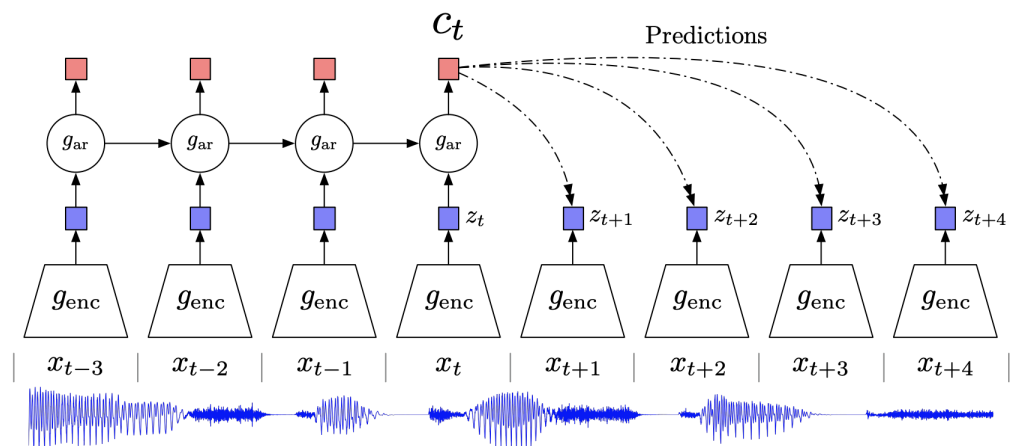


# TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2021

## Mutual Information Coding

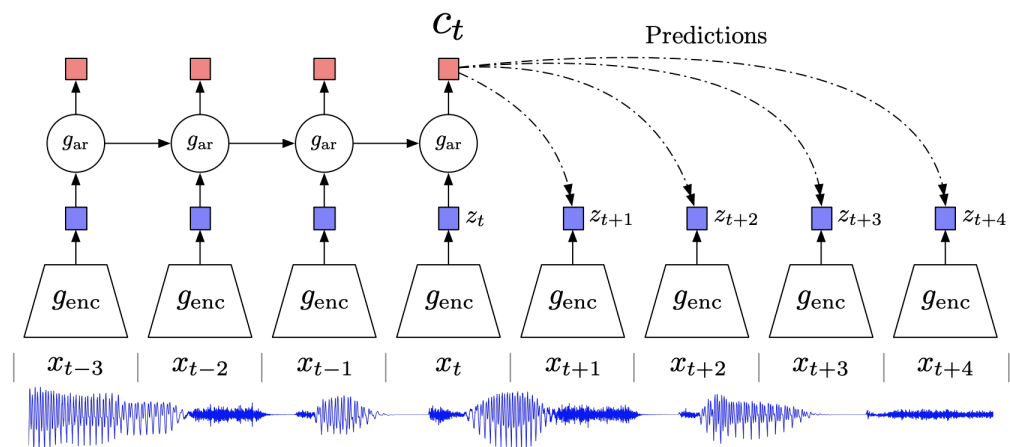
# Mutual Information Coding



van den ORD et al. 2018

Consider the problem of learning phonetic representations of speech sounds. In the figure each  $z_t$  is a symbol representing the sound at time  $t$ .

# Mutual Information Coding



van den ORD et al. 2018

Unlike VAEs, mutual information coding is about **predicting latent variables**. There is no attempt to model the input speech sound. Intuitively we want to separate signal from noise and avoid modeling noise.

## wav2vec 2.0

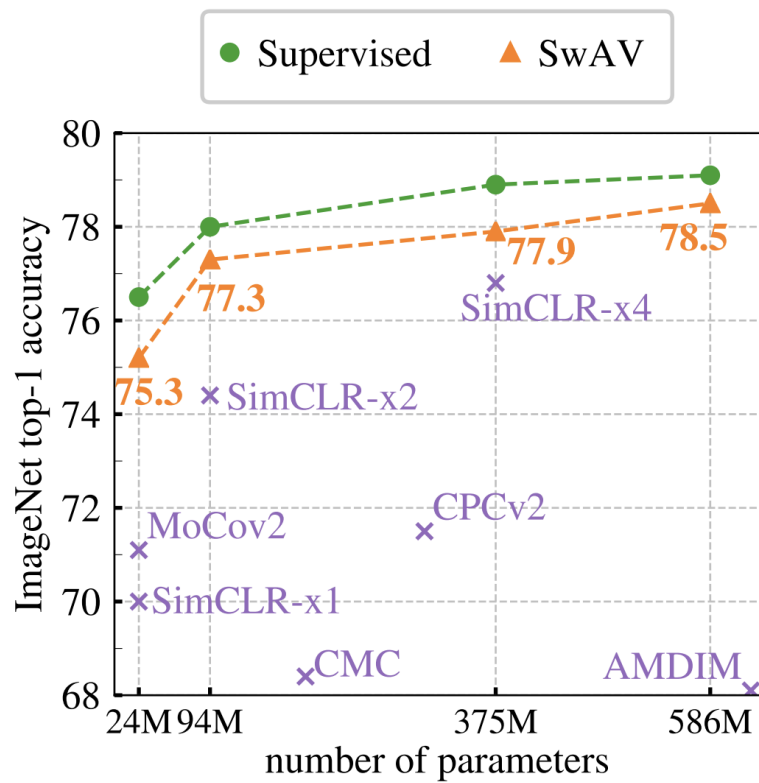
Trained on 53k hours of **unlabeled** audio they convert speech to a sequence of symbols they call “pseudo-text units”.

Using this pre-trained transcription of speech into pseudo-text they reduce the amount of labeled data needed for a given accuracy in speech recognition by a factor of 100.

Baevski et al., 2020

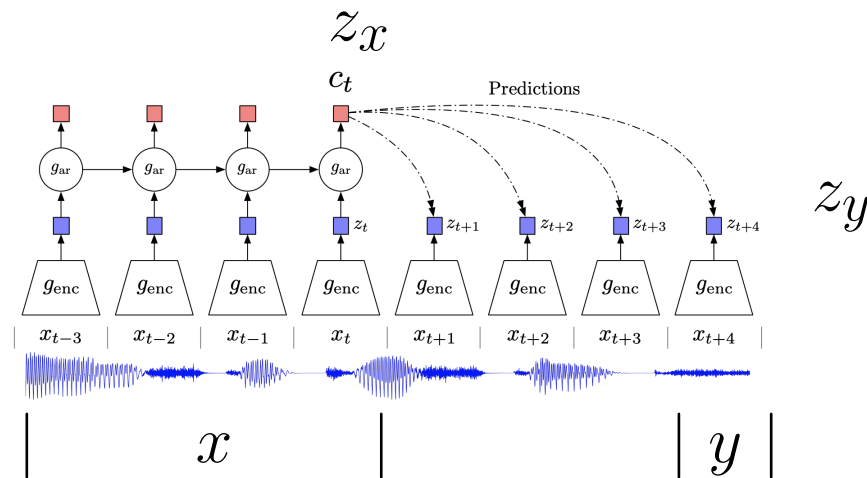
# SwAV

Mutual information coding as pretraining of image features.



Caron et al. 2021

# Mutual Information Coding: General Formulation



Consider a population distribution on pairs  $\langle x, y \rangle$ .

We are interested in extracting latent variables  $z_x$  and  $z_y$  from  $x$  and  $y$  respectively that preserve the mutual information between  $x$  and  $y$ .

## Mutual Information Coding General Formulation

For a population on  $\langle x, y \rangle$  we introduce latent variables  $z_x$  and  $z_y$  defined by mappings  $z_x(x)$  and  $z_y(y)$  where these mappings are defined by parameters  $\Phi_x$  and  $\Phi_y$  respectively.

$$\Phi_x^*, \Phi_y^* = \operatorname{argmax}_{\Phi_x, \Phi_y} I(z_x, z_y)$$

This has a degenerate solution of  $z_x(x) = x$  and  $z_y(y) = y$  but this can be avoided with various types of restrictions on  $z_x$  and  $z_y$  as described below.

## Contrastive Predictive Coding (CPC)

For  $z_x$  and  $z_y$  vectors, and for  $N \geq 2$ , we define a distribution on tuples  $(z_x, z_y^1, \dots, z_y^N, i)$  by drawing pairs  $(x_1, y_1), \dots, (x_n, y_n)$  from the population, and  $i$  uniformly from 1 to  $N$ , and constructing

$$(z_x(x_i), z_y(y_1), \dots, z_y(y_N), i).$$

We then train a model to predict  $i$ .

$$\Phi^* = \operatorname{argmin}_{\Phi} E_{(z_x, z_y^1, \dots, z_y^N, i)} [-\ln P_{\Phi}(i | (z_x, z_y^1, \dots, z_y^N))]$$

$$P_{\Phi}(i | z_x, z_y^1, \dots, z_y^N) = \operatorname{softmax}_i z_x^{\top} z_y^i$$



## The CPC Theorem

For any distribution on pairs  $(z_x, z_y)$ , if CPC probabilities are computed by

$$P_{\Phi}(i|z_x, z_y^1, \dots, z_y^N) = \operatorname{softmax}_i s(z_x, z_y^i)$$

then

$$I(z_x, z_y) \geq \ln N - E_{(z_x, z_y^1, \dots, z_y^N, i)} [-\ln P_{\Phi}(i|(z_x, z_y^1, \dots, z_y^N))]$$

Chen et al., On Variational Bounds of Mutual Information,  
May 2019.

## The CPC Restriction on $z_x(x)$ and $x_y(y)$

$$P_{\Phi}(i|z_x, z_y^1, \dots, z_y^N) = \operatorname{softmax}_i z_x^{\top} z_y^i$$

By only using  $z_x$  and  $z_y$  in inner products at the final layer we force the features to carry information in a shallow (linear) representation.

# Contrastive Predictive Coding for Images

(SimCLR:) A Simple Framework for Contrastive Learning of Visual Representations, Chen et al., Feb. 2020 (self-supervised leader as of February, 2020).

They construct a distribution on pairs  $\langle x, y \rangle$  defined by drawing an image from ImageNet and then drawing  $x$  and  $y$  as random “augmentations” (modifications) of the image.

The training maximizes the contrastive lower bound on  $I(x, y)$ .

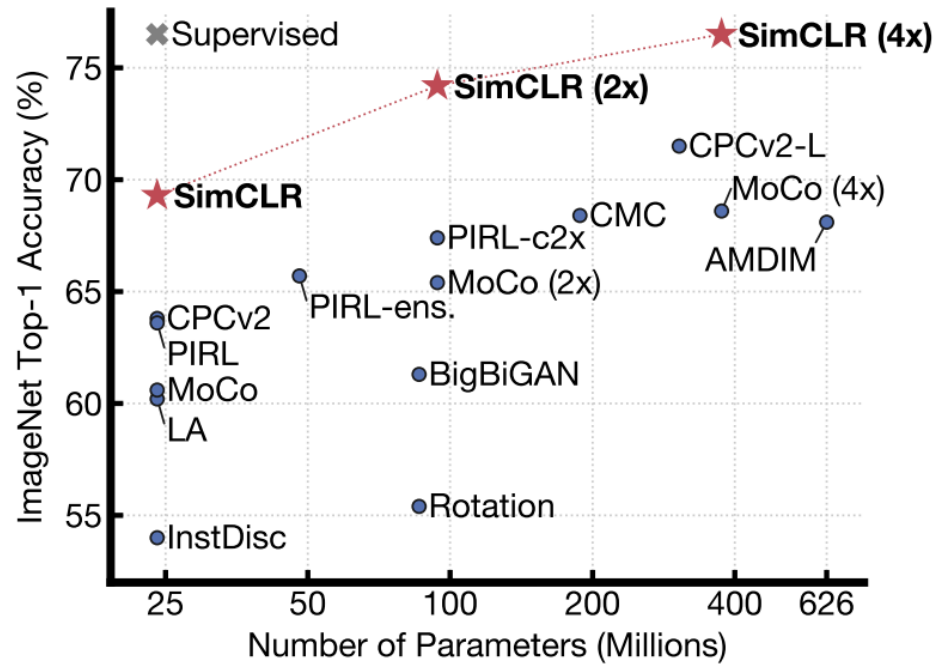
## Contrastive Predictive Coding for Images

A resulting feature map  $z_\Phi$  on images is extracted from this training.

The feature map  $z_\Phi$  is tested by using a **linear** classifier for ImageNet based on these features.

This is called linear probing.

# SimCLR



Chen et al. 2020

## A Weakness of CPC

$$I(z_x, z_y) \geq \ln N - E_{(z_x, z_y^1, \dots, z_y^N, i)} [-\ln P_\Phi(i | (z_x, z_y^1, \dots, z_y^N))]$$

The discrimination problem is too easy.

The guarantee can never be stronger than  $\ln N$  where  $N$  is the number of choices in the discrimination task.

## Direct Mutual Information (MI) Coding

For a population on  $\langle x, y \rangle$  we introduce latent variables  $z_x$  and  $z_y$  defined by mappings  $z_x(x)$  and  $z_y(y)$  where these mappings are defined by parameters  $\Phi_x$  and  $\Phi_y$  respectively.

$$\begin{aligned}\Phi_x^*, \Phi_y^* &= \operatorname{argmax}_{\Phi_x, \Phi_y} I(z_x, z_y) \\ &= \operatorname{argmax}_{\Phi_x, \Phi_y} H(z_y) - H(z_y|z_x) \\ &= \operatorname{argmin}_{\Phi_x, \Phi_y} H(z_y|z_x) - H(z_y)\end{aligned}$$

## Direct MI Coding

$$\Phi_x^*, \Phi_y^* = \underset{\Phi_x, \Phi_y}{\operatorname{argmin}} H(z_y|z_x) - H(z_y)$$

We can block the solution of  $z_x(x) = x$  and  $z_y(y) = y$  by requiring that  $z_y$  is a symbol from a limited finite vocabulary.

This typically ensures  $H(z_y) \ll H(y)$ .

We can then allow  $H(z_x)$  to be large, for example  $z_x$  might be a vector representation of the a symbol sequence  $z_1, \dots, z_t$ .



## Direct MI Coding

$$\Phi_x^*, \Phi_y^* = \operatorname{argmin}_{\Phi_x, \Phi_y} H(z_y|z_x) - H(z_y)$$

If  $z_y$  is a symbol from a limited vocabulary we can estimate  $H(z_y)$  from an empirical histogram over the symbols.

We can use a model  $\Psi$  to predict  $z_y$  from  $z_x$  and train on cross-entropy loss.

## Direct MI Coding

$$\Phi_x^*, \Phi_y^*, \Psi^* = \operatorname{argmin}_{\Phi_x, \Phi_y, \Psi} E_{(x,y) \sim \text{Pop}} \left[ -\ln P_{\Psi}(z_y|z_x) + \ln \hat{P}(z_y) \right]$$

where  $\hat{P}(z_y)$  is an estimate (perhaps an exponential moving average) of the probability of  $z_y$  over the draw of  $(x, y) \sim \text{Pop}$ .

## Direct MI Coding Theorem

$$\Phi_x^*, \Phi_y^*, \Psi^* = \operatorname{argmin}_{\Phi_x, \Phi_y, \Psi} E_{(x,y) \sim P_{\text{op}}} [-\ln P_{\Psi}(z_y|z_x) + \ln P(z_y)]$$

$$I(z_x, z_y) \geq H(z_y) - \hat{H}(z_y|z_x)$$

$$\hat{H}(z_y|z_x) = E_{(x,y) \sim P_{\text{op}}} [-\ln P_{\Psi}(z_y(y)|z_x(x))]$$

## Direct MI Coding Theorem

For  $H(z_y) = \ln K$  where  $K$  is the number of symbols, and  $\hat{H}(z_y|z_x) = 0$  we get

$$I(z_x, z_y) \geq \ln K$$

Which typically improves significantly on the best possible bound  $I(z_x, z_y) \geq \ln N$  from CPC.

# SwAV

SwAV uses direct MI coding rather than SimCLR's CPC.



Caron et al. 2021

**END**