UNDERSTANDING VISUAL CONCEPTS WITH CONTINUATION LEARNING

William F. Whitney, Michael Chang, Tejas Kulkarni, and Joshua B. Tenenbaum
Department of Brain and Cognitive Science
Massachusetts Institute of Technology
{wwhitney,mbchang,tejask,jbt}@mit.edu

ABSTRACT

We introduce a neural network architecture and a learning algorithm to produce factorized symbolic representations. We propose to learn these concepts by observing consecutive frames, letting all the components of the hidden representation except a small discrete set (gating units) be predicted from the previous frame, and let the factors of variation in the next frame be represented entirely by these discrete gated units (corresponding to symbolic representations). We demonstrate the efficacy of our approach on datasets of faces undergoing 3D transformations and Atari 2600 games.

INTRODUCTION

Deep learning has led to remarkable breakthroughs in solving perceptual tasks such as object recognition, localization and segmentation using large amounts of labeled data. However, the problem of learning abstract representations of images without manual supervision is an open problem in machine perception. Existing unsupervised learning techniques have tried to address this problem (Hinton and Salakhutdinov 2006; Ranzato et al. 2007; Lee et al. 2009) but lack the ability to produce latent factors of variations or symbolic visual concepts from raw data. Computer vision has historically been formulated as the problem of producing symbolic descriptions of scenes from input images (Horn 1986). Without disentangled and symbolic visual concepts, it is difficult to interpret or re-use representations across tasks as no single component of the representation vector has a semantic meaning by itself. Traditionally, it has been difficult to adapt neural network architectures to learn such representations from raw data. In this paper, we introduce a neural network architecture and a learning algorithm to produce factorized symbolic representations given consecutive images. We demonstrate the efficacy of our approach on datasets including faces undergoing 3D transformations, moving objects in 2D worlds, and Atari 2600 games.

RELATED WORK

A number of generative models have been proposed in the literature to learn abstract visual representations including RBM-based models (Hinton and Salakhutdinov 2006, Lee et al. (2009)), variational auto-encoders (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014; Kulkarni et al. 2015), convolution based encoder-decoders (Ranzato et al. 2007; Lee et al. 2009), and generative adversarial networks (Goodfellow et al. 2014; Radford, Metz, and Chintala 2015). However, the representations produced by most of these techniques are entangled, without any notion of symbolic concepts. The exception to this is more recent work by Hinton et al. (Hinton, Krizhevsky, and Wang 2011) on 'transforming auto-encoders' which use a domain-specific decoder with explicit visual entities to reconstruct input images. Inverse graphics networks (Kulkarni et al. 2015) have also been shown to disentangle interpretable factors of variations, albeit in a semi-supervised learning setting. Probabilistic program induction has been recently applied for learning visual concepts in the hand-written characters domain (Lake, Salakhutdinov, and Tenenbaum 2015). However, this approach requires the specification of primitives to build up the final conceptual representations. The tasks we consider in this paper contain great conceptual diversity and it is unclear if there exist a simple set of primitives. Instead we propose to learn these concepts by observing consecutive frames, letting all the components of the hidden representation except a small discrete set (gating units) be

predicted from the previous frame, and let the factors of variation in the next frame be represented entirely by these discrete gated units (corresponding to symbolic representations).

MODEL

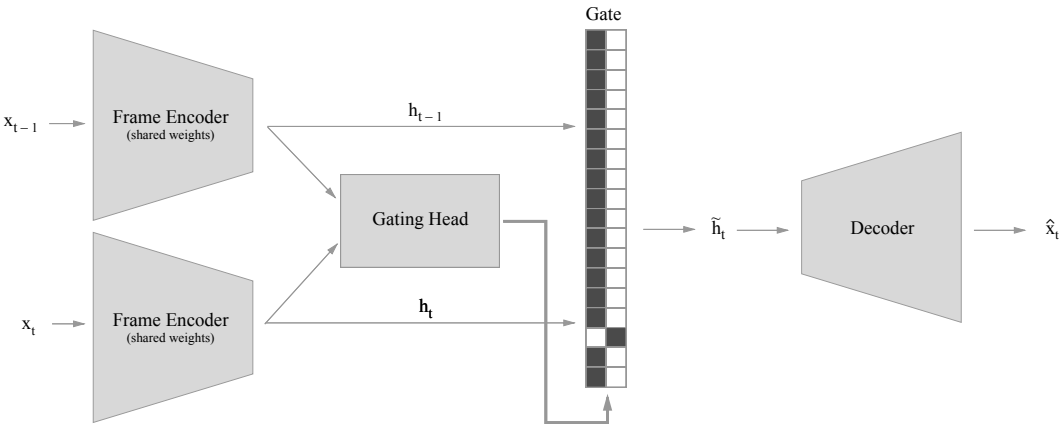

Figure 1: The gated model. Each frame encoder produces a representation from its input. The gating head examines both these representations, then picks one component from the encoding of time t to pass through the gate. All other components of the hidden representation are from the encoding of time $t-1$. As a result, each frame encoder predicts what it can about the next frame and encodes the "unpredictable" parts of the frame into one component.

This model is a deep convolutional autoencoder (Hinton and Salakhutdinov 2006; Bengio 2009; Masci et al. 2011) with modifications to accommodate multiple frames and encourage a particular factorization in the latent space. Given two frames in sequence, x_{t-1} and x_t, the model first produces respective latent representations h_{t-1} and h_t through a shared encoder. The model then combines these two representations to produce a hidden representation $\tilde{h}_t$ that is fed as input to a decoder.

We train the model using a novel objective function: given the previous frame x_{t-1} of a video and the current frame x_t, reconstruct the current frame as $\hat{x}_t$.

To produce $\tilde{h}_t$, we introduce a *gating* in the encoder (see Fig. 1) that select a small set of *gating units* that characterize the transformation between x_{t-1} and x_t. For clarity, in this paper we describe our model under the context of one gating unit. Concretely, the encoder learns to use a *gating head* that selects one index i of the latent representation vector as the gating unit, and then $\tilde{h}_t$ is constructed as h_{t-1}, with the ith component of h_{t-1} swapped out for the ith component of h_t.

Because the model must learn to reconstruct the current frame t from a representation that is primarily composed of the components of the representation of x_{t-1}, the model is encouraged to represent the attributes of t that are different from that of x_{t-1}, such as the lighting or pose of a face, in a very compact form that is completely disentangled from the invariant parts of the scene, such as the facial features. Thus, the model isolates the transformation from x_{t-1} to x_t from other latent features via the component i selected by the gating head.

CONTINUATION LEARNING

To learn the gating function, we use a technique first described in (W. Whitney 2016) for smoothly annealing a soft weighting function into a binary decision. Ordinarily, a model which produces a hard decision to gate through a single component (out of e.g. 200) would be difficult to train; in this case, it would require many forward passes through the decoder to calculate the expectation of the loss for each of the possible decisions. However, a model which uses a soft weighting over all the components can be trained with gradient descent in a single forward-backward pass.

In order to create a continuation between these two possibilities, we use a scheduling for *weight sharpening* (Graves, Wayne, and Danihelka 2014) combined with noise on the output of the gating

head. Given a weight distribution w produced by the gating head and a sharpening parameter γ which is proportional to the training epoch, we produce a sharpened and noised weighting:

$$w_i' = \frac{\left(w_i + \mathcal{N}(0, \sigma^2)\right)^\gamma}{\sum_j w_j^\gamma}$$

This formulation forces the gating head to gradually concentrate more and more mass on a single location at a time over the course of training, and in practice results in fully binary gating distributions by the end of training. This gating distribution thus selects a single component of h_t to use in $\tilde{h}_t$.

RESULTS

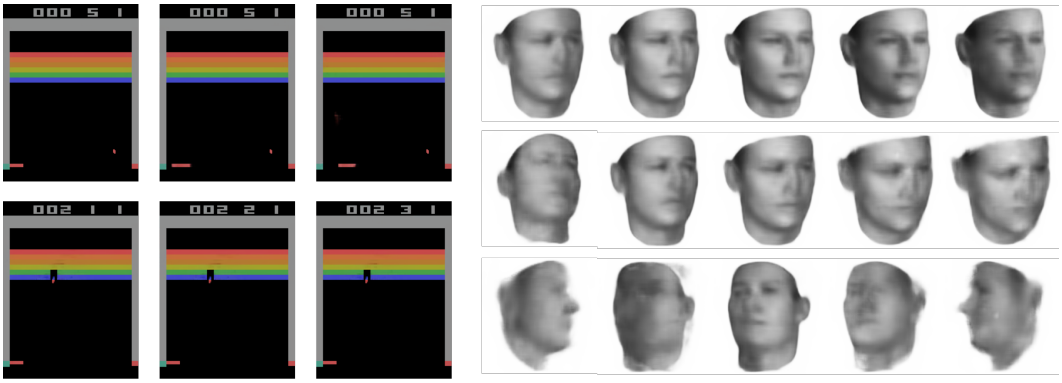

Figure 2: **Manipulating the hidden representation.** Each row was generated by encoding an input image, then changing the value of a single component of the latent representation before rendering it with the decoder. **Top left:** a single unit controls the position of the paddle in Breakout. **Bottom left:** another unit controls the count of the remaining lives in the score bar. **Top right:** one unit controls the direction of lighting. **Middle right:** a unit that controls the elevation of the face. **Bottom right:** a unit controls the azimuth of the face, though this transformation is not smooth. All input images are from the test set.

ATARI FRAMES

Our first dataset is frames from playing the Atari 2600 game Breakout. The model is given as input two frames which occurred in sequence, then reconstructs the second frame. This dataset was generated with a trained DQN network (Mnih et al. 2015). Since the model can only use a few components of its representation from the second frame, these components must contain all information necessary to predict the second frame given the first. For this dataset we use three gating heads, allowing three components of h_t to be included in $\tilde{h}_t$.

SYNTHETIC FACES

We trained the model on faces generated from the Basel face model (Paysan et al. 2009) and prepared as in (Kulkarni et al. 2015). The input is two images of the same face between which only one of {lighting, elevation, azimuth} changes. For this dataset we use a single gating head, so the model must represent all differences between these two images in one unit only.

DISCUSSION

We have shown that it is possible to train a model which learns a factorized, symbolic representation of the factors of variation in image sequences from raw data. Such a model uses temporal continuity to understand visual concepts at a high level, representing objects and motion instead of raw pixels. Future work can extend this model to more complex settings with an arbitrary number of factors of variation.

REFERENCES

Bengio, Yoshua. 2009. "Learning Deep Architectures for AI." *Foundations and Trends in Machine Learning* 2 (1). Now Publishers Inc.: 1–127.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems*, 2672–80.

Graves, Alex, Greg Wayne, and Ivo Danihelka. 2014. "Neural Turing Machines." *ArXiv Preprint ArXiv:1410.5401*.

Hinton, Geoffrey E, and Ruslan R Salakhutdinov. 2006. "Reducing the Dimensionality of Data with Neural Networks." *Science* 313 (5786). American Association for the Advancement of Science: 504–7.

Hinton, Geoffrey E, Alex Krizhevsky, and Sida D Wang. 2011. "Transforming Auto-Encoders." In *Artificial Neural Networks and Machine Learning–ICANN 2011*, 44–51. Springer.

Horn, Berthold. 1986. *Robot Vision*. MIT press.

Kingma, Diederik P, and Max Welling. 2013. "Auto-Encoding Variational Bayes." *ArXiv Preprint ArXiv:1312.6114*.

Kulkarni, Tejas D, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. 2015. "Deep Convolutional Inverse Graphics Network." In *Advances in Neural Information Processing Systems*, 2530–8.

Lake, Brenden M, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. "Human-Level Concept Learning Through Probabilistic Program Induction." *Science* 350 (6266). American Association for the Advancement of Science: 1332–8.

Lee, Honglak, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. 2009. "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 609–16. ACM.

Masci, Jonathan, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. 2011. "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction." In *Artificial Neural Networks and Machine Learning–ICANN 2011*, 52–59. Springer.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, et al. 2015. "Human-Level Control Through Deep Reinforcement Learning." *Nature* 518 (7540). Nature Publishing Group: 529–33.

Paysan, P., R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. 2009. "A 3D Face Model for Pose and Illumination Invariant Face Recognition." *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*. Genova, Italy: IEEE.

Radford, Alec, Luke Metz, and Soumith Chintala. 2015. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks." *ArXiv Preprint ArXiv:1511.06434*.

Ranzato, M, Fu Jie Huang, Y-L Boureau, and Yann LeCun. 2007. "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition." In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 1–8. IEEE.

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra. 2014. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models." *ArXiv Preprint ArXiv:1401.4082*.

Whitney, William. 2016. "Disentangled Representations in Neural Models." *ArXiv Preprint ArXiv:1602.02383*.