TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Mutual Information Coding

Deep Co-Training

For a population on $\langle x, y \rangle$ and a "feature map" z_Φ we optimize Φ by

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}}\ I(z_\Phi(x), z_\Phi(y)) - \beta H(z_\Phi(x))$$

Here we can think of $z_\Phi(x)$ as what we remember about a past x to carry information about a future y while maintaining low memory requirements.

Deep Co-Training

$$
\begin{aligned}
\Phi^* &= \operatorname*{argmax}_{\Phi}\ (1-\beta)\hat{H}_\Phi(z_\Phi(x)) - \hat{H}_\Phi(z_\Phi(x)|z_\Phi(y)) \\
\hat{H}_\Phi(z_\Phi(x)) &= E_x\ -\ln\ P_{\Psi^*(\Phi)}(z_\Phi(x)) \\
\Psi^*(\Phi) &= \operatorname*{argmin}_{\Psi}\ E_x\ -\ln P_\Psi(z_\Phi(x)) \\
\hat{H}_\Phi(z_\Phi(x)|z_\Phi(y)) &= E_{x,y}\ -\ln P_\Phi(z_\Phi(x)|z_\Phi(y))
\end{aligned}
$$

Here we only model distributions on z. Unlike VAEs, there is no attempt to model distributions on x or y.

Mutual Information Objectives

CPC represents a fundamental shift in the self-supervised training objective.

GANs and VAEs are motivated by modeling $\mathrm{Pop}(y)$.

But in CPC there is no attempt to model $\mathrm{Pop}(y)$.

CPC can be viewed as training a feature map z_Φ so as to maximize the mutual information $I(z_\Phi(x), z_\Phi(y))$ while, at the same time, making $z_\Phi(x)$ useful for linear classifiers.

Relationship to Noise Contrastive Estimation

CPC is noise contrastive estimation (NCE) with "noise" generated by drawing y unrelated to x. By the NCE theorems, universality implies

$$P_{\Phi^*}(i|z_1, \ldots, z_N, z_x) = \operatorname*{softmax}_i \ \ln \frac{\mathrm{Pop}(z_i|z_x)}{\mathrm{Pop}(z_i)}$$

and also

$$\begin{aligned}
\mathcal{L}_{\mathrm{CPC}} &\geq \ln N - \frac{N-1}{N}(KL(\mathrm{Pop}(z_y|z_x), \mathrm{Pop}(z_y)) + KL(\mathrm{Pop}(z_y), \mathrm{Pop}(z_y|z_x))) \\
&= \ln N - \frac{N-1}{N}(I(z_x, z_y) + KL(\mathrm{Pop}(z_y), \mathrm{Pop}(z_y|z_x)))
\end{aligned}$$

Contrastive Predictive Coding (CPC)

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

For example x and y might be video frames separated by 10 seconds in a video.

For simplicity we will assume that the marginal distributions on x and y are the same — the probability that an image occurs as a first frame is the same as the probability that image occurs as a second frame.

In CPC we draw a pair $\langle x, y\rangle$ and minimize a discriminator loss for distinguishing $z_\Phi(y)$ from $z_\Phi(\tilde{y})$ for $\tilde{y} \sim \mathrm{Pop}(y)$. The discriminator gets to see x.

Contrastive Predictive Coding (CPC)

For $N \geq 2$ let $\tilde{P}^{(N)}$ be the distribution on tuples $\langle i, y_1, \ldots, y_N, x\rangle$ defined by the following process.

- draw a pair $\langle x, y\rangle$ from the population.
- drawn a sequence of $N-1$ "distractor values" from the marginal distribution $\mathrm{Pop}(y)$. These are unrelated to x.
- insert y at a random position among the distractors to get the sequence $y_1, \ldots, y_N$.
- return the tuple $\langle i, y_1, \ldots, y_N, x\rangle$ where i is the index of y among the distractors.

Contrastive Predictive Coding (CPC)

$$\Phi^* = \operatorname*{argmin}_{\Phi} \mathcal{L}_{\mathrm{CPC}}(\Phi)$$

$$\mathcal{L}_{\mathrm{CPC}}(\Phi) = E_{\langle i, y_1, \ldots, y_N, x\rangle \sim \tilde{P}^{(N)}} - \ln P_{\mathrm{CPC}}(i | z_\Phi(y_1), \ldots, z_\Phi(y_N), z_\Phi(x))$$

$$P_{\mathrm{CPC}}(i|z_1, \ldots, z_N, z_x) = \operatorname*{softmax}_{i} \; z_i^\top z_x$$

Contrastive Predictive Coding (CPC)

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \; \mathcal{L}_{\mathrm{CPC}}(\Phi)$$

$$P_\Phi(i|z_1, \ldots, z_n, z_x) = \underset{i}{\operatorname{softmax}} \; z_i^\top z_x$$

As N gets larger the contrastive discrimination task gets harder.

The task is also made difficult by the requirement that the score is defined to be an inner product of feature vectors.

Contrastive Predictive Coding (CPC)

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

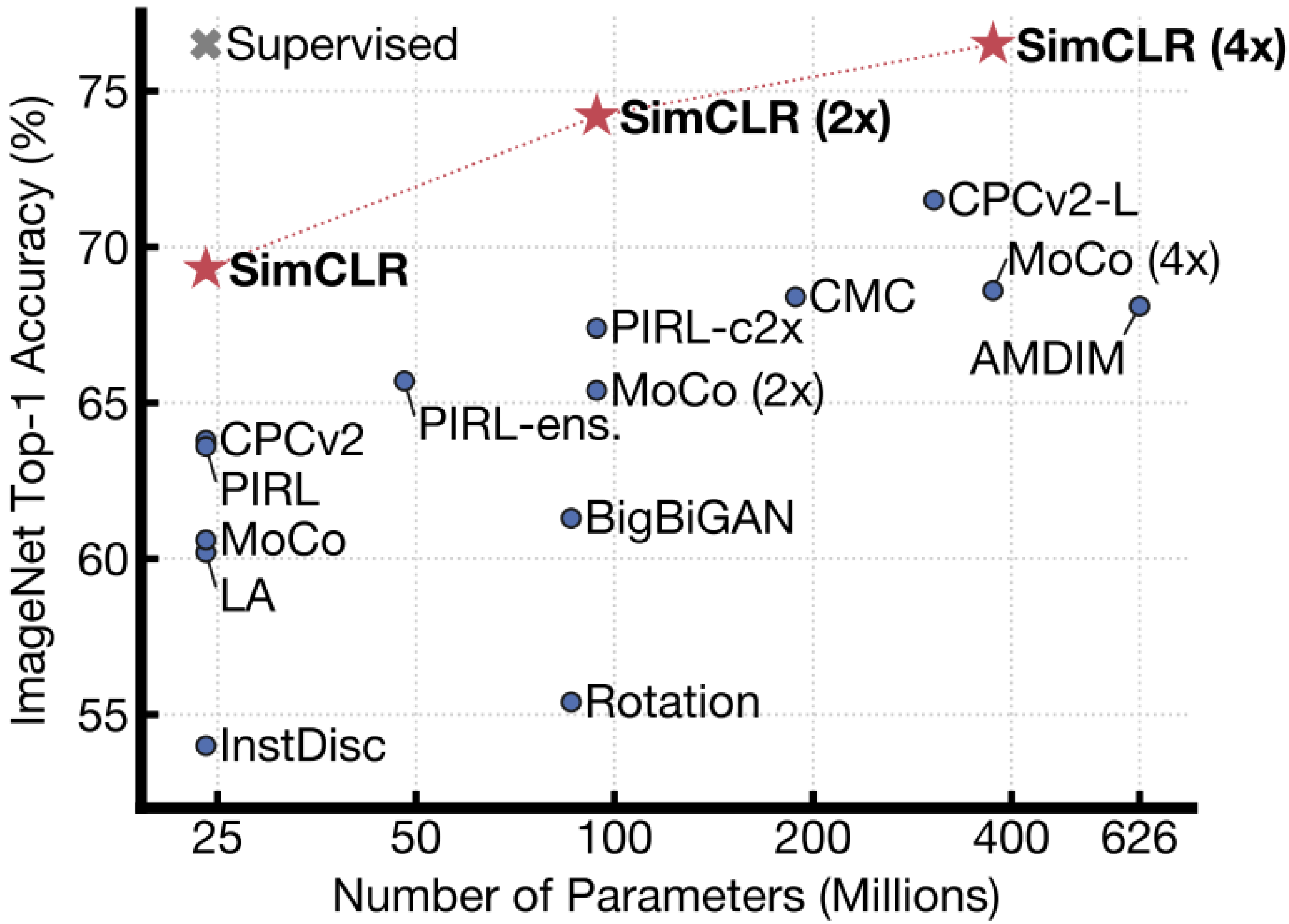
They use a distribution on pairs $\langle x, y \rangle$ defined by drawing an image s from ImageNet and then drawing x and y as random "augmentations" (modifications) of the image s — either a random translation, rotation, color jitter, masking, edge image, or a composition of these modifications.

Contrastive Predictive Coding (CPC)

The feature map z_Φ can then be applied to the images of ImageNet.

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

SimCLR


Supervised
SimCLR (4x)
SimCLR (2x)
SimCLR
CPCv2-L
MoCo (4x)
CMC
PIRL-c2x
AMDIM
MoCo (2x)
PIRL-ens.
CPCv2
PIRL
BigBiGAN
MoCo
LA
Rotation
InstDisc
75
70
65
60
55
25
50
100
200
400
626
ImageNet Top-1 Accuracy (%)
Number of Parameters (Millions)

END