1 Motivations

1. Universal unsupervised learning from high-dimensional data. Universal as in domain-agnostic.
2. The goal is robust and generic representation learning. They argue that the problem of learning representation from labelled data is that the representation learnt is not generic: when pretraining a model on image classification, the features obtained might be good for related tasks, but does not include information that are more relevant to other tasks, such as color or object counting.
3. They combine predictive coding and noise contrastive estimation.

2 Main Components

2.1 Predictive Coding

1. In predictive coding, the task is to predict the future, missing or contextual information. The label is denoted x and the input is denoted c.
2. The assumption is that the latent representation learnt for predictive coding will also be useful for downstream tasks, because predictive coding is such a generic task.
3. They argue for the need to predict many timesteps into the future, instead of only one step, to extract slowly changing global structure in the data and discards low-level details and noise.
4. They mention that the main challenge in modeling and predicting high-dimensional data is the need to model every details in the data to maximize $p(x|c)$, thus wasting model capacity. Thus, modeling $p(x|c)$ might not be a good method to extract the structure and information that are shared between x and c.
5. They thus propose to encode x and c into a compact vector representation and optimize the encoding such that it maximizes the mutual information between the encoded x and c, thus maximizing the mutual information between x and c.

$$I(x;c) = \sum_{x,c} p(x,c) \log \frac{p(x|c)}{p(x)}$$

6. This is because the mutual information between the encoded variables is upper-bounded by the mutual information between the input x and c into the encoder.
7. By maximizing the mutual information between the encoded variables, they extract the shared information between x and c in the form of latent variable.
8. In the paper, x is future data in a sequence of data and c is the data up to a point in the sequence. The architecture is as follows.
9. Given an input sequence x_t, they use an encoder to map each item in the sequence to its latent representation $z_t = g_{\text{enc}}\left(x_t\right)$.
10. Next, an autoregressive model takes in all $z_{\leq t}$ and produces the and produce a latent representation of the context $c_t = g_{\text{ar}}(z \leq t)$.
11. As they have argued against predicting the future given the context with a generative model $p(x|c)$, to maximize the mutual information, which involves this conditional probability, they instead form a proxy term, which should be proportional to the mutual information under the same distribution:

$$f_k\left(x_{t+k}, c_t\right) \propto \frac{p\left(x_{t+k}|c_t\right)}{p\left(x_{t+k}\right)}$$

12. They choose f to be a log-bilinear model:

$$f_k\left(x_{t+k}, c_t\right) = \exp\left(z_{t+k}^T W_k c_t\right)$$

13. They argue that the use of the proxy term removes the need to model the complicated high-dimensional conditional distribution $p(x|c)$.

2.2 InfoNCE

1. Both the encoder and the autoregressive model are jointly trained to maximize a loss based on NCE, which they refer to as InfoNCE.

2. Given a set $X = \{x_1, \ldots, x_N\}$ of N random samples, containing one positive sample from $p\left(x_{t+k} \mid c_t\right)$ and $N-1$ negative example from a proposal distribution that does not conditioned on c_t $p\left(x_{t+k}\right)$, they optimize for

$$\mathcal{L}_{\mathrm{N}} = -\underset{X}{\mathbb{E}}\left[\log \frac{f_k\left(x_{t+k}, c_t\right)}{\sum_{x_j \in X} f_k\left(x_j, c_t\right)}\right]$$

3. They argue that optimizing for this loss will lead f to estimate the density ratio of interest in the mutual information between the unencoded x and c.

4. The mutual information between the variable c_t and x_{t+k} can be evaluated by

$$I\left(x_{t+k}, c_t\right) \geq \log(N) - \mathcal{L}_N$$

5. Thus, minimizing the loss above is equivalent to maximizing a lower bound on the mutual information.

3 Experiments

1. They perform experiments on audio, vision, nlp and rl to demonstrate the generality of their approaches. Below list interesting tibits from the experimental section.

2. In the audio experiment, they show that the prediction task becomes much harder as the prediction window increases. This shows that the objective function is neither trivial or impossible.

3. They found that not all the information encoded by the CPC networks are linearly separable. This was demonstrated by comparing the performance of a linear classifier with a non-linear one with the input being the CPC encoding for a classification task. The non-linear classifier performs much better and closer to the performance of the fully supervised model.

4. In the audio experiment, their model processes information that is slightly longer than one second, which shows that there is potentially a lot more mileage to be gained by scaling up the model.

4 Other interesting bits

1. They argue that the representation obtained by either the encoder or the autoregressive model can be used for downstream tasks. The representation obtained by the autoregressive model can be used if extra context from the past might be useful.

5 Question

1. Is it the trend that with more data, a more domain-agnostic approach will perform better than a domain-specific approach?

2. Is there any particular reason why they choose to use the log-bilinear model?

3. I don't quite get the mathematical argument for the InfoNCE loss yet. Need to go through their appendix to see step-by-step derivation.