

TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Mutual Information Coding

Mutual Information Objectives

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

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

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

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

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

and also

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

Deep Co-Training

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

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

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

$$\Phi^* = \operatorname{argmax}_{\Phi} (1 - \beta) \hat{H}_{\Phi}(z_{\Phi}(x)) - \hat{H}_{\Phi}(z_{\Phi}(x)|z_{\Phi}(y))$$

$$\hat{H}_{\Phi}(z_{\Phi}(x)) = E_x - \ln P_{\Psi^*(\Phi)}(z_{\Phi}(x))$$

$$\Psi^*(\Phi) = \operatorname{argmin}_{\Psi} E_x - \ln P_{\Psi}(z_{\Phi}(x))$$

$$\hat{H}_{\Phi}(z_{\Phi}(x)|z_{\Phi}(y)) = E_{x,y} - \ln P_{\Phi}(z_{\Phi}(x)|z_{\Phi}(y))$$

Here, as in CPC, we only model distributions on z . There is no attempt to model distributions on x or y .

END