CPC v1:

Representation Learning with Contrastive Predictive Coding

Aaron van den Oord, Yazhe Li, Oriol Vinyals

DeepMind

Sungman, Cho.

Introduction

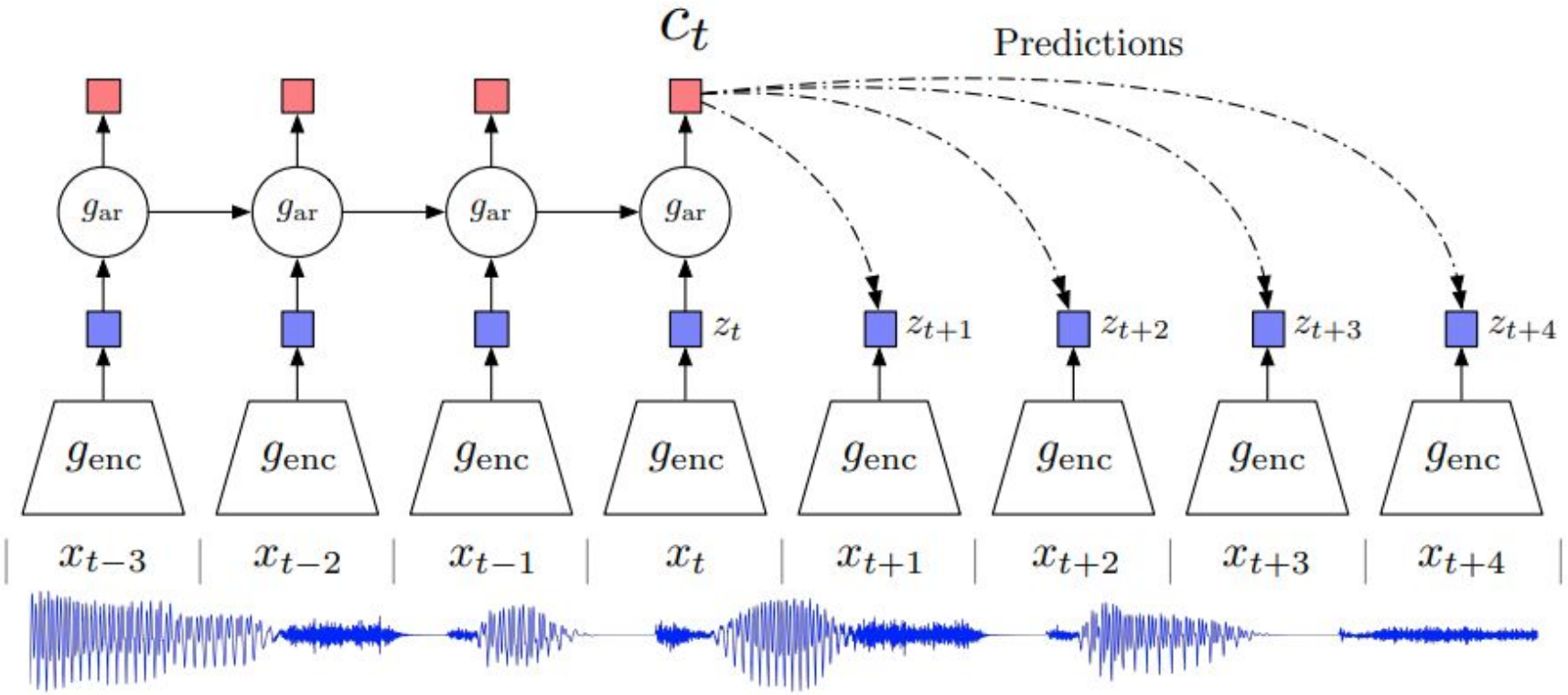
- Improving representation learning requires **features that are less specialized towards solving a single supervised task.**

- One of the most **common strategies for unsupervised learning** has been to **predict future.**

Introduction

- **Compress** high-dimensional data into compact **latent embedding space.**
- Use **powerful autoregressive models** in this latent space to make predictions.
- Rely on Noise-Contrastive Estimation for the loss function.

Introduction



Why CPC ?

- 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.**

  At the same time **it discards low-level information and noise** that is more local.

- **Slow features** : invariant to time steps.

Why CPC ?

- **Unimodal losses** (Mean Squared Error, Cross Entropy) are not appropriate loss function to learn representations.

- **Generative models** is powerful. but computationally intense, and waste capacity at modeling the complex relationships in the data, often ignoring the context.

Why CPC ?

- In ImageNet,
    - Input image : 224 x 224 x 3 x 8 bit → x
    - Class : 10bit (1,000 classes) → c
- Modeling $p(x|c)$ directly may not be optimal for the purpose of extracting shared information between x and c .

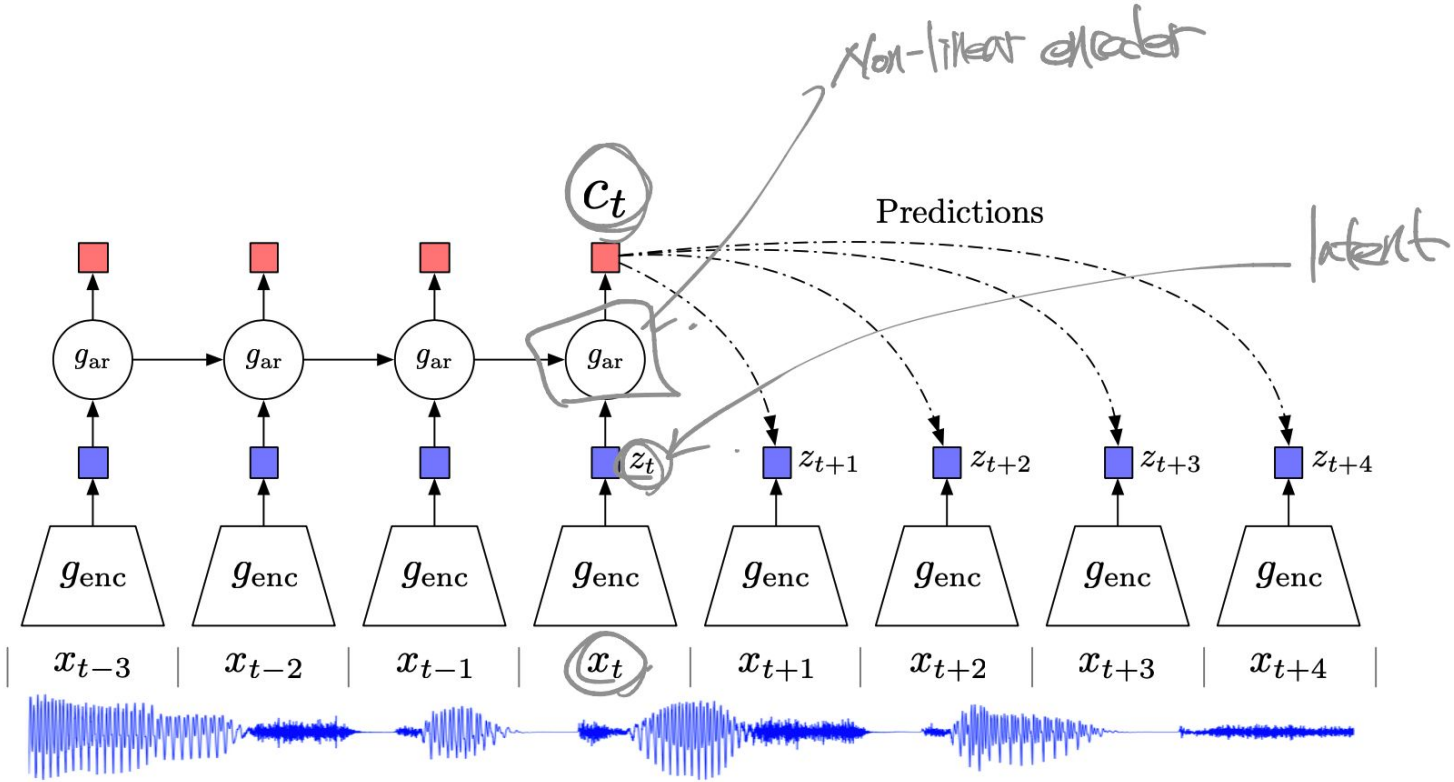
Why CPC ?

- Target : x (future), Context : c (present)

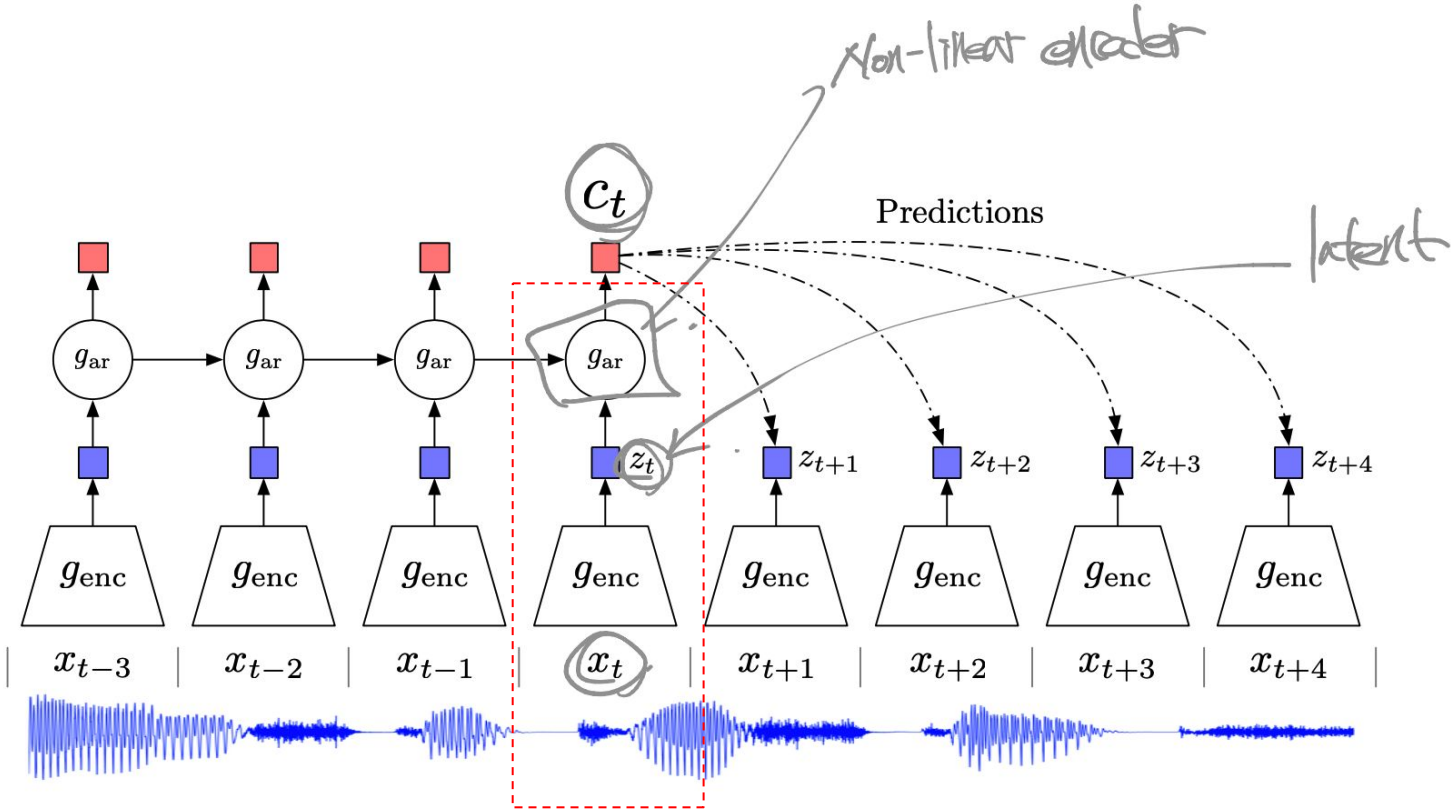
$$I(x;c) = \sum_{x,c} p(x,c) \log \frac{p(x|c)}{p(x)}.$$

- By maximizing the MI(Mutual Information), we extract the underlying latent variables the inputs have in common.

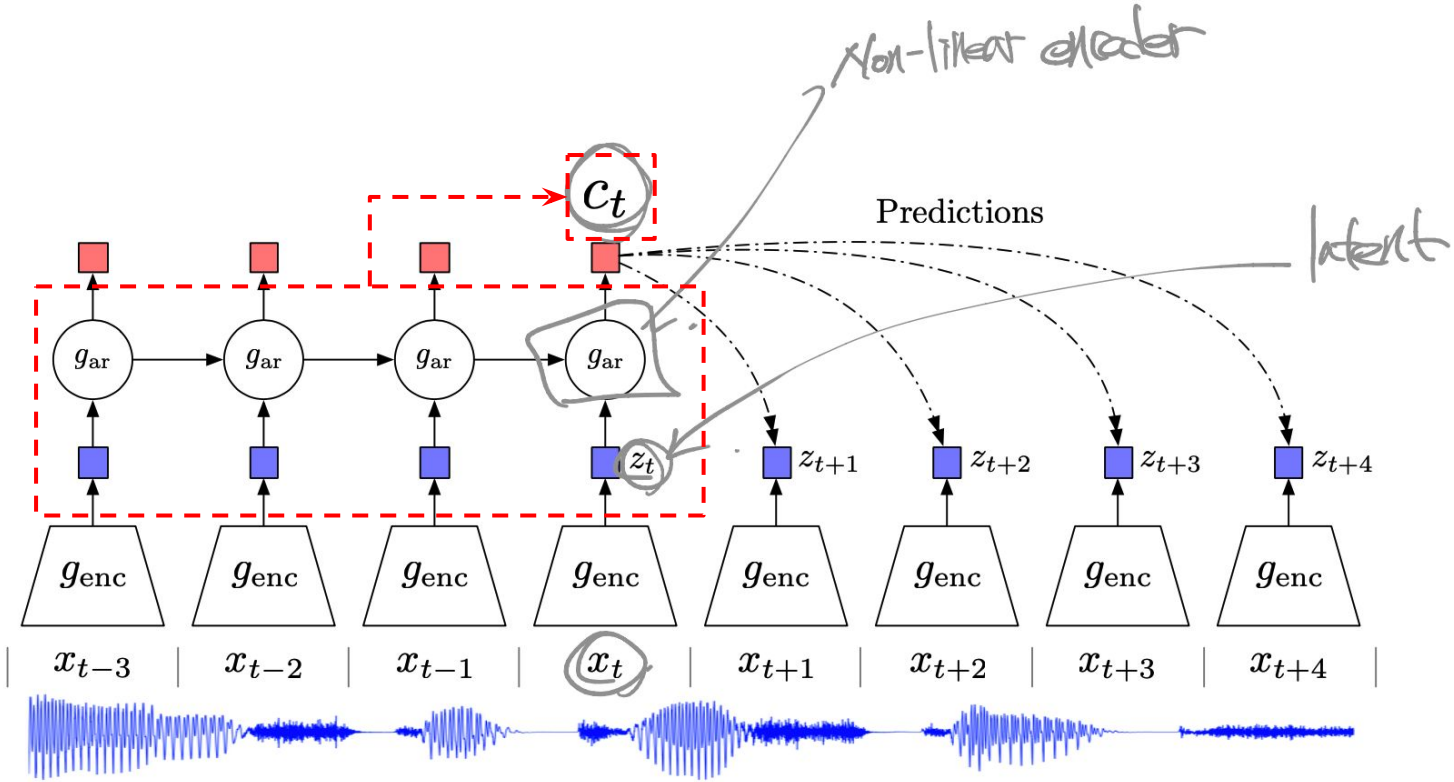
CPC ?


Non-linear encoder
Predictions
Intent
c_t
g_{ar}
g_{enc}
z_t
z_{t+1}
z_{t+2}
z_{t+3}
z_{t+4}
x_{t-3}
x_{t-2}
x_{t-1}
x_t
x_{t+1}
x_{t+2}
x_{t+3}
x_{t+4}

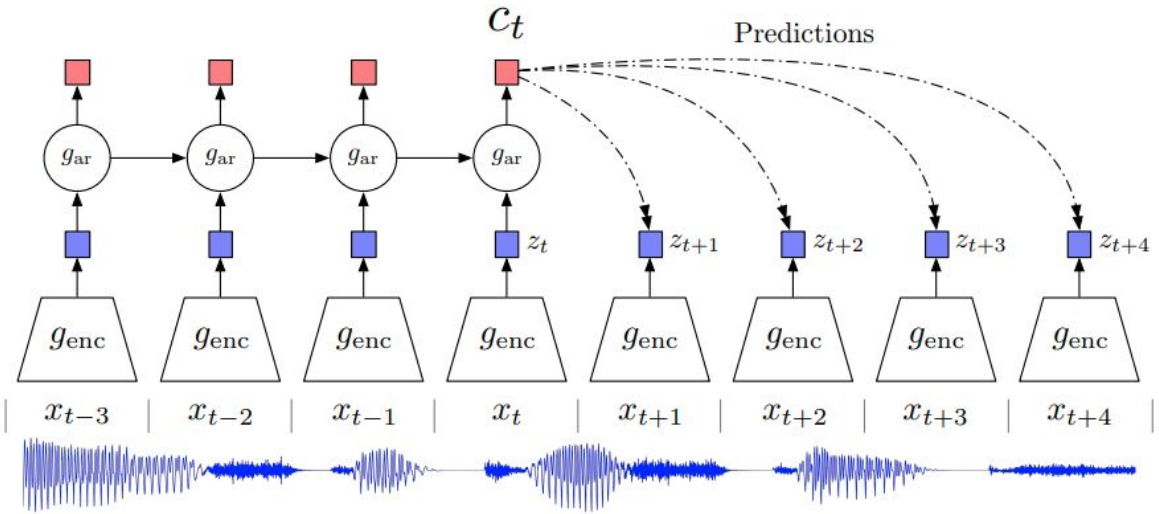
CPC ?


Non-linear encoder
Predictions
Intent
c_t
g_{ar}
g_{enc}
z_t
z_{t+1}
z_{t+2}
z_{t+3}
z_{t+4}
x_{t-3}
x_{t-2}
x_{t-1}
x_t
x_{t+1}
x_{t+2}
x_{t+3}
x_{t+4}

CPC ?


Predictions
c_t
g_{ar}
g_{ar}
g_{ar}
g_{ar}
z_t
z_{t+1}
z_{t+2}
z_{t+3}
z_{t+4}
g_{enc}
g_{enc}
g_{enc}
g_{enc}
g_{enc}
g_{enc}
g_{enc}
g_{enc}
x_{t-3}
x_{t-2}
x_{t-1}
x_t
x_{t+1}
x_{t+2}
x_{t+3}
x_{t+4}

CPC ?

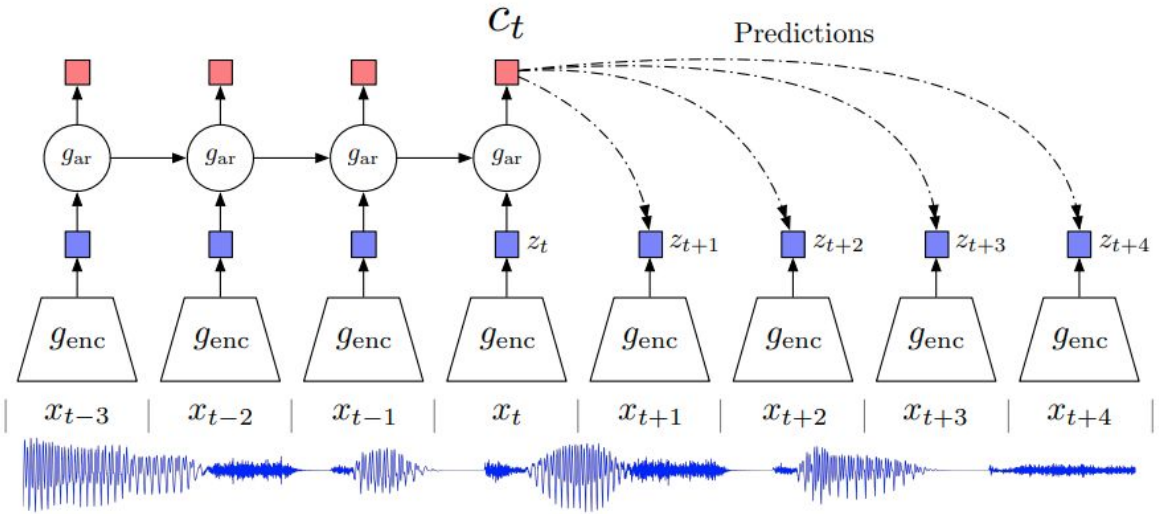


$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$$

Log-bilinear model

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

CPC ?



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

Log-bilinear model

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

Transformation
(linear, nonlinear, recurrent)

CPC ?

Transformation
(linear, nonlinear, recurrent)

$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})} \quad \text{Log-bilinear model} \quad f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right),$$

Can use sampling techniques such as Noise-Contrastive Estimation and Importance Sampling

Representation for downstream task.

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

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

Experiments

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.

Experiments

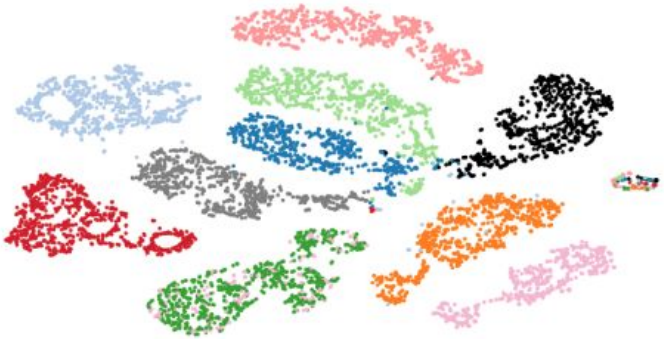


Figure 2: t-SNE visualization of audio (speech) representations for a subset of 10 speakers (out of 251). Every color represents a different speaker.

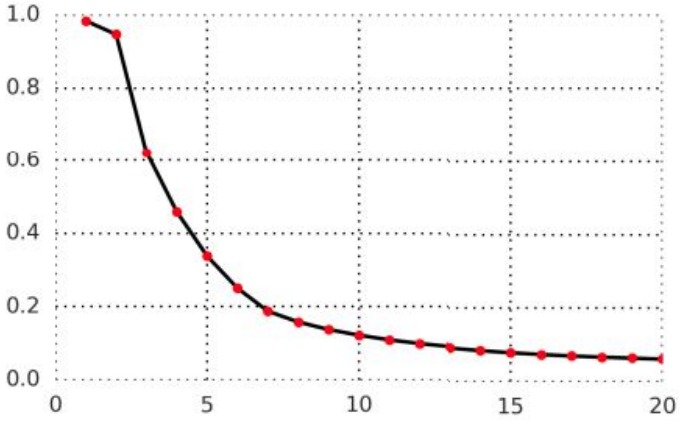
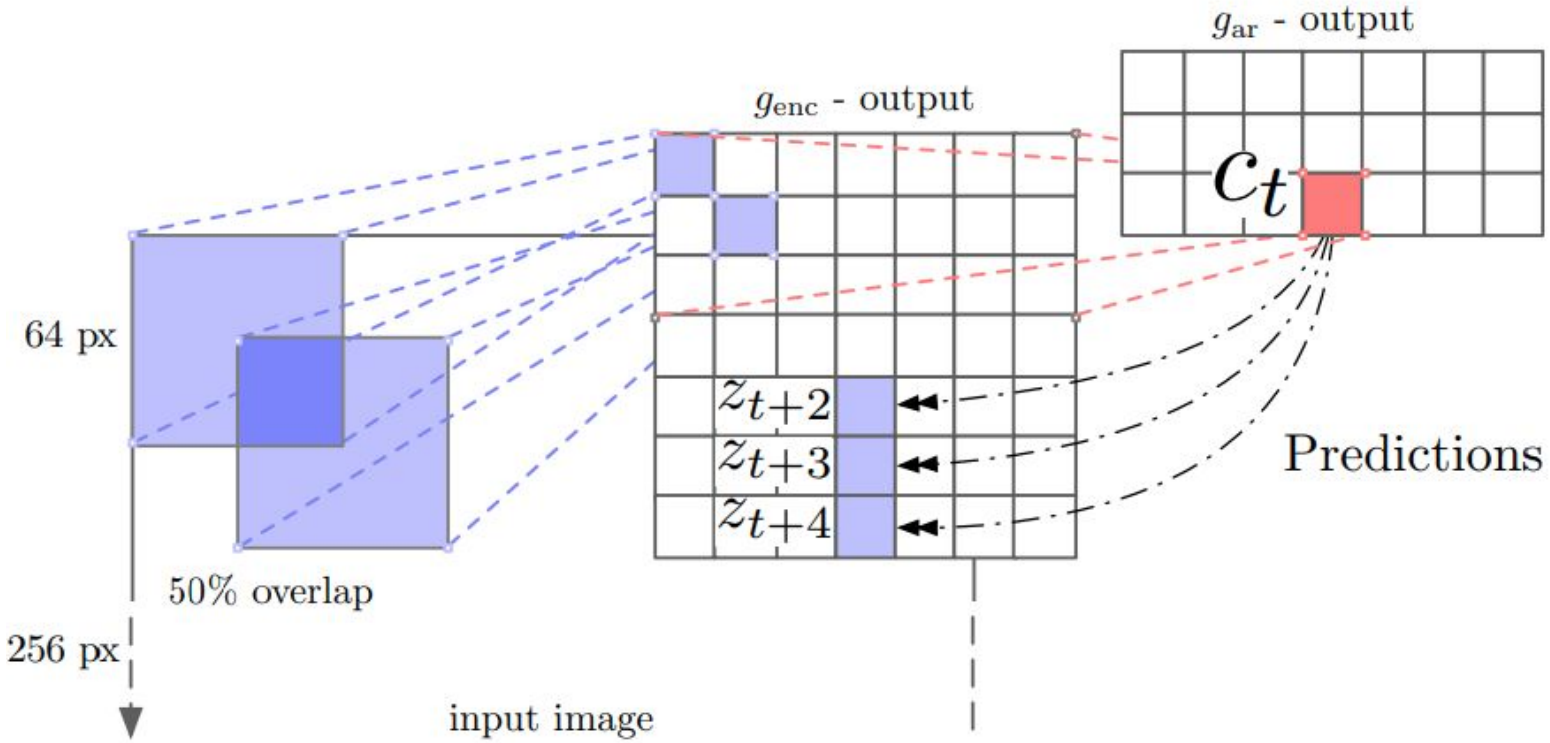


Figure 3: Average accuracy of predicting the positive sample in the contrastive loss for 1 to 20 latent steps in the future of a speech waveform. The model predicts up to 200ms in the future as every step consists of 10ms of audio.

Experiments



Experiments

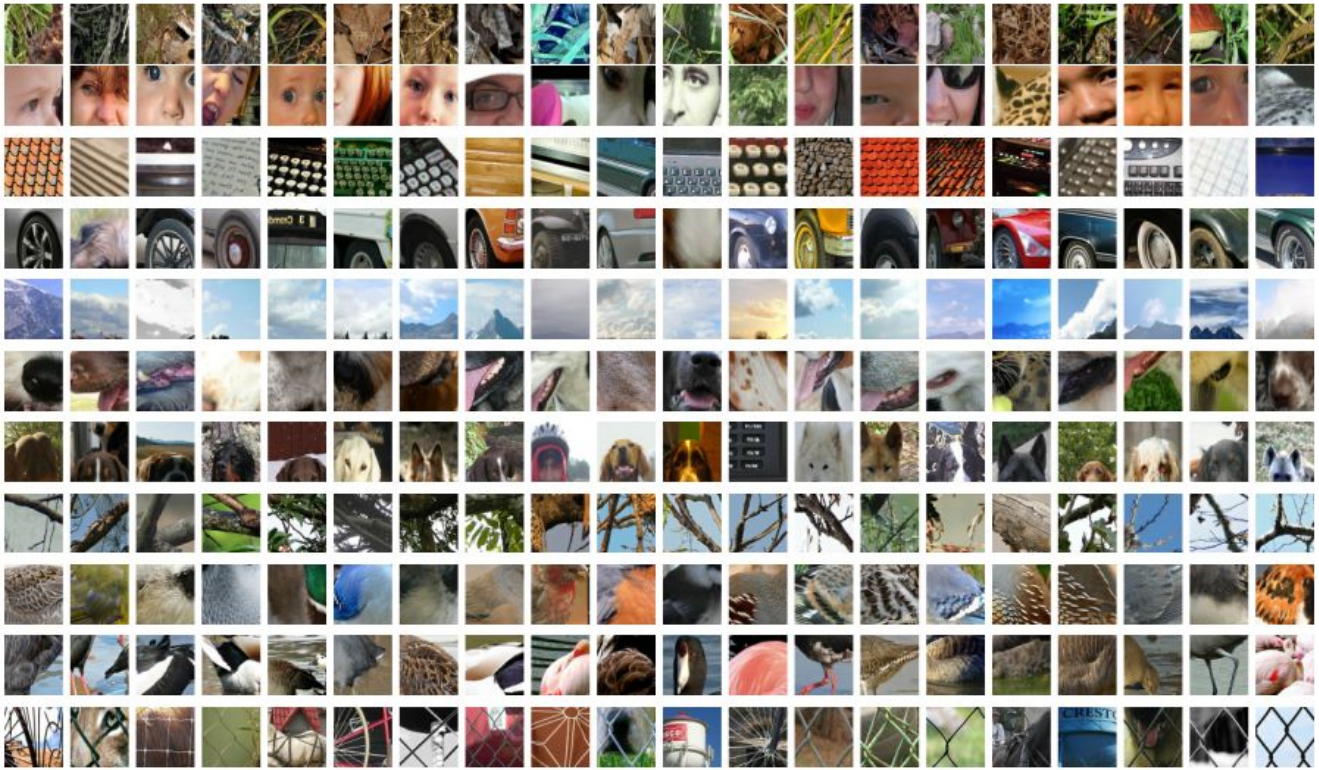


Figure 5: Every row shows image patches that activate a certain neuron in the CPC architecture.

Thank you.