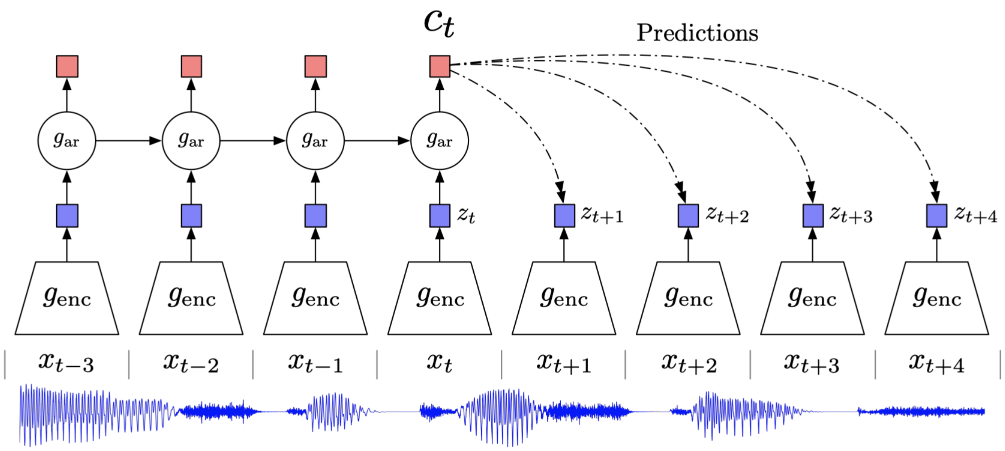
# TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2021

# Mutual Information Coding

# Mutual Information Coding



van den ORD et al. 2018

We want to find codes that maximize the mutual information between the codes for different but related inputs. In the figure we want the codes to maximize the mutual information between $C_t$ and $z_{t+k}$.

# wav2vec 2.0

Mutual information coding as pretraining of speech features.

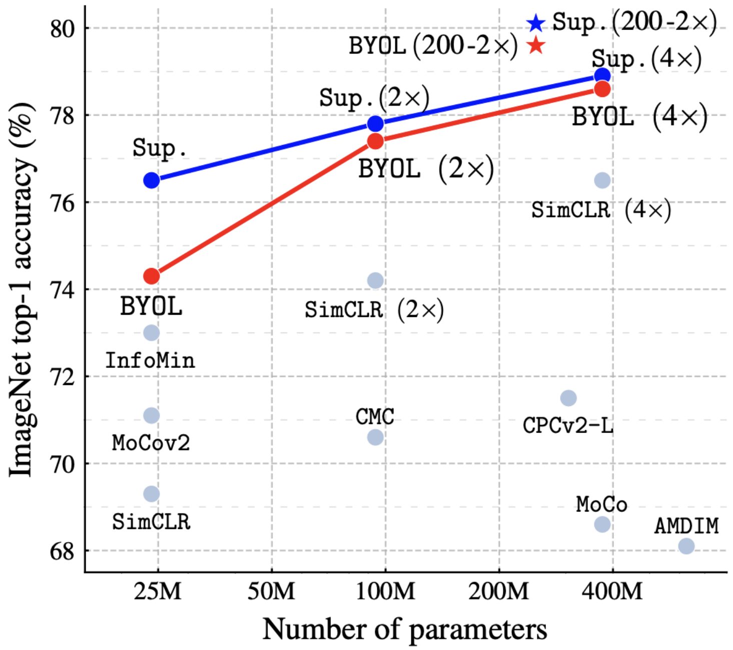
Trained on 53k hours of unlabeled audio (no text) they convert speech to a sequence of discrete quantized vectors they call "pseudo-text units".

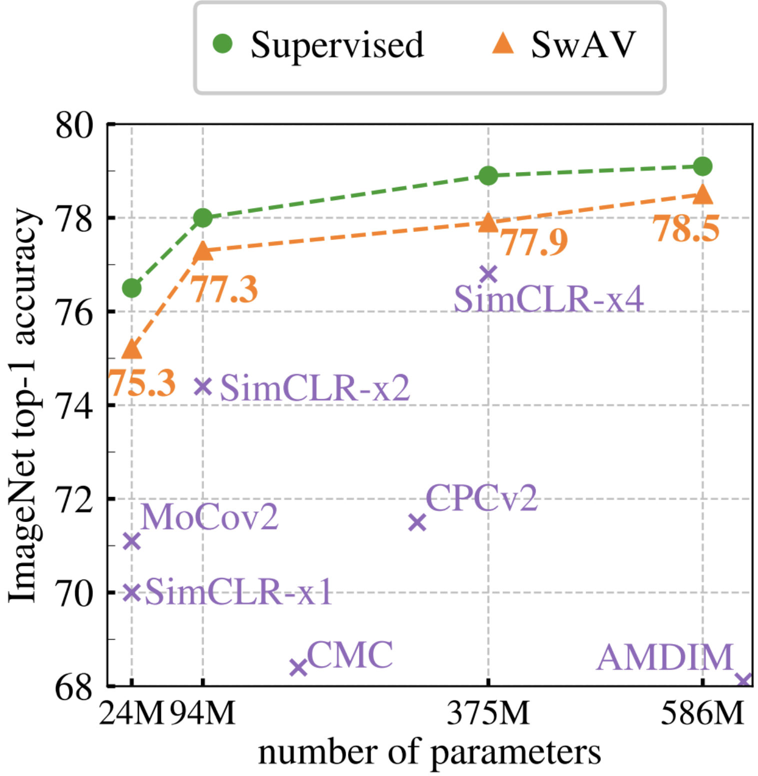
Using the pre-trained transcription of speech sound into pseud-text they achieve the previous state of the art for training on 100 hours of transcriped text but using only one hour instead of 100 hours.

Baevski et al., 2020

# SimCLR, BYOL and SwAV

Mutual information coding as pretraining of image features.





# Mutual Information Coding

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

For example $x$ might be an image and $y$ a sound wave.

We are interested in extracting latent variables $z_x$ and $z_y$ from $x$ and $y$ respectively.

For example $z_x$ might be a bag of words extracted from the image and $z_y$ a bag of words extracted from the sound wave.

# Mutual Information Coding

For a population on $\langle x, y\rangle$ we introduce two discrete latent variables $z_x$ and $z_y$ defined by models $P_\Phi(z_x|x)$ and $P_\Phi(z_y|y)$.

$$\Phi^* = \operatorname*{argmax}_\Phi \; I_{\mathrm{Pop},\Phi}(z_x, z_y) - \beta(H_{\mathrm{Pop},\Phi}(z_x) + H_{\mathrm{Pop},\Phi}(z_y))$$

Here we are asking to maximize the mutual information while (intuitively) limiting the information in $z_x$ and $z_y$.

In the bag of words example we are asking to maximize the mutual information between the two probability distributions on bags of words while limiting the information in the bags.

# Mutual Information Coding

$$\Phi^* = \underset{\Phi}{\mathrm{argmax}}\; I_{\mathrm{Pop},\Phi}(z_x, z_y) - \beta(H_{\mathrm{Pop},\Phi}(z_x) + H_{\mathrm{Pop},\Phi}(z_y))$$

Limiting the information in $z_x$ and $z_y$ prevents the trivial solution of $z_x = x$ and $z_y = y$.

**Here we only model distributions on $z_x$ and $z_y$. Unlike GANs and VAEs, there is no attempt to model distributions on the observables $x$ and $y$.**

# Mutual Information Coding

$$
\begin{aligned}
\Phi^* &= \operatorname*{argmax}_{\Phi}\; I_{\mathrm{Pop},\Phi}(z_x, z_y) - \beta(H_{\mathrm{Pop},\Phi}(z_x)))\\
&= \operatorname*{argmax}_{\Phi}\, (1-\beta)H_{\mathrm{Pop},\Phi}(z_x) - \; H_{\mathrm{Pop},\Phi}(z_x|z_y)\\
&= \operatorname*{argmin}_{\Phi}\; H_{\mathrm{Pop},\Phi}(z_x|z_y) - (1-\beta)H_{\mathrm{Pop},\Phi}(z_x)
\end{aligned}
$$

# Mutual Information Coding

We can estimate entropies by cross-entropies

$$\begin{aligned}
\Phi^* &= \operatorname*{argmin}_{\Phi}\; H_{\mathrm{Pop},\Phi}(z_x|z_y) - (1-\beta)H_{\mathrm{Pop},\Phi}(z_x) \\
&\leq \operatorname*{argmin}_{\Phi}\; \hat{H}_{\mathrm{Pop},\Phi,\Psi}(z_x|z_y) - (1-\beta)H_{\mathrm{Pop},\Phi}(z_x) \\
&\approx ?\; \operatorname*{argmin}_{\Phi}\; \hat{H}_{\mathrm{Pop},\Phi,\Psi}(z_x|z_y) - (1-\beta)\hat{H}_{\mathrm{Pop},\Phi,\Theta}(z_x)
\end{aligned}$$

Here the cross-entropy models $\Psi$ ans $\Theta$ are adversarial.

# Contrastive Predictive Coding

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}}\ I_{\mathrm{Pop},\Phi}(z_x, z_y)$$

It turns out that we can give a lower bound on the mutual information term using **noise contrastive estimation**.

# A Contrastive Lower Bound

We now give a contrastive lower bound for general mutual information $I(z, w)$ given only the ability to sample from the joint distribution on $z$ and $w$.

For $N \geq 2$ let $c_{z,w}$ be the density defined by drawing pairs $(z_1, w_1), \ldots (z_n, w_n)$ from the population and then constructing the tuple $(i, z_1, \ldots, z_N, w_i)$ where $i$ is drawn uniformly from 1 to $N$.

# A Constrastive Bound

We train the codes to so that we can predict $i$ from $(z_1, \ldots, z_n, w)$.

$$\begin{aligned}
\Phi^* &= \operatorname*{argmin}_\Phi \; E_{(i,z_1\ldots,z_N,w)\sim c_{z,w}} \;\; - \ln P_\Phi(i|z_1, \ldots, z_n, w) \\
&= \operatorname*{argmin}_\Phi \; \mathcal{L}(\Phi) \\
P_\Phi(i|z_1, \ldots z_n, w) &= \operatorname*{softmax}_i \; s_\Phi(z_i, w) \quad \text{(required)} \\
I(z, w) &\geq \ln N - \mathcal{L}(\Phi)
\end{aligned}$$

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

## Forcing $z_x$ and $z_y$ to be Useful

In the objective

$$\Phi^* = \operatorname*{argmax}_{\Phi} \; I_{\mathrm{Pop},\Phi}(z_x, z_y) - \beta(H_{\mathrm{Pop},\Phi}(z_x) + H_{\mathrm{Pop},\Phi}(z_y))$$

the limitation on the entropy of $z_x$ and $z_y$ block the trivial solution of $z_x = x$ and $z_y = y$.

CPC applications have used an alternative.

# Forcing $z_x$ and $z_y$ to be Useful

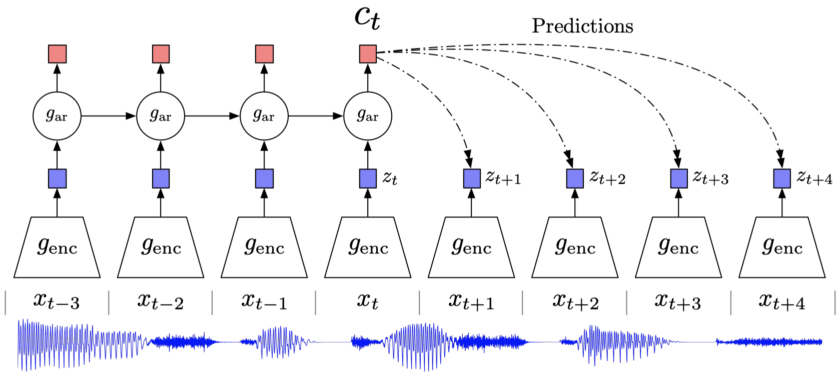
$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{(i,z_x^1 \ldots, z_x^n, z_y) \sim c_{z_x, z_y}} \; -\ln P_\Phi(i | z_x^1, \ldots, z_x^n, z_y)$$

$$P_\Phi(i | z_x^1, \ldots, z_x^n, z_y) = \operatorname*{softmax}_i \; z_y^\top z_x^i$$

Requiring that the score be a simple inner product blocks $z_x = x$ and $z_y = y$ and forces $z_x$ and $z_y$ to carry the information in a linearly extractible way.

# Contrastive Predictive Coding for Speech

wav2vec uses contrastive predictive coding (CPC)



# 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.

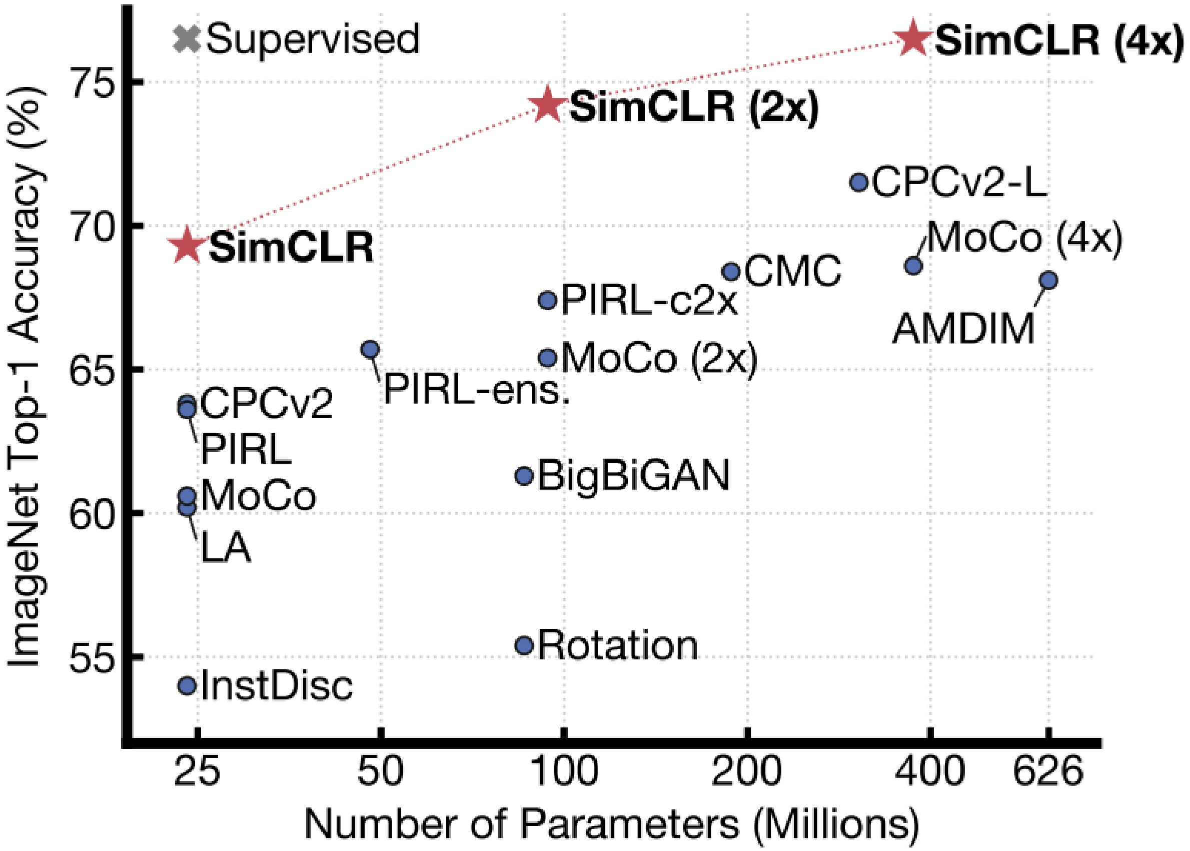
The training maximizes the contrastive lower bound on $I(x, y)$.

# 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.

# SimCLR



Chen et al. 2020

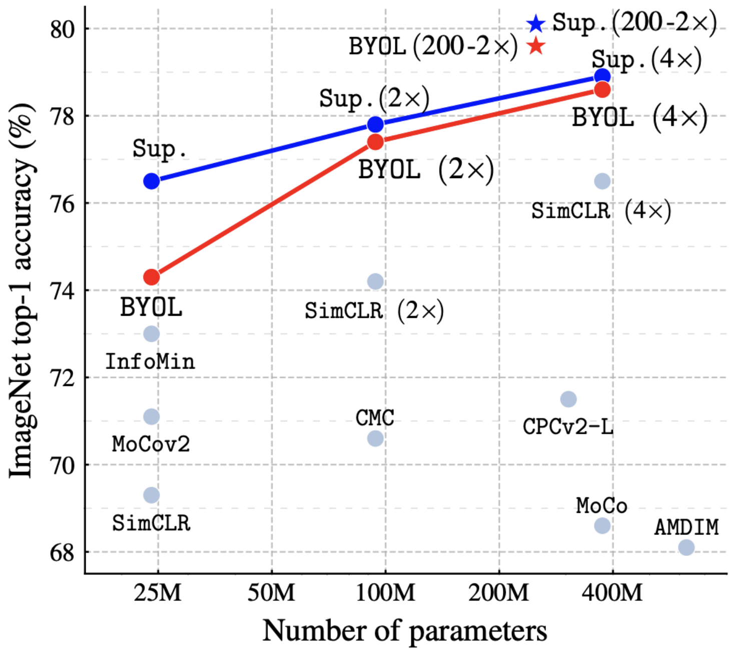
# BYOL and SwAV do not use CPC

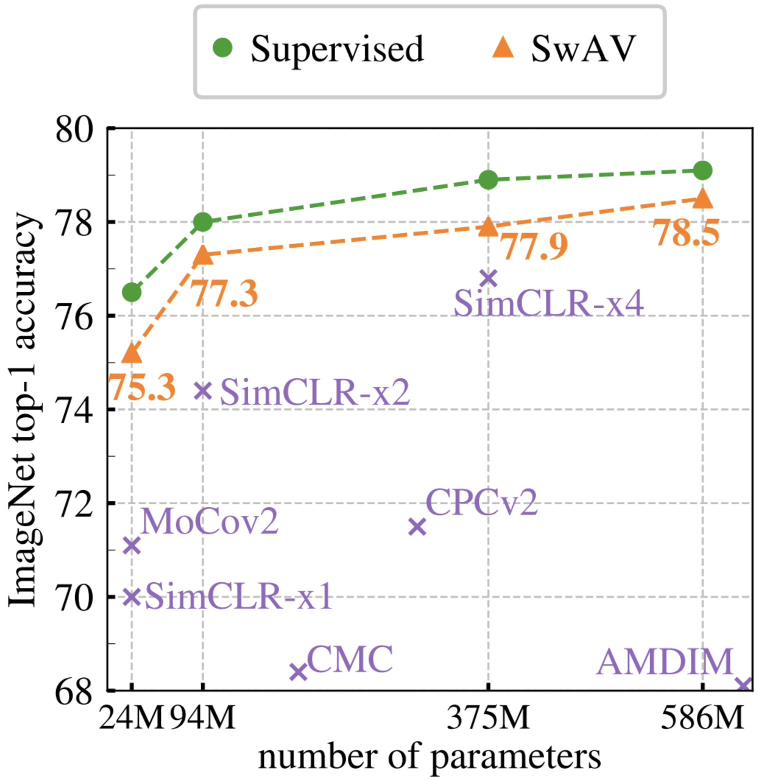
BYOL and SwAV train only on cross-entropy conditional probability

$$\mathcal{L}(\Psi) = -\ln P_\Psi(z_x|z_y)$$

This has a degenerate solution where the code for $z_x$ is constant independent of $x$. Complex hacks are used to cause this not to happen but performance is improved over the contrastive SimCLR.

# BOYL and SwAV





Grill et al. 2020, Caron et al. 2021

**END**