Unsupervised Learning of Physical Concepts
ECE 6970, Project Report

Niko Grupen, George Karagiannis

December 2019

1 Introduction

From a very early age (∼2-6 months), humans possess a remarkable ability to reason about physical events. Intuitive physics—the name commonly used to describe this phenomenon—has been identified as a core ingredient of human intelligence and one that bootstraps learning throughout development [38]. For example, seminal studies in the developmental psychology literature have shown that a foundational understanding of basic object concepts (e.g. object individuation [39, 21], object permanence [5], spatiotemporal continuity [35]) guides later learning of more complex relationships between objects, such as collision, containment, and support [3, 4, 18]. Equally remarkable is the fact that infants learn this object-centric representation of the world without any direct supervision. Though infants may learn semantic information in a supervised fashion later in development (e.g. a parent pointing to and naming objects in a scene), knowledge of intuitive physics is largely acquired by simply observing sequences of events.

An AI that exhibits "human-like" learning and reasoning [25] therefore must not only possess the ability to explain observations in terms of physical concepts, but must be able to learn such concepts in a completely unsupervised (or self-supervised) fashion. Recent work has examined this problem primarily by modelling individual aspects of intuitive physics that suit specific applications [1, 26] or by drawing parallels between the approximate probabilistic reasoning capabilities of infants and the physics engines found in modern day simulators [8, 6]. Though both classes of prior work do not require supervision, we instead desire a complete representation of an environment's physical "state" (i.e. the physical properties present in the scene) that does not assume *a priori* knowledge of how the world works (e.g. physics engines). This goal is effectively an application of the representation learning problem [10], which seeks to learn a distribution over latent factors of an observed input. Specifically, because our data consists of 2D videos of objects interacting (in a rudimentary attempt to capture the passive observation an infant experiences in the first few months of development), we are interested in the next-step prediction variant of representation learning.

In this report, we examine one such approach, Contrastive Predictive Coding (CPC)[31], which modifies the next-step prediction task by creating an objective that instead maximizes the mutual information between future observations and the observations we have made thus far. By constructing a loss function that meets this objective, the authors show that learning representations in this manner is indeed tractable. Working within this framework, the question of interest becomes: given video scenes of objects interacting, do the latent representations learned by CPC encode physical concepts that will be useful for other tasks? We provide preliminary empirical results that serve as a first step towards answering this question.

2 Related Work

2.1 Computational Accounts of Intuitive Physics

The origin of an infant's knowledge of the physical world is a widely debated subject in cognitive science, with theories of cognitive development spanning nearly the entire spectrum between Piaget's theory of experiential stage transformation [32, 33] and Spelke's theory of innate core knowledge [38]. It is no surprise then that there is a wide range of computational approaches to modeling intuitive physics as well. We focus here on the three most popular categories of approach. The first posits that understanding intuitive physics is equivalent to defining a scene in a simulator and allowing it to play out over a series of time steps [7]. Such approaches, often referred to as "intuitive physics engines", specify a set of objects and possible interactions [6, 7, 8], modelling successive roll-outs of this "simulation" as a stochastic process; later performing inference to predict future states. Recent work replaces some of the hand-defined components of this paradigm with learned object properties and relations [13, 43]. A second popular approach, probabilistic program induction, similarly adds structure to the problem by specifying an abstract description from which valid "programs" can be defined to describe objects, their parts, and relations between objects [23, 24]. Though primarily intended to study compositionality in concept learning, these approaches can certainly apply to physical concepts as well. Finally, a number of approaches have taken a bottom-up, end-to-end learning approach that models individual aspects of intuitive physics directly from observations [1, 26, 41]. Our approach seeks to remove pre-defined "rules" from the modelling process while simultaneously capturing information about many general physical concepts in a single representation.

2.2 Unsupervised Representation Learning

In the representation learning community, auxiliary tasks based on next-step prediction have provided effective self-supervision for learning generative models of a given data distribution. Unlike standard probabilistic generative models that specify a conditional probability distribution $P(X|Y{=}y)$ and a prior $P(Y)$ for a set of input samples $X{=}\{x_1, ..., x_i\}$ and labels $Y{=}\{y_1, ..., y_j\}$, self-supervised models do not have access to a corresponding label for each data sample. Instead, such methods assume X to be a temporal sequence and model the conditional probability $P(x_t|x_1, ..., x_{t-1})$, where t represents the current time-step and $[x_{t-1}, ..., x_1]$ represents the sequence of observations preceding x_t. The joint probability of the entire sequence can then be recovered trivially as the product of the aforementioned conditional probabilities $P(x_t, ..., x_1) = \prod_{n=1}^{t} P(x_n|x_n, ..., x_1)$. Though typically used to model text for natural language processing applications [11, 28, 34, 22], next-step prediction methods have proven effective for frame prediction in videos [15, 27, 40] and even the prediction of action-conditioned states in Atari games [30, 16]. CPC combines next-step prediction with a contrastive loss function, which has proven useful in a number of representation learning techniques [37, 36, 42, 14]. CPC also falls into a recent class of methods that learn features by maximizing the mutual information between random variables [2, 19, 16], most of which stem from MINE [9].

3 Method

We first review a few concepts that are key to understanding CPC and its formulation.

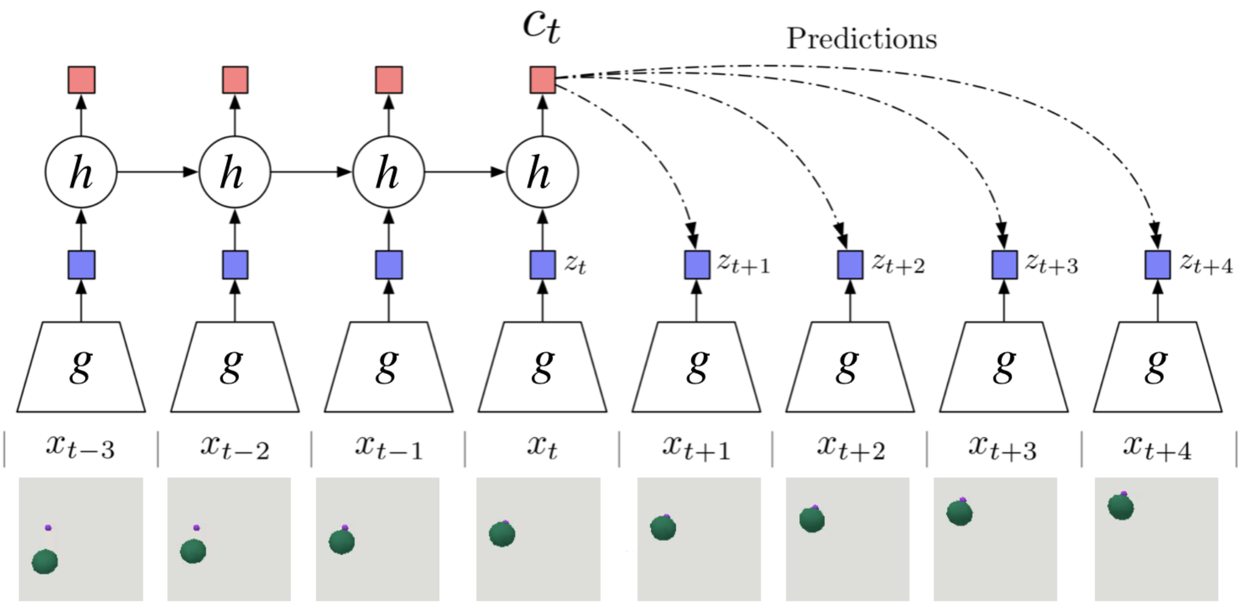

Figure 1: An intuitive view of Contrastive Predictive Coding.

3.1 Mutual Information

Let $X{\sim}P_X$ and $Y{\sim}P_Y$ be two random variables with joint P_{XY} and PMFs p_X and p_Y, respectively. We define mutual information as the Kullback-Leibler (KL-) divergence between the joint P_{XY} and the product of the marginals $P_X \otimes P_Y$ as follows:

$$I(X;Y) := D_{KL}(P_{XY}||P_X \otimes P_Y) \tag{1}$$

$$= \mathop{\mathbb{E}}_{P_{XY}}\left[\log\frac{dP_{XY}}{dP_X \otimes dP_Y}\right] \tag{2}$$

If we assume both P_X and P_Y are discrete distributions, we can rewrite Equation 2 as a sum:

$$I(X;Y) = \sum_{x,y \in P_{XY}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \tag{3}$$

$$= \sum_{x,y \in P_{XY}} P_{XY}(x,y)\log\frac{P_X(x|y)}{P_X(x)} \tag{4}$$

where the conditional form follows nicely in Equation 4. Mutual information quantifies the dependence between X and Y where, in Equation 1, a larger divergence between P_{XY} and $P_X \otimes P_Y$ represents a stronger dependence. Importantly, we can also represent mutual information as the decrease in entropy of X when introducing Y:

$$I(X;Y) := H(X) - H(X|Y) \tag{5}$$

Spaces	Functions	Other Notation
$\mathcal{X}$: sample space	$g : \mathcal{X} \to \mathcal{Z}$	t: current time-step
$\mathcal{Z}$: space of sample encodings	$h : \mathcal{Z} \to \mathcal{C}$	k: number of prediction steps
$\mathcal{C}$: space of context vectors	$w : \mathcal{C} \to \mathcal{Z}$	

Table 1: Table of notation for CPC.

3.2 Log Bilinear Model

Given a set of sequential samples $V = \{w_1, ..., w_N\}$, a log bilinear model computes a distribution over possible future observations, given all of the observations made up to this point:

$$P(w_i = w | w_1, ..., w_{i-1}) = \frac{\exp(g(w)^T c_i)}{\sum_{w' \in V} \exp(g(w')^T c_i)} \tag{6}$$

where $g(w_i)$ is a vector representation of the observation w_i and c_i is a function summarizing observations up to the current time-step, typically computed as:

$$c_i = \sum_{j=1}^{i-1} \alpha_j g(w_j) \tag{7}$$

Here, c represents a linear combination of $[x_{i-1}, ..., x_1]$, but more complex functions can be used, as we will see in the case of CPC. Also note that a different $\alpha_i \in A$ is used for each time-step. We use the notation V and w here to mirror the literature on log bilinear models, in which samples w_i represent words in a vocabulary V, but the same model holds for any time-series data.

3.3 Contrastive Predictive Coding

Let $\mathcal{X} = \{X_1, ..., X_N\}$ be a set of temporally ordered samples. For a single sample $X_i \in \mathcal{X}$, we define the following functions:

$$z_i = g(X_i) \tag{8}$$

$$C_i = h(Z_{\leq i}) = \begin{cases} h(Z_i, C_{i-1}) & \text{if } i > 1. \\ h(Z_i, \mathbf{0}) & \text{otherwise.} \end{cases} \tag{9}$$

where $g(X_i)$ is a non-linear transformation of X_i into a (typically lower-dimensional) vector and $h(g(X_i), C_{i-1})$ is a context vector summarizing any previously encountered samples $[X_{i-1}, ..., X_1]$, as in Equation 7.

The authors posit that it is possible to learn a rich representation of an input sample from the current time-step $X_t \in \mathcal{X}$ by constructing a learning objective that maximizes $I(X_{t+k}; C_t)$; the mutual information between an input sample from a future time-step X_{t+k} and the context at the current time-step C_t. What does such a learning objective[1] look like? We know from Equation 5 that:

$$I(X_{t+k}; C_t) = H(X_{t+k}) - H(X_{t+k} | C_T)$$

[1] This motivation and the descriptions that follow represent our interpretation of CPC. The story and notation differs from that of the original paper, but we feel that expressing it in this manner is crucial to understanding the underlying ideas. This interpretation is a work in progress.

and, therefore, that the mutual information is upper bounded as:

$$I(X_{t+k}; C_t) \leq H(X_{t+k}) \leq \log N$$

where $N = |\mathcal{X}|$. In the absolute best case, then, $I(X_{t+k}; C_t) - \log N = 0$. This provides intuition that our loss function should be of the following form:

$$\mathcal{L}_N = I(X_{t+k}; C_t) - \log N \tag{10}$$

$$= \mathop{\mathbb{E}}_{\mathcal{X}} \left[\log \frac{P(X_{t+k}|C_t)}{P(X_{t+k})} \right] - \log N \tag{11}$$

at which point mutual information is maximized. Of course we know that in general mutual information will not be maximal, $\mathcal{L}_N = I(X_{t+k}; C_t) - \log N \geq 0$, and L_N will be a quantity we want to minimize. The question now becomes: is there a way to easily approximate the RHS of Equation 10? There are multiple ways that this problem can be solved, with one such example being MINE [9]. The authors instead define a new loss, InfoNCE, inspired by prior work on log bilinear models for time-series data (outlined in 3.2).

The authors define InfoNCE loss as:

$$\mathcal{L}_{I_{\text{NCE}}} = -\mathop{\mathbb{E}}_{\mathcal{X}} \left[\log \frac{f_k(X_{t+k}, C_t)}{\sum_{X_j \in \mathcal{X}} f_k(X_j, C_t)} \right] \tag{12}$$

which becomes precisely a log bilinear model if we define f_k as follows:

$$f_k(X_{t+k}, C_t) = \exp(Z_{t+k}^T W_k C_t) \tag{13}$$

The authors claim[2] that the optimal value of $\mathcal{L}_{I_{\text{NCE}}}$ is proportional to $\frac{p(X_{t+k}|C_t)}{P(X_{t+k})}$. If we are clever about how we expand Equation 11, we find that this loss formulation does indeed approximate the mutual information objective we are interested in:

$$\mathcal{L}_N = \mathop{\mathbb{E}}_{\mathcal{X}} \left[\log \frac{P(X_{t+k}|C_t)}{P(X_{t+k})} \right] - \log N \tag{14}$$

$$= \mathop{\mathbb{E}}_{\mathcal{X}} \left[\log \frac{\frac{P(X_{t+k}|C_t)}{P(X_{t+k})}}{N} \right] \tag{15}$$

$$= \mathop{\mathbb{E}}_{\mathcal{X}} \left[\log \frac{\frac{P(X_{t+k}|C_t)}{P(X_{t+k})}}{N \mathop{\mathbb{E}}_{X_j} \frac{P(X_j|C_t)}{P(X_j)}} \right] \tag{16}$$

$$= \mathop{\mathbb{E}}_{\mathcal{X}} \left[\log \frac{\frac{P(X_{t+k}|C_t)}{P(X_{t+k})}}{\sum_{X_j \in \mathcal{X}} \frac{P(X_j|C_t)}{P(X_j)}} \right] \tag{17}$$

Disclaimer In practice, solving Equation 12 can become computationally prohibitive for large $N = |\mathcal{X}|$, as it requires computing the inner product between many Z's and C's. The authors use Noise Contrastive Estimation [17, 29, 20] and Importance Sampling [12]—methods that instead

[2] We have note verified this claim in its entirety yet.

approximate the expectation with an easier-to-sample-from set $\mathcal{S} = \{s_1, ..., s_N\}$ containing one "positive" sample from $P(X_{t+k}|C_t)$ and $N-1$ "negative" samples from $P(X_{t+k})$—to address this limitation. We do not cover this step in the report [3].

4 Discussion

We have presented an introduction to Contrastive Predictive Coding and its application to the unsupervised learning of intuitive physics. Our re-derivation of this paper is certainly a work in progress, but we are excited by the level of understanding we have gained about this paper, compared to just a few days ago. We feel that studying this approach in the manner that we have brings us closer to our goal of building more principled machine learning algorithms that begin to scratch the surface of systems that can learn and reason like humans.

We have a number of future steps planned, which begin with finishing our re-derivation of the paper. We hope that this will inspire a better implementation of the algorithm and ideas about how to improve it moving forward. We are also interested in training CPC on a more complex datasets that includes additional object relationships like containment and stability.

References

[1] Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. In *Advances in Neural Information Processing Systems*, pages 5074–5082, 2016.

[2] Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. *arXiv preprint arXiv:1906.08226*, 2019.

[3] Renée Baillargeon. Infants' physical world. *Current directions in psychological science*, 13(3):89–94, 2004.

[4] Renée Baillargeon, Jie Li, Weiting Ng, and Sylvia Yuan. An account of infants' physical reasoning. *Learning and the infant mind*, pages 66–116, 2009.

[5] Renee L Baillargeon, Elizabeth S. Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20:191–208, 1985.

[6] Christopher Bates, Peter Battaglia, Ilker Yildirim, and Joshua B Tenenbaum. Humans predict liquid dynamics using probabilistic simulation. In *CogSci*, 2015.

[7] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016.

[8] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

[3] We are in the process of re-deriving these steps and did not want to include information that we were not completely comfortable with yet

[9] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[10] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[11] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[12] Yoshua Bengio and Jean-Sébastien Senécal. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, 19(4):713–722, 2008.

[13] Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016.

[14] Sumit Chopra, Raia Hadsell, Yann LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546, 2005.

[15] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in neural information processing systems*, pages 64–72, 2016.

[16] Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo A Pires, Toby Pohlen, and Rémi Munos. Neural predictive belief representations. *arXiv preprint arXiv:1811.06407*, 2018.

[17] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[18] Susan J Hespos and Renée Baillargeon. Young infants' actions reveal their developing knowledge of support variables: Converging evidence for violation-of-expectation findings. *Cognition*, 107(1):304–316, 2008.

[19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[20] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[21] Philip J. Kellman and Elizabeth S. Spelke. Perception of partly occluded objects in infancy. *Cognitive Psychology*, 15(4):483 – 524, 1983.

[22] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[23] Brenden Lake, Ruslan Salakhutdinov, and Joshua Tenenbaum. Concept learning as motor program induction: A large-scale empirical study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, 2012.

[24] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[25] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[26] Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*, 2016.

[27] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[28] Andriy Mnih and Geoffrey Hinton. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648. ACM, 2007.

[29] Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*, 2012.

[30] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in neural information processing systems*, pages 2863–2871, 2015.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[32] Jean Piaget. *The construction of reality in the child*. Routledge, 2013.

[33] Jean Piaget and Margaret Cook. *The origins of intelligence in children*, volume 8. International Universities Press New York, 1952.

[34] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

[35] Elizabeth S. Spelke, Roberta Kestenbaum, Daniel Simons, and Debra Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, 13:113–142, 06 1995.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[37] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[38] Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992.

[39] Elizabeth S Spelke, Claes von Hofsten, and Roberta Kestenbaum. Object perception in infancy: Interaction of spatial and kinetic information for object boundaries. *Developmental Psychology*, 25(2):185, 1989.

[40] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

[41] Zhihua Wang, Stefano Rosa, Bo Yang, Sen Wang, Niki Trigoni, and Andrew Markham. 3d-physnet: Learning the intuitive physics of non-rigid object deformations. *arXiv preprint arXiv:1805.00328*, 2018.

[42] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[43] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems*, pages 127–135, 2015.

A Preliminary Results

We present preliminary results[4] obtained by training CPC on a custom dataset composed of 2D scenes in which objects move and interact in arbitrary ways. As in the paper, we train CPC models and probe the learned representations (both z_t and c_t) with a physics-based linear classification task. In practice, we implement g as a non-linear encoder (CNN) and h as an autoregressive model (GRU).

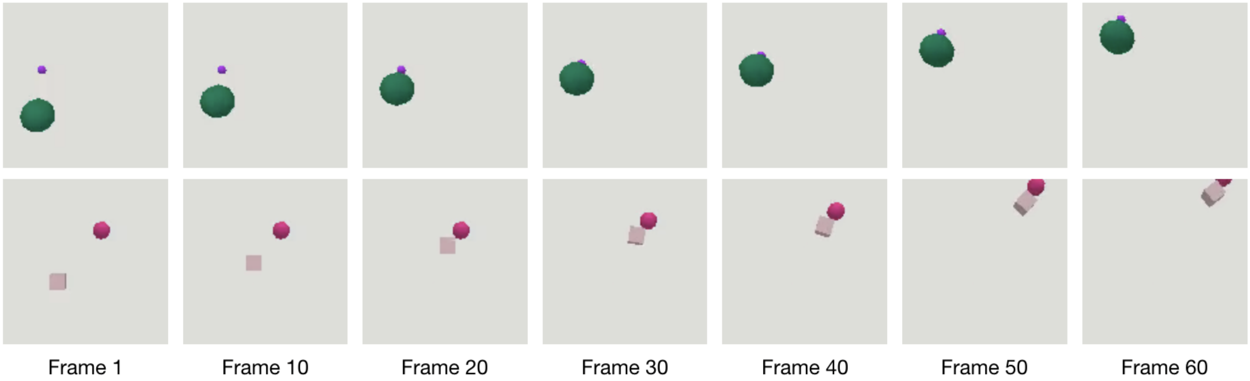

Figure 2: Example videos from dataset.

A.1 Dataset

Our dataset, samples from which are displayed in figure XYZ, consists of 5,500 videos with a length of 150 frames each, yielding a total of 825,000 frames. Each video is initialized with two objects of random shape (square or circle), size, and color (from the RGB spectrum). An external force

[4] Following our discussion earlier this week, we thought it best to place our preliminary results in the appendix, as we expect that our implementation of CPC will change as does our understanding of the method.

Object Properties	Object Relations
Size	Collision
Color	
Shape	
Position	
Velocity	

Table 2: Table of object properties and relations collected in our dataset. Note that object properties are collected for both the static and moving objects.

of random magnitude perturbs one of the objects, yielding a scene with one static object and one moving object. With probability α, the force pushes the moving object on a trajectory that collides with the static object; and the object is pushed on a non-collision trajectory with probability $1-\alpha$. We found that $\alpha = 0.75$ produces a sufficient number of object collisions. This combination of initial conditions and effects yields a set of suitably rich interactions from which our model can learn. We also collect each object's properties in each frame to serve as a set of labels from which we can test the learned CPC representations. The total set of information collected from each frame is summarized in Table 2.

A.2 Evaluation

To evaluate, we compared the effect of both the number of constrastive samples (N) and the number of forward prediction steps (k) on training. We hypothesize that both a larger N and larger k should result in C_t capturing a richer representation that performs better on downstream probing tasks. In each case, the probing task was based on classifying physical properties in the scene, as outlined in Table 2. *Note*: all models were only trained for 10 epochs, regardless of N or k values, due to computational/time constraints.

Effect of Negative Samples (N) Figure 3 shows the results of training CPC on three values of N: $N = 64$, $N = 128$, $N = 256$. All three models appear close to convergence at the end of 10 epochs. Despite minor fluctuations in accuracy, it seems performance on the linear probing task remains relatively constant across different values of N. This would indicate that the number of contrastive samples has little effect on the information encoded in C_T. We note, however, that we were only able to test N values up to a value of 256 before running into memory constraints. We would be interested in exploring whether or not this effect is more pronounced with a larger value of N. We also note that there are a few exceptions here, as it appears that accuracy in predicting object color increases with N. It may also be the case that the scenes in our dataset are too simplistic. We are interested in examining the results of these tests for a more complex (possibly 3D) dataset.

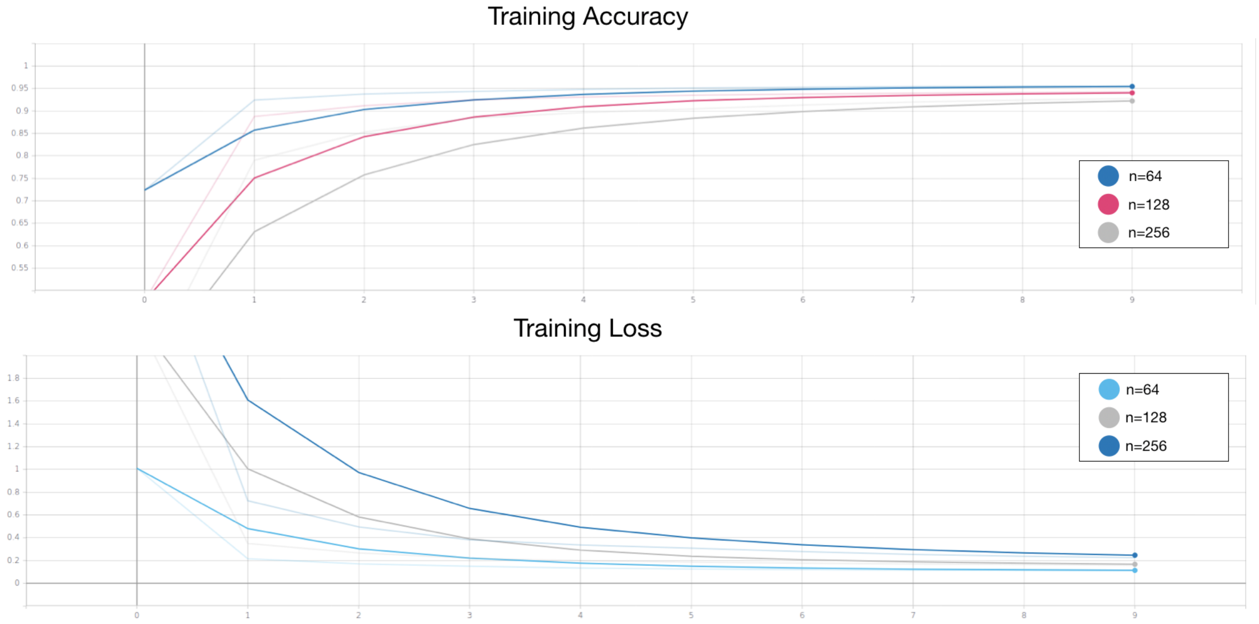

Figure 3: Training accuracy and loss curves for CPC trained with N values of: $N = 64$, $N = 128$, $N = 256$.

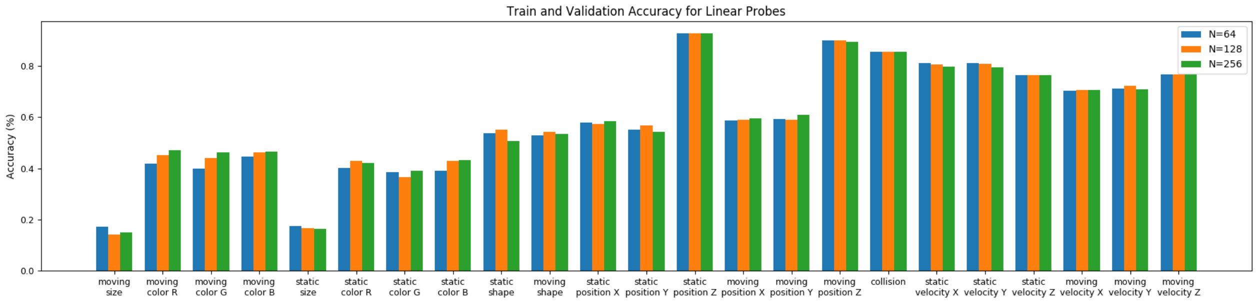

Figure 4: Results for linear probing using C_t from the trained CPC models with N values of: $N = 64$, $N = 128$, $N = 256$.

Effect of Prediction Steps (k) Figure 5 shows the results of training CPC on three values of k: $k = 4$, $k = 10$, $k = 20$. Unfortunately, it does not appear that the model for $k = 20$ converged within 10 epochs. However, the learned representation C_t, shown in Figure 6, seems to perform as well as the $k = 4$ and $k = 20$, despite this. We would be interested to see if a fully trained $k = 20$ model would outperform the others in linear probing. This would give evidence that larger values of k do indeed produce a richer representation by forcing the model to learn more global features.

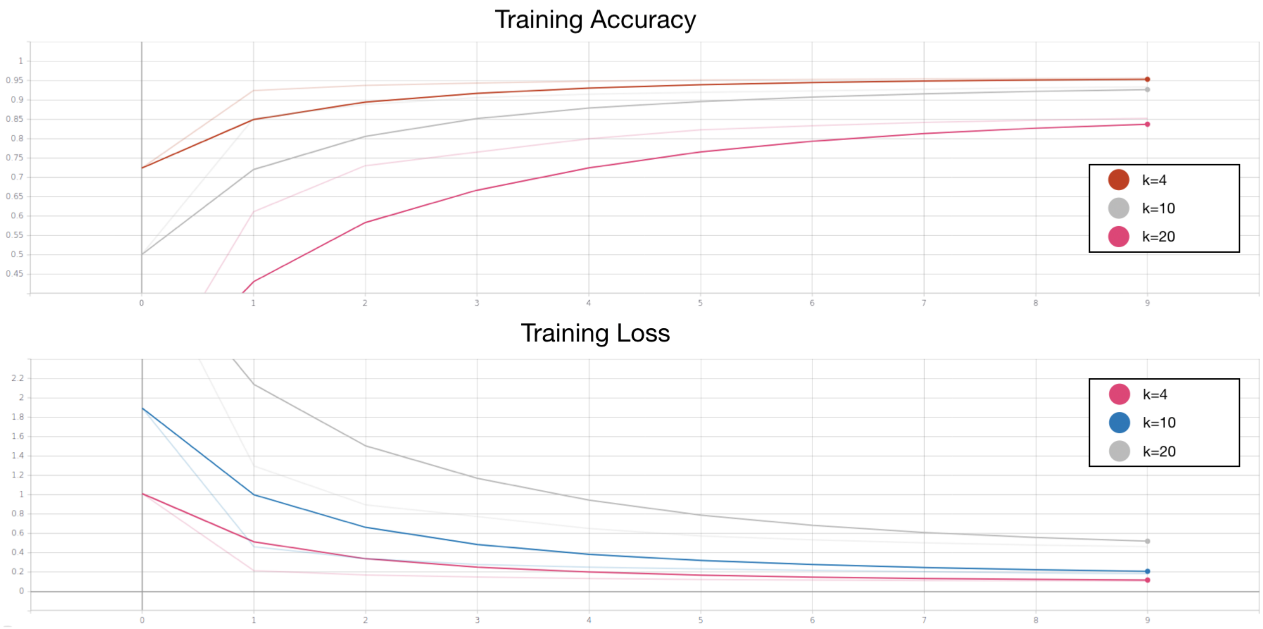

Figure 5: Training accuracy and loss curves for CPC trained with k values of: $k = 4$, $k = 10$, $k = 20$.

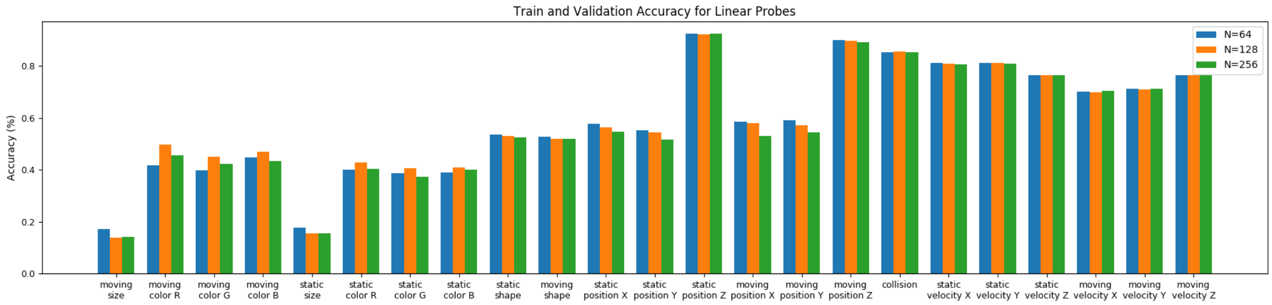

Figure 6: Results for linear probing using C_t from the trained CPC models with k values of: $k = 4$, $k = 10$, $k = 20$.