

TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Contrastive Predictive Coding

Maximizing Mutual Information

We consider the distribution on x , y , z_x and z_y defined by drawing $\langle x, y \rangle \sim \text{Pop}$, $z_x \sim P_\Phi(z_x|x)$ and $z_y \sim P_\Phi(z_y|y)$.

We are interested in optimizing $P_\Phi(z_x|x)$ and $P_\Phi(z_y|y)$ under the following objective.

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

Maximizing Mutual Information

$$\begin{aligned}\Phi^* &= \operatorname{argmax}_{\Phi} I_{\text{Pop},\Phi}(z_x, z_y) - \beta(\textcolor{red}{H}_{\text{Pop},\Phi}(z_x) + \textcolor{red}{H}_{\text{Pop},\Phi}(z_y)) \\ &\geq \operatorname{argmax}_{\Phi} I_{\text{Pop},\Phi}(z_x, z_y) - \beta(\hat{H}_{\text{Pop},\Phi}(z_x) + \hat{H}_{\text{Pop},\Phi}(z_y))\end{aligned}$$

$$\textcolor{red}{H}_{\text{Pop},\Phi}(z_x) = E_{\text{Pop},\Phi} - \ln P_{\textcolor{red}{Pop},\textcolor{red}{\Phi}}(z_x)$$

$$\hat{H}_{\text{Pop},\Phi}(z_x) = E_{\text{Pop},\Phi} - \ln P_{\textcolor{red}{\Phi}}(z_x)$$

$$\textcolor{red}{H}_{\text{Pop},\Phi}(z_x) \leq \hat{H}_{\text{Pop},\Phi}(z_x)$$

Maximizing Mutual Information

$$\Phi^* = \operatorname{argmax}_{\Phi} I_{\text{Pop}, \Phi}(z_x, z_y) - \beta(\hat{H}_{\Phi}(z_x) + \hat{H}_{\Phi}(z_y))$$

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 $(z_1, w_1), \dots, (z_n, w_n)$ from the population and then constructing the tuple (i, z_1, \dots, z_N, w) where i is drawn uniformly from 1 to N and $w = w_i$ is the value of w paired with z_i .

A Contrastive Lower Bound

$$\begin{aligned}\Phi^* &= \operatorname{argmin}_{\Phi} E_{(i, z_1, \dots, z_N, w) \sim c_{z, w}} - \ln P_{\Phi}(i | z_1, \dots, z_n, w) \\ &= \operatorname{argmin}_{\Phi} \mathcal{L}(\Phi)\end{aligned}$$

$$P_{\Phi}(i | z_1, \dots, z_n, w) = \operatorname{softmax}_i s_{\Phi}(z_i, w) \quad (\text{required})$$

$$I(z, w) \geq \ln N - \mathcal{L}(\Phi)$$

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

Forcing z_x and z_y to be Useful

In the objective

$$\Phi^* = \operatorname{argmax}_{\Phi} I_{\text{Pop}, \Phi}(z_x, z_y) - \beta(\textcolor{red}{H}_{\text{Pop}, \Phi}(z_x) + \textcolor{red}{H}_{\text{Pop}, \Phi}(z_y))$$

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.

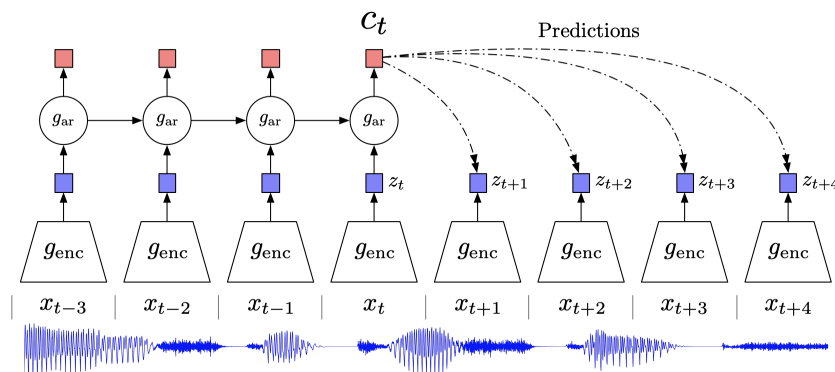
Forcing z_x and z_y to be Useful

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

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

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



van den Oord et al., 2018

We seek to train an auto-regressive g_{ar} and encoder g_{enc} by

$$g_{ar}^*, g_{enc}^* = \underset{g_{ar}, g_{enc}}{\operatorname{argmax}} E_t \sum_{k=1}^K I(c_t, z_{t+k})$$

The training maximizes the contrastive lower bound on $I(c_t, z_{t+k})$

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.

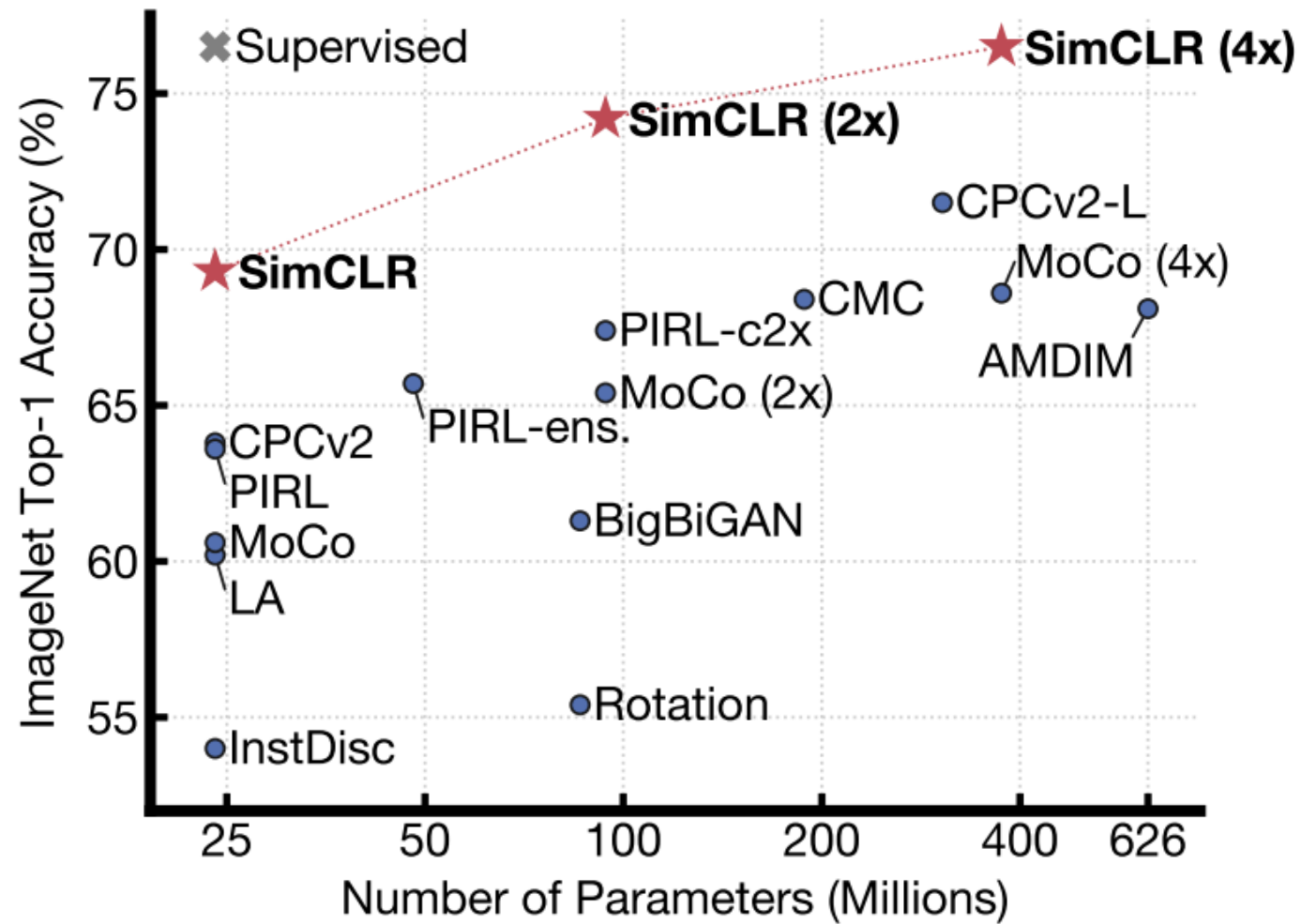
The training maximizes the contrastive lower bound on $I(x, y)$.

Contrastive Predictive Coding for Images

A resulting feature map z_Φ on images is extracted from this training.

The feature map z_Φ is tested by using a **linear** classifier for ImageNet based on these features.

SimCLR



END