Representation Learning with Contrastive Predictive Coding

DeepMind

정영석

Introduction

▪ Motivation of Pre-training

- Improving representation learning requires features that are less specialized towards solving a single supervised task.
  - ✓ Example) Image Classification : Image Texture
    Automatic Speech Recognition : Speaker id
  - ✓ unsupervised learning is an important stepping stone towards robust and generic representation learning.
- One of the most common strategies for unsupervised learning has been to predict future, missing or contextual information.

Introduction

Proposal Methods

- We compress high-dimensional data into a much more compact latent embedding space in which conditional predictions are easier to model.
- We use powerful autoregressive models in this latent space to make predictions many steps in the future.
- We rely on Noise-Contrastive Estimation for the loss function in similar ways that have been used for learning word embeddings in natural language models, allowing for the whole model to be trained end-to-end.

Methodology

▪ Contrastive Predictive Coding (CPC)

- Motivation and Intuitions

✓ Slow Feature (Global Information)

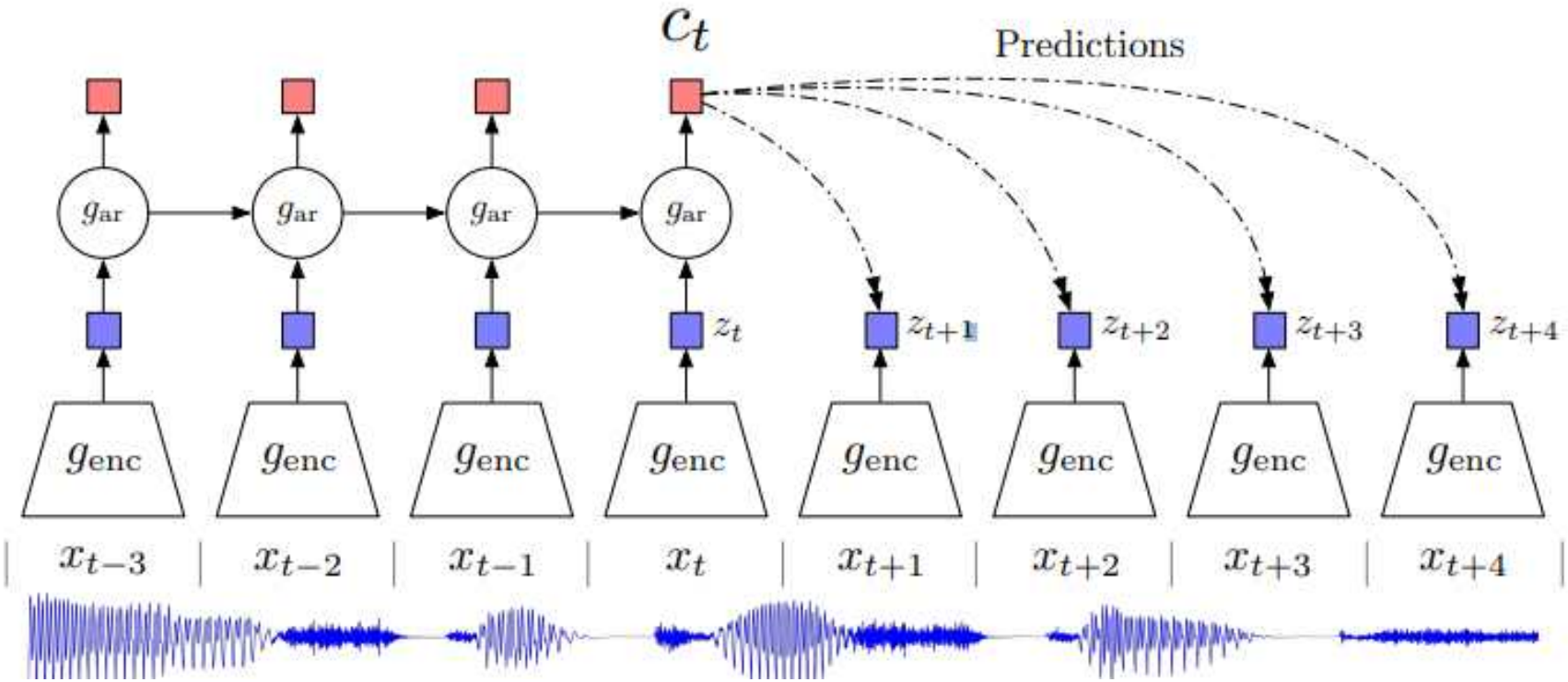
✓ The main intuition behind our model is to learn the representations that encode the underlying shared information between different parts of the (high-dimensional) signal.

✓ Ex) story line in books, objects in images

✓ Models trained via MSE or cross entropy are computationally intense, and waste capacity at modeling the complex relationships in the data x, often ignoring the context c.

Methodology

- **Contrastive Predictive Coding (CPC)**



Methodology

▪ Contrastive Predictive Coding (CPC)

- Instead of direct predicting x, the model predict density ratio which preserve mutual information.

$$I(x;c) = \sum_{x,c} p(x,c) \log \frac{p(x|c)}{p(x)}. \quad f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$$

- Simple log-bilinear model

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

Methodology

▪ Contrastive Predictive Coding (CPC)

- Why the f(·) convergence to density proportion.

$$p(d=i|X,c_t) = \frac{p(x_i|c_t)\prod_{l\neq i} p(x_l)}{\sum_{j=1}^{N} p(x_j|c_t)\prod_{l\neq j} p(x_l)} = \frac{\frac{p(x_i|c_t)}{p(x_i)}}{\sum_{j=1}^{N}\frac{p(x_j|c_t)}{p(x_j)}}.$$

- InfoNCE Loss

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

Methodology

▪ Contrastive Predictive Coding (CPC)

- Mutual information between c and x

$$
\begin{aligned}
\mathcal{L}_{\mathrm{N}}^{\mathrm{opt}} &= -\underset{X}{\mathbb{E}} \log \left[\frac{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})}}{\frac{p(x_{t+k}|c_t)}{p(x_{t+k})} + \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)}} \right] \\
&= \underset{X}{\mathbb{E}} \log \left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} \sum_{x_j \in X_{\mathrm{neg}}} \frac{p(x_j|c_t)}{p(x_j)} \right] \\
&\approx \underset{X}{\mathbb{E}} \log \left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} (N-1) \underset{x_j}{\mathbb{E}} \frac{p(x_j|c_t)}{p(x_j)} \right] \\
&= \underset{X}{\mathbb{E}} \log \left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} (N-1) \right] \\
&\geq \underset{X}{\mathbb{E}} \log \left[\frac{p(x_{t+k})}{p(x_{t+k}|c_t)} N \right] \\
&= -I(x_{t+k}, c_t) + \log(N),
\end{aligned}
$$

$$
\rightarrow I(x_{t+k}, c_t) \geq \log(N) - \mathcal{L}_{\mathrm{N}},
$$

Experiment

- **Audio**

Figure 2: t-SNE visualization of audio (speech) representations for a subset of 10 speakers (out of 251). Every color represents a different speaker.

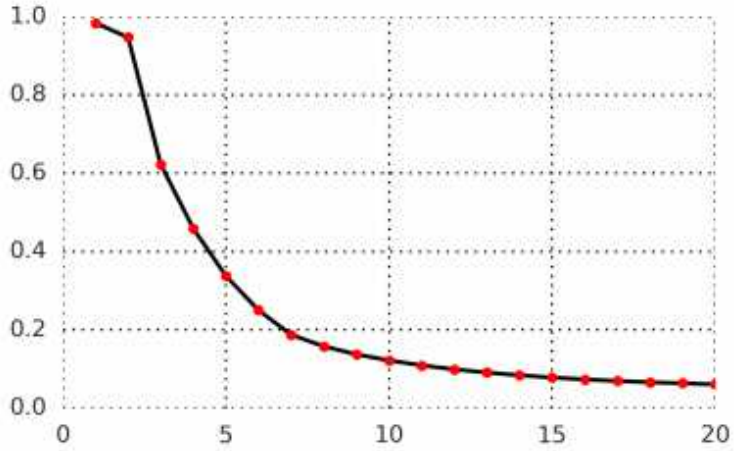


Figure 3: Average accuracy of predicting the positive sample in the contrastive loss for 1 to 20 latent steps in the future of a speech waveform. The model predicts up to 200ms in the future as every step consists of 10ms of audio.

Experiment

- **Audio**

Method	ACC
Phone classification	
Random initialization	27.6
MFCC features	39.7
CPC	64.6
Supervised	74.6
Speaker classification	
Random initialization	1.87
MFCC features	17.6
CPC	97.4
Supervised	98.5

Table 1: LibriSpeech phone and speaker classification results. For phone classification there are 41 possible classes and for speaker classification 251. All models used the same architecture and the same audio input sizes.

Method	ACC
#steps predicted	
2 steps	28.5
4 steps	57.6
8 steps	63.6
12 steps	64.6
16 steps	63.8
Negative samples from	
Mixed speaker	64.6
Same speaker	65.5
Mixed speaker (excl.)	57.3
Same speaker (excl.)	64.6
Current sequence only	65.2

Table 2: LibriSpeech phone classification ablation experiments. More details can be found in Section 3.1.

Experiment

- Vison pre-training

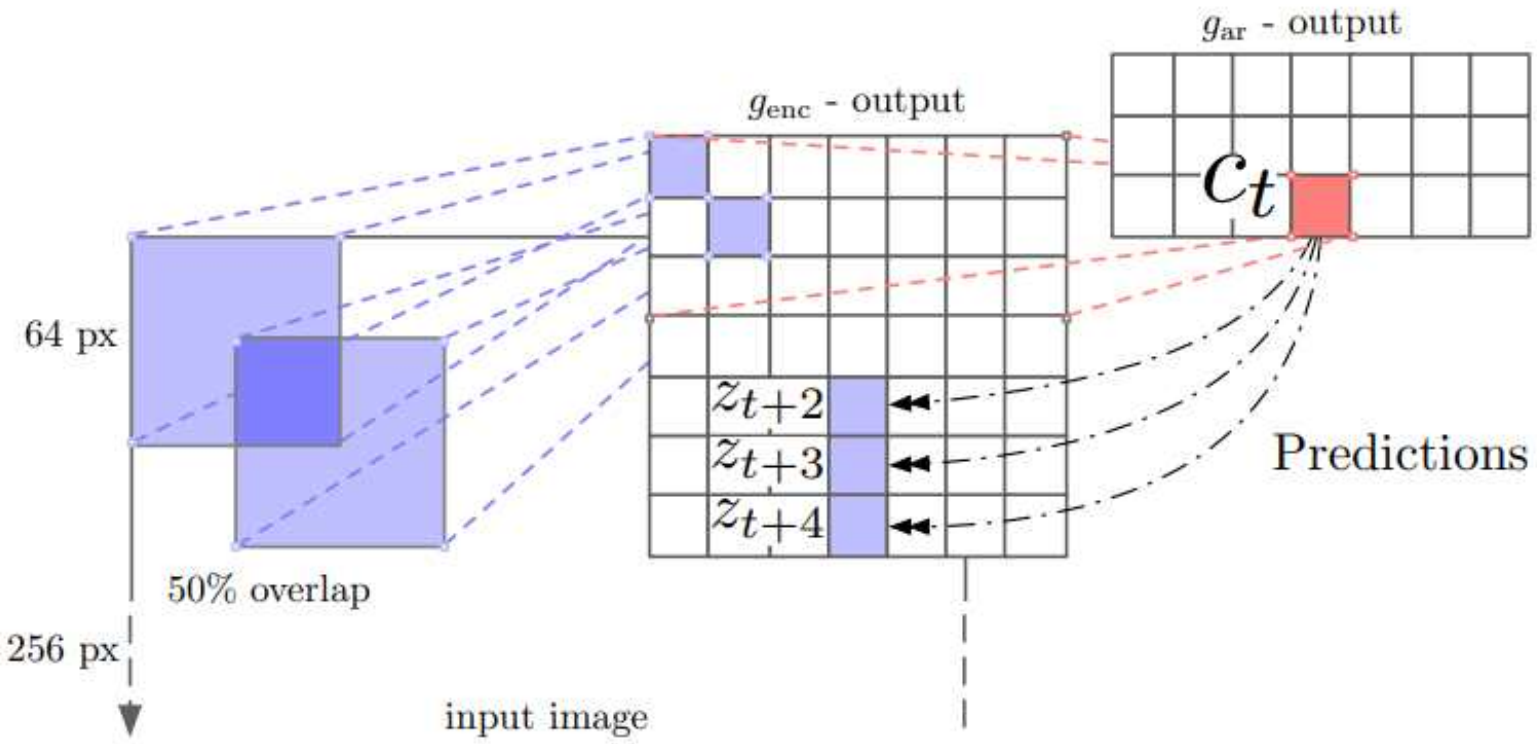


Figure 4: Visualization of Contrastive Predictive Coding for images (2D adaptation of Figure 1).

Experiment

- Vision

Method	Top-1 ACC
Using AlexNet conv5	
Video [28]	29.8
Relative Position [11]	30.4
BiGan [35]	34.8
Colorization [10]	35.2
Jigsaw [29] *	38.1
Using ResNet-V2	
Motion Segmentation [36]	27.6
Exemplar [36]	31.5
Relative Position [36]	36.2
Colorization [36]	39.6
CPC	**48.7**

Table 3: ImageNet top-1 unsupervised classification results. *Jigsaw is not directly comparable to the other AlexNet results because of architectural differences.

Method	Top-5 ACC
Motion Segmentation (MS)	48.3
Exemplar (Ex)	53.1
Relative Position (RP)	59.2
Colorization (Col)	62.5
Combination of MS + Ex + RP + Col	69.3
CPC	**73.6**

Table 4: ImageNet top-5 unsupervised classification results. Previous results with MS, Ex, RP and Col were taken from [36] and are the best reported results on this task.