 INTRODUCTION

When observing a particular interaction some behaviors tend to repeat over time following specific dynamics. Understand such behaviors and differentiating the dynamics that define them is a relevant task that allows to characterize the interactions, acquire knowledge from them, and build reactive systems that adapt to evolving situations. Those interactions could be, for example, human activities captured in video, data from a vehicle-mounted camera, or the motion of an agent of interest in a given environment.

If we consider a video in which different kinds of action sequences can be observed, the task we aim at for a learning system would be to separate the diverse types of dynamics in the observed sequences and embed them in representational states. For example, imagine a camera mounted on a car. In a video from such device, one would observe predictable changes on the frames for particular actions, for instance, there would be a certain dynamics when the car goes straight, and other where it curves. Similarly, when observing a human performing different kinds of actions in a given scenario, the dynamics followed by the person for each action is to be differentiated. However, such separations should be performed actively during an on-line observation, therefore the system, and particularly its internal representations, should adapt dynamically to the changing data.

A viable process to achieve that goal implies representing every observation, e.g., each frame, so that estimating the dynamic evolution of such representations consequently means abstracting the dynamics of the observations. For example, when observing people performing sets of actions, one would describe frames regarding the position and the pose of the person acting. Nonetheless, given an unsupervised framework, such information is not available. So, the way in which the representations are defined is to depend on the observed dynamics. That is so since the relevance of what is to be represented comes from its relation to the evolution of the observations, e.g., the actions being executed.

Therefore we define our primary goal as to the acquisition of representational states observations and their dynamics simultaneously in an unsupervised way.

Accordingly, the definition of representations is central and determines how learning is to be understood. In particular, representational states are to be defined as dynamic and capable of adjusting themselves to changing environments and uncertainty in sensory data. Taking into account such constraints, we consider an active process where changes in the world and internal states play a primary role interpreting the observed data. That is in opposition to an entirely passive process where an input is transformed into static representations, e.g., a classifier. So, the representational process should be understood in the temporal domain as a mechanism that responds to perceived changes.

To implement that, it would be necessary that a system, e.g., a neural network (NN), adapts itself over time to the observed data. With NNs the primary way to achieve similar behaviors is through recurrent networks. When recurrence is involved, the states of previous time steps affect the interpretation of the current inputs, which could include information from different time steps as in the case of NNs based on LSTM units (Hochreiter & Schmidhuber, 1997).

Nonetheless, it is also possible to define the adaptability of a NN regarding its predictive accuracy. In general, the prediction error of the network's output is only used for adapting the NN during training by adjusting its parameters through, for example, backpropagation. However, a more dynamic view would include a capability of such kind as part of the inference process. That is, the NN could benefit from a feature that allows it to modify some of its internal states dynamically to model the sensed data based on feedback from its prediction error in a backpropagation-like way. That would allow an active process in which the interpretation of the environment depends also on previous states, or beliefs, therefore making the system capable of actively adapting to changing scenarios.

The ideas of actively interpreting and adapting to observed data coincide with dynamic views on conceptual representations in the cognitive science. From such perspectives, representations are seen more as dynamic and distributed structures than as symbols or static categories. In particular, Kiefer & Pulvermüller (2012) conclude from neuroimaging studies that concepts might be flexible, experience-dependent and modality-specific, while distributed across the sensory-motor systems. In particular, for them, the flexibility is crucial for the capability of adapting to diverse situations.

Olier et al. (2017b) elaborate on how the definition of concepts has evolved and how it impacts the way in which learning is understood, and how that consequently affects the design of artificial learning agents. In particular, they argue that concepts are not to be seen as the encapsulation of knowledge in symbols, but as the structure on which the emergence of behavior occurs. Therefore, how we represent should be seen as dynamic and time dependent, that is, representations make sense only when embedded in the interaction process.

Moreover, Olier et al. (2017b) analyze differences between several views on concepts by linking categorization based approaches to the computational views of cognition, while ideas of concepts as flexible, distributed and context-dependent to many aspects of embodied Wilson & Golonka (2013) and grounded cognition (Barsalou, 2008). They describe an approach in which representing implies an act of actively interpreting and adapting to the world.

Barsalou (2008), from the perspectives of grounded cognition, has elaborated on how simulation is fundamental for concept's acquisition and processing, referring to simulation as the re-enactment of sensorimotor modalities. That can be linked to the ideas on predictive coding (Rao & Ballard, 1999), in which top-down information in the cortex carries predictions about lower levels, while feed-forward connections carry residual errors. Those notions are further developed by Friston (2010) with the free energy principle, where it is argued that the primary function of the brain is to minimize free energy or suppress prediction error.

Those ideas have been developed and interpreted in different ways as algorithms. Frequently, implementations aim at systems that update internal beliefs about causes of perceived information from prediction error. Some approaches, particularly given the probabilistic characteristics of the free energy principle, are based on Bayesian methods and generative models, which are argued to account for contextual reactions and causality given temporal relations (Chater et al., 2006). Here we explore some existing techniques and propose a method based on generative models that aims at constructing representations of observed and its dynamics. Particularly we propose a generative model that works simultaneously as an encoder, or its own inverse model, by the use of prediction error to update internal states.

2 PREVIOUS WORK

Different approaches based on predictive coding and prediction error minimization have been presented recently. The ideas by (Tani, 2014) focus on such principles to train recurrent networks. In his work, hierarchical architectures are proposed, where internal states at each level change in different time scales through leaky recursions. Each level in that hierarchy encodes more abstract and stable representations of observed sequences. Based on those ideas, Choi & Tani (2017) propose a similar approach based on convolutional neural networks (CNN). In those works, the input to the system is the prediction error.

Similarly, Lotter et al. (2017) have proposed an architecture based on convolutional LSTMs. In such model, each level of a hierarchy processes the prediction error of lower ones and sends information down to lower levels for predicting future frames. In turn, Canziani & Culurciello (2017) and Ilin et al. (2017) also propose hierarchical architectures where the activations of layers at the same level of a coder and a generator, are used laterally, sending information about the last prediction during inference, and about the inference for the generation. In all of those works, the learning is unsupervised.

Eyjolfsdottir et al. (2017) present a semi-supervised approach, where motion prediction error and classification error are minimized simultaneously during training. It is a model based on a recurrent neural network with LSTMs that simultaneously classifies actions and predicts future motion of observed agents in video data.

None of those methods focus on building variational representations explicitly. Recent works have proposed models that combine recurrent NNs with state space models, building methods under the ideas of Bayesian networks, and particularly of Kalman Filters (Haarnoja et al., 2016; Karl et al., 2017). Some approaches along those lines learn dynamic latent variables that adapt to the environment explicitly from prediction error signals (Olier et al., 2017a). In those works, the ideas of variational Bayes or Variational Auto-Encoders (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) are used to model data in terms of probability distributions. Architectures of that kind have been used to build representations useful for predicting behaviors (Olier et al., 2017c). Nonetheless, in those methods mentioned so far, knowledge about the dynamics observed is embedded implicitly in the recursions of the NNs and thus not differentiated nor represented explicitly.

One of the primary goals of the present work is to define internal states that represent different dynamics, or ways in which the observations change over time. We aim at getting as a result some internal representational states that are more stable over periods where such dynamics are relatively constant and change when a new model is needed.

The representations of the dynamics are to be variational, and the best way to achieve that is through generative models such as the VAE. Those architectures are typically based on an encoder and a generator. Nonetheless, recently Bojanowski et al. (2017) have proposed an architecture that without the need for an encoder, aims at obtaining similar advantages to the generative adversarial networks (GAN) (Goodfellow et al., 2014) by estimating a latent representation of data capable of producing good samples. That yields more compact models and a direct coupling between representational states and data only through the generator.

An advantage of that to the problem at hand is that it is not assumed that the parameters of the distributions are to be defined deterministically by the observation through an encoder, but that can depend on the relation between the predictions and the data itself. In Bojanowski et al. (2017) that is done to define the best distribution to generate the data during training; however, that concept could be extended to the inference process continuing to modify the distributions based on the prediction error.

3 METHOD

The model presented is divided in two hierarchical levels and is thought as a Bayesian network. It is based on a generative model that, in opposition to other methods as VAE or GAN does not assume an encoder, but all the processing is based only on the decoder or generator, similarly to the method by Bojanowski et al. (2017).

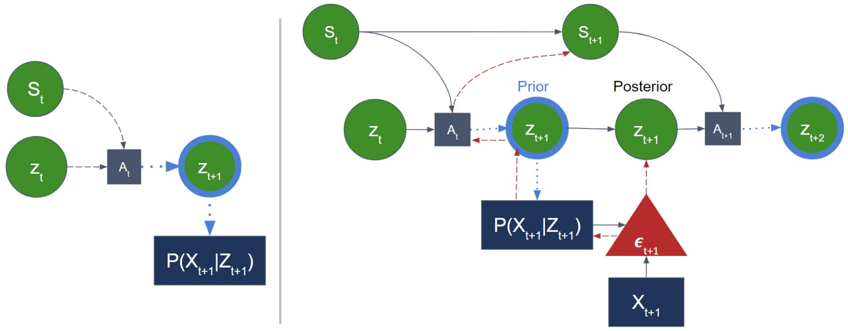

Figure 1: Left, the general diagram of the architecture, where a prior of Z_{t+1} is generated from the current state through the transition model A_t generated from S_t (dynamics level). The predicted observation is denoted as $P(X_{t+1}|Z_{t+1})$ (observation level). Right, the prediction process and the update of Z and S from prediction error.

As can be seen in figure 1 left, in the two levels we consider generative models. We denominate the lower one the ”observation level”, and in it the observations X_t are generated from the latent variables Z_t. In the second level, denominated ”dynamics level”, a transition models A_t are generated from S_t. Such models are understood as the dynamics defining the transition from Z_t to Z_{t+1}.

To develop the ideas we want to introduce, the problem is divided in two parts; a first one where the inputs are constant and the representations have to adapt to the input over several iterations; and a second case where the input data is changing over time and internal states are being adapted to represent the inputs and their dynamics at the same time. In that first case no transition model is considered, thus only the observation level is described; In the second case, the two levels are considered.

3.1 CONSTANT INPUT

As the model is based on predictive coding, and we do not consider a specific recognition model or encoder, the way in which Z_t is updated given an observation depends on the prediction error. In that sense, if X_t was to remain constant, Z_t should be adjusted over time to converge to the best distribution that represents X_t given the knowledge stored in the generator. During both, training and inference, the updating of Z works as backpropagation, but optimizing Z_t and not the weights. This process can be thought as an active exploration of the latent space guided by the uncertainty in the knowledge acquired by the generator until a point in time. Moreover, as the latent space is explored mainly by the dynamic adaptation from prediction error, an inverse transformation is implicitly learned, which allows to recover the latent representations from a generated sample or to estimate them from a real one.

The observations X_t are represented through a set of independent Gaussian distributions $Z_t \sim \mathcal{N}\left(\mu_{Z_t}, diag(\sigma^2_{Z_t})\right)$ as in Kingma & Welling (2014); Rezende et al. (2014). We assume that there is a generative model φ_X that generates $\widehat{X}^i_t|Z_t \sim \mathcal{N}\left(\widehat{\mu}_{X^i_t}, diag(\widehat{\sigma}^2_{X^i_t})\right)$, for each element i in the observations, e.g., each color component of every pixel in an image. Then, the parameters of such distributions are defined as $\widehat{\mu}_{X^i_t}, \widehat{\sigma}_{X^i_t} = \varphi_X(Z_t)$.

Assuming the case where the observations are constant ($X_t = X$), Z_t is to be adjusted to reduce prediction error over time and match the best states to generate X (see figure 2). The error is defined as the mean Log-Likelihood over all the predicted component i in X, denoted as $\log L(X^i_t|\widehat{X}^i_t)$. From now on, we assume that X are images with N pixels and c color components. Thus, each

i represent a color component of a pixel in X, that is, $c * N$ components by image. We define $\epsilon_t = \frac{1}{c*N} \sum_i \log L(X_t^i | \widehat{X}_t^i)$, and the proportional correction to each component j in $\mu_{Z_t^j}$ given the observation as $\delta Z_t^j = \frac{\partial \epsilon_t}{\partial Z_t^j}$. However, to achieve a better stability, we define the correction in terms of the one in the previous time step as a way of momentum:

$$\delta Z_t^j = \frac{1}{2} \left(\delta Z_{t-1}^j + \frac{\partial \epsilon_t}{\partial Z_t^j} \right) \tag{1}$$

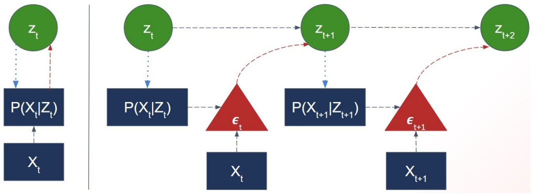

Figure 2: Left, the general description of the update process, where a prediction is generated and the measured error is used to update Z_t. On the right, the same process is unrolled in time as a recurrent network.

Given that such correction is dependent on the weights of the generator at a given time during training, the values could produce variations leading to instabilities, or osculations around the optimal point if used directly. Then, we use a normalized version that is proportional to the error. To normalize the correction we take into account all the components of Z. First we rescale δZ_t as $\Delta Z_t^j = \delta Z_t^j / argmax_j(\delta Z_t^j)$ and we define a normalized version of the variances as $\varsigma^j Z_t = \sigma_{Z_t}^j / argmax_j(\sigma_{Z_t}^j)$

From these normalizations one can defined the correction $\mu_{Z_{t+1}^j} = \mu_{Z_t^j} - \Delta Z_t^j \varsigma^j Z_t$. That however, would not be proportional to the prediction error. To that end, a normalized measure of the error is used. In this case we use the normalized likelihood of the predicted image $N_{Lt} = \frac{1}{cN} \sum_i \exp \left[\frac{-(\mu_{X_t^i} - X_t^i)^2}{(2(\sigma_{X_t^i})^2)} \right]$, which resides in the interval $[0, 1]$.

Defining $\overline{N}_{Lt}$ is defined as the mean of N_{Lt} over all i in X, and $L_{Wt} = 1 - \overline{N}_{Lt}$, the correction of the expected values of Z becomes:

$$\mu_{Z_{t+1}^j} = \mu_{Z_t^j} - L_{Wt} \Delta Z_t^j \varsigma^j Z_t \tag{2}$$

which means that the value of the correction diminishes as the prediction error gets lower.

In order to update the values of σ_Z, we assume that it is to increase if the correction goes beyond the current variance, and get reduced otherwise. That is achieved by:

$$\sigma_{Z_{t+1}}^j = \sigma_{Z_t}^j \left(0.5 + L_{Wt} \frac{|\Delta^j Z_t|}{\varsigma^j Z_t} \right) \tag{3}$$

Loss The loss is calculated as the mean log-likelihood and a regularization defined as the Kullback-Leibler divergence between Z and a prior $P_\theta(Z)$, which we assume as a normal distribution for this constant input case. Then the loss function to be minimized for T iterations of Z updates is given by:

$$\mathcal{L}_c = \frac{1}{T} \sum_{t=1}^{T} \left[-\epsilon_t + KL \left(P(Z_t) || P_\theta(Z) \right) \right] \tag{4}$$

3.2 Transition model and prediction

When X_t is changing over time, updating Z at time t will direct the distribution towards the best representation of the current frame. Nonetheless, as it is necessary to predict the following frame, a transition model is needed to encode how the states Z change, and implicitly how X changes over time. Then, we include the generative model that given S_t produces the transition model A_t.

S_t in the model in figure 1 right acts as a switching variable that determines the transition model in the observation level. Here, such transition is considered as linear of the kind $Z_{t+1} = A_t Z_t + C$, where A_t is a matrix, and C is noise. C in this case will be modelled by the sampling of Z_t and S_t, which are to encode the variance in the observations and transitions.

Considering a video, series of frames are fed as observations, and then the model should represent the sequences in the training data as a specific semantics that makes sense for the observed interaction. Such segmentation is to be achieved by S regarding how observations change over time in a given situation.

The proposed process is depicted in figure 1 right. It starts with generating $A_t|S_t$, from which a prior $\widehat{Z}_{t+1}$ can be estimated given the lineal model assumed. From such prior, a prediction of the observation in the following time step $\widehat{X}_{t+1}|\widehat{Z}_{t+1}$ can be generated. As before, $\widehat{X}_{t+1}|\widehat{Z}_{t+1} \sim \mathcal{N}\left(\widehat{\mu}_{X_{t+1}}, diag(\widehat{\sigma}^2_{X_{t+1}})\right)$ and $\widehat{\mu}_{X_{t+1}}, \widehat{\sigma}_{X_{t+1}} = \varphi_X(\widehat{Z}_{t+1})$. Equally, $S_t \sim \mathcal{N}\left(\mu_{S_t}, diag(\sigma^2_{S_t})\right)$

The prior distribution $\widehat{Z}_{t+1} \sim \mathcal{N}\left(\widehat{\mu}_{Z_{t+1}}, diag(\widehat{\sigma}^2_{Z_{t+1}})\right)$ is estimated as follows: For the expected value, $\widehat{\mu}_{Z_{t+1}} = A_t \mu_{Z_t}$, where $A_t|S_t = \varphi_A(S_t)$ is a generative model. For the variance, a function to be learned is defined such that $\widehat{\sigma}_{Z_{t+1}} = \varphi_\sigma(\sigma_{Z_t})$.

As for Z, the updates for the expected values of S are calculated in relation to the prediction error as:

$$\delta S^k_t = \frac{1}{2}\left(\delta S^k_{t-1} + \frac{\partial \epsilon_{t+1}}{\partial S^k_t}\right) \tag{5}$$

Then again, the result of such differentiation should be normalized to ensure more stability and convergence. Moreover, as we are trying to model switching variables with S, we would need some kind of non-linearity to make the changes dependent on certain situations, and so S would not vary constantly as Z. In this case, such condition would be a constant increase in prediction error over a given period.

To achieve that, we also define the normalized correction of the components of S as $\Delta S^k_t = \delta S^k_t / argmax_k(\delta S^k_t)$, and $\varsigma^j S_t = \sigma^j_{S_t} / argmax_j(\sigma^j_{S_t})$. The update of the expected values of S as:

$$\mu_{S^k_{t+1}} = \mu_{S^k_t} - \gamma_{S^k_t} \Delta S^k_t \varsigma S^k_t \tag{6}$$

$\gamma_{S^k_t}$ is a weight that determines whether there should be a change, and the amplitude of such change.

$$\gamma_{S^k_t} = L_{Wt} W_{S_t} |\delta S^k_t| \tag{7}$$

W_{S_t} is the way in which the non-linearity is introduced. In this case we use a sigmoid that performs a gating function, allowing a change when needed and keeping $\mu_{S^k_t}$ constant otherwise; it is defined as:

$$W_{S_t} = 1/\left(1 + \exp\left[-\left((1 - p\overline{N}_{Lt}) - (\alpha - 2\sigma_{Lw})\right)\beta\right]\right) \tag{8}$$

In that expression, α and σ_{Lw} are moving averages of the expected value and variance of L_{Wt} weighted over the last 100 iterations. $p\overline{N}_{Lt}$ is a moving average of $(\overline{N}_{Lt})^2 / p\overline{N}_{Lt-1}$ weighted over the last 10 iterations. β in turn is defined as $\ln\left((1/0.75 - 1)/(\sigma_{Lw}/2)\right)$.

These parameters are defined to activate the sigmoid when prediction error starts to accumulate beyond learned normal levels. α and σ_{Lw} are learned statistics of the error, and can be seen as the threshold when combined in the exponential. In other words, it is expected that the sigmoid gets active when the error starts rising beyond the mean plus two standard deviations. β is set so that the sigmoid reaches, in average, 0.75 when the values is half standard deviation over the threshold. Finally, $p\overline{N}_{Lt}$ is the variable that is keeping track of the prediction error over time; so it will rise if there is an increasing or constantly high prediction error, which would activate the sigmoid if higher than the threshold.

That can be interpreted as switching between transition models, changing them when the current one is leading to constantly high prediction errors. In other words, the system identifies when the situation has changed, and estimates in which situation it is, which is represented by S.

The variance of S is updated such that it diminishes when there is no change, meaning that the system is more certain about the dynamic model, and would increase when there is a change. That is achieved by:

$$\sigma_{S^k_{t+1}} = \ell\sigma_{S^k_t} + \left(W_{S_t} L_{Wt} |\Delta S^k_t|\right) \tag{9}$$

ℓ is a positive number in $(0, 1)$; we have set it to 0.9 for the experiments.

Loss As before the loss is composed by the prediction error and a regularization that minimizes the divergence between the prior and the posterior of Z. So for a sequence of T time steps the loss is defined as:

$$\mathcal{L}_t = \frac{1}{T}\sum_{t=1}^{T}\left[-\epsilon_t + KL\left(P(Z_t)||P(\widehat{Z}_t)\right)\right] + KL\left(P(Z_T)||P_\theta(Z)\right) \tag{10}$$

The last term is added so that the transition models are forced to maintain the values of the priors low, as otherwise the matrix would increase the values over time.

3.3 Model implementation details

The generative model of the observation level φ_X is a NN composed of a 500 units fully connected layer, which inputs are a sample of Z_t during training, or the expected value during test. It is followed by four layers of transposed convolutions. The receptive fields of all the convolutions are 3 by 3, with stride 2, and the stack size of the filters are 128, 64, 64 and 32. The generator φ_A is composed by two separated fully connected layer of three layers with input size 2 (the size of S), the first two layers have 500 units, and with output of size 500 (size of Z). The inner product of the output of these two NNs yields the matrix A. The variance of the prior $\widehat{Z}_t$ is estimated by φ_σ which is composed of three fully connected layers, with input and output of size 500, and middle layers of size 1000. The update process is disconnected from the general backpropagation, as the weights are optimized for the generation process. The models are optimized using the AdaMax algorithm Kingma & Ba (2014).

4 Experiments

Two experiments are set to evaluate the situations described in 3, constant input and prediction based on changing transitions models.

4.1 Constant input

For testing the first case, an image is maintained constant as the input to the network in figure 2 during 10 time steps. The aim is to test how the evolution of the internal representations lead to the best reconstruction possible given the knowledge stored in the generator. That is to be measured regarding the quality of the samples and the prediction error. The results are also compared to a VAE based on the same architecture as the network, though with two by two pooling in the encoder instead of stride. We compare to a VAE as it is a generative model in which representations are determined regarding probability distributions, and where the loss function is defined in the same way as we do. As has been explored in other works, sharper images can be achieved with methods based on GANs, and in general using more complex loss functions as Bojanowski et al. (2017) have shown. Nonetheless, those approaches do not follow the criteria of variational representations we are focused on here.

The dataset used is the aligned images of the CelebA Dataset (Liu et al., 2015), consisting of aligned faces. 4000 images are used for training.

In figure 3 the differences between the results achieved with our method, and with a VAE are depicted. In that figure one can see the original image on the left, and the output of 5 of the 10 iterations in the update process of the observation level given the static input. Clearly, the active exploration of the latent space leads to better results particularly in terms of color details both of the persons and the background, as well as in terms of sharper details in the faces. Equally, in cases where the poses are less common, these are emulated with the iteration, but are not achieved by the VAE.

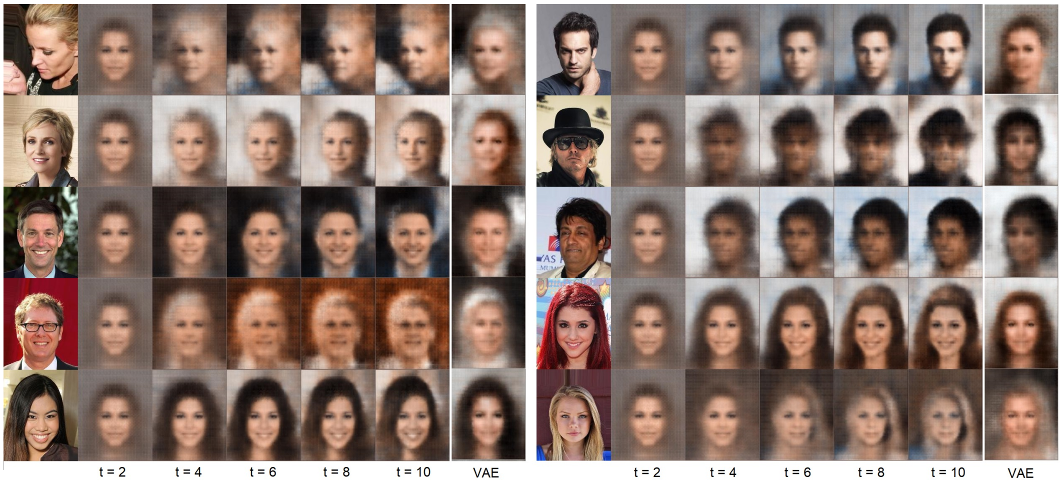

Figure 3: Two sets of examples comparing the results achieved with our method and with the VAE. Five time steps in the iteration process are shown, and the image produced by the VAE compared at the end.

Table 1: Log-likelihood of reconstructions

Method	**Log-likelihood**
Ours	-0.466
VAE	-1.34

4.2 Dynamic input and prediction

To test the complete model, including the dynamics level, video sequences of a person performing different actions are used. The aim is to determine if the changes in S over time correspond to the variations in motion patterns of the person being observed. We use a dataset where a person sitting at a table passes an object to other persons, or place it on defined points on the table. The dataset has been introduced by Schydlo et al.. The scenario is a table with four marked spots on the table, and four participants are sat around it. Only one participant is fully visible, and is the one performing the actions. That person performs six different actions that consist os sequences of sub-action as handing a ball to someone else, placing the ball on a marked spot, or resting the hand on the table. Sequences of those sub actions are classified, and the initial time of each actions have been labelled in the dataset. The six actions are performed randomly in sequence. A whole video is divided in three, one of the sequences is used for training, and the others for validation and testing.

We have trained the method proposed with these video sequences, and the results show a high correlations between the changes in the values of S, and the initial times marked for each action in the ground truth.

To visualize the results, we aligned the frames of the video to the values of μ_{S_t} over time. That can be seen in figure 4. The initial action is handing the ball to someone on the right (from $t = 0$ to $t = 20$), followed by the action of receiving the ball back (from $t = 20$ to $t = 55$). Later, the ball is placed on the table to then be handed to the person on the bottom of the frame (from $t = 125$ to $t = 155$), who later returns it (from $t = 165$ to $t = 200$). Finally, the ball is placed on the table again.

This shows how the model is successfully segmenting the actions in the video sequence in a way that can be even semantically useful for a human observer.

In figure 5 the values of μ_{S_t} over time are plotted together with the action labels of the ground truth. This shows the relation between the changes in S and the predefined actions. Though they do not coincide exactly, most of the changes in S happen in synchrony with the ground truth. That implies

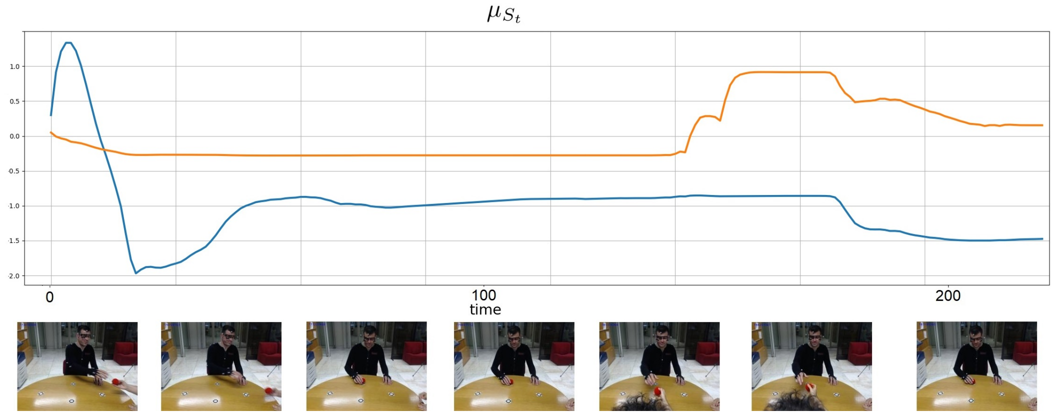

Figure 4: The values of μ_{S_t} over time and alignd frames depicting the corresponding actions.

that the model is extracting the regularity of the predefined actions and learning transition models for each of them. However, there are actions that are similar among them, and thus are represented with very similar values in S.

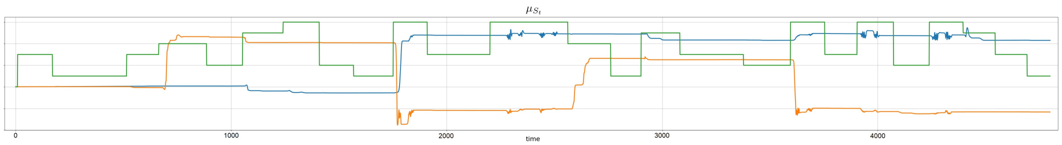

Figure 5: The values of μ_{S_t} over time and the actions given as the ground truth rescaled to fit in the same graph.

5 Conclusions

We have presented a method for representing dynamic data, and we have tested it on videos of interactions. The states are organized in two levels representing the observations and their dynamics respectively. It has been shown that the method proposed is capable of learning generative models by exploring the latent space through an active adaptation based on prediction error propagation. In the model, the representations make sense in the temporal domain, as to serve their function they have to evolve with the observed data dynamically.

Two experiments have been performed to test the model. The first one on static data showing how the adaptation leads to better results than an entirely static model, e.g., a VAE. The process evaluated in that case takes more processing that the static method, yet it shows that the accuracy is not only dependent on the generalization capability of the model, but also on its ability to adapt temporally to the data. In a second experiment, videos of actions performed in a given scenario are used to learn representations of the images and the dynamics of the activities observed. The results show that the model is capable of extracting a semantics similar to the one defined as ground truth for the data used.

These ideas have been connected with definitions positions from different branches of the cognitive and brain sciences, which state that interpreting the world is an active process. That suggests that a possible path towards better machine learning algorithms may imply understanding representations and their processing as a temporal process embedded in a dynamic interaction with the environment, or the evolution of the data itself.

ACKNOWLEDGMENTS

REFERENCES

Lawrence W Barsalou. Grounded cognition. *Annual Review of Psycholgy.*, 59:617–645, 2008.

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*, 2017.

Alfredo Canziani and Eugenio Culurciello. Cortexnet: a generic network family for robust visual temporal representations. *arXiv preprint arXiv:1706.02735*, 2017.

Nick Chater, Joshua B Tenenbaum, and Alan Yuille. Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences*, 10(7):287–291, 2006.

M. Choi and J. Tani. Predictive coding for dynamic vision: Development of functional hierarchy in a multiple spatio-temporal scales rnn model. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 657–664, May 2017. doi: 10.1109/IJCNN.2017.7965915.

Eyrun Eyjolfsdottir, Kristin Branson, Yisong Yue, and Pietro Perona. Learning recurrent representations for hierarchical behavior modeling. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Tuomas Haarnoja, Anurag Ajay, Sergey Levine, and Pieter Abbeel. Backprop kf: Learning discriminative deterministic state estimators. In *Advances in Neural Information Processing Systems*, pp. 4376–4384, 2016.

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

Alexander Ilin, Isabeau Prémont-Schwarz, Tele Hotloo Hao, Antti Rasmus, Rinu Boney, and Harri Valpola. Recurrent ladder networks. *arXiv preprint arXiv:1707.09219*, 2017.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. 2017.

Markus Kiefer and Friedemann Pulvermüller. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*, 48(7):805–825, 2012.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Juan Sebastian Olier, Emilia Barakova, Matthias Rauterberg, Lucio Marcenaro, and Carlo Regazzoni. Grounded representations through deep variational inference and dynamic programming. *Proceedings of the 7th ICDL-EPIROB conference*, 2017a.

Juan Sebastian Olier, Emilia Barakova, Carlo Regazzoni, and Matthias Rauterberg. Re-framing the characteristics of concepts and their relation to learning and cognition in artificial agents. *Cognitive Systems Research*, 44:50 – 68, 2017b. ISSN 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2017.03.005. URL `http://www.sciencedirect.com/science/article/pii/S1389041717300402`.

Juan Sebastian Olier, Damian Andres Campo, Lucio Marcenaro, Emilia Barakova, Carlo Regazzoni, and Matthias Rauterberg. Active estimation of motivational spots for modeling dynamic interactions. In *The 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2017)*. IEEE, 2017c.

Juan Sebastian Olier, Pablo Marín-Plaza, David Martín, Lucio Marcenaro, Emilia Barakova, Matthias Rauterberg, and Carlo Regazzoni. Dynamic representations for autonomous driving. In *The 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2017)*. IEEE, 2017d.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

Danilo J Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1278–1286, 2014.

Paul Schydlo, Mirko Rakovi, and Jos Santos-Victor. Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction. In *Unpublished*. URL `http://vislab.isr.ist.utl.pt/datasets/`.

Jun Tani. Self-organization and compositionality in cognitive brains: A neurorobotics study. *Proceedings of the IEEE*, 102(4):586–605, 2014.

Andrew Wilson and Sabrina Golonka. Embodied cognition is not what you think it is. *Frontiers in Psychology*, 4:58, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00058. URL `http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00058`.