Information Theoretic Co-Training

David McAllester
TTI-Chicago

February 22, 2018

Abstract

This paper introduces an information theoretic co-training objective for unsupervised learning. We consider the problem of predicting the future. Rather than predict future sensations (image pixels or sound waves) we predict "hypotheses" to be confirmed by future sensations. More formally, we assume a population distribution on pairs (x, y) where we can think of x as a past sensation and y as a future sensation. We train both a predictor model $P_\Phi(z|x)$ and a confirmation model $P_\Psi(z|y)$ where we view z as hypotheses (when predicted) or facts (when confirmed). For a population distribution on pairs (x, y) we focus on the problem of measuring the mutual information between x and y. By the data processing inequality this mutual information is at least as large as the mutual information between x and z under the distribution on triples (x, z, y) defined by the confirmation model $P_\Psi(z|y)$. The information theoretic training objective for $P_\Phi(z|x)$ and $P_\Psi(z|y)$ can be viewed as a form of co-training where we want the prediction from x to match the confirmation from y.

1 Intuition and Formulation

We Consider the problem of predicting the future from the past. Intuitively we are not interested in predicting raw future sense data such as image pixels. Rather we are interested in predicting facts about the future as will be inferred from future sensations. Here we consider the joint problem of (1) learning to convert sensation to facts, and (2) learning to predict future facts. Information theoretic co-training aims to measure the mutual information between past sensation and the future sensation by demonstrating the ability to predict future facts.

We formulate information theoretic co-training by letting $\mathcal{X}$ be a space of possible past sensations, $\mathcal{Y}$ be a space of possible future sensations, and $\mathcal{Z}$ be a space of facts. We assume a population distribution $\mathcal{P}$ on sensation pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We assume models $P_\Phi(z|x)$ and $P_\Psi(z|x)$ to predict future facts from past and future sensations respectively. Here Φ and Ψ are parameter vectors and we assume that the probabilities are differentiable in the parameters.

We assume that z is most accurately estimated from y. We then define a distribution on triples (x, z, y) where (x, y) is drawn from the population distribution $\mathcal{P}$ and z is drawn from $P_\Psi(z|y)$. We note that by the data processing inequality we have

$$I(x, y) \geq I_\Psi(x, z) = H_\Psi(z) - H_\Psi(z|x).$$

Here $H_\Psi(z)$ and $H_\Psi(z|x)$ are determined by the distribution on triples which is itself determined by Ψ.

The information theoretic co-training object takes into account the difficulty of empirically measuring the entropies $H_\Psi(z)$ and $H_\Psi(z|x)$. In the phonetics experiment we have that z has only 64 possible values which implies that the entropy $H_\Psi(z)$ is at most six bits. In this case $H_\Psi(z)$ can be approximated directly by the empirical marginal over z in a large minibatch. The entropy $H_\Psi(z|x)$ cannot in general be measured directly. We assume that we can sample (x, y) from the population, and sample z from $P_\Psi(z|y)$, but have no way computing $P_\Psi(z|x)$. However, following standard practice we can upper bound the entropy $H_\Psi(z|x)$ by the cross-entropy $H^+_{\Psi,\Phi}(z|x)$ where the model probability $P_\Phi(z|x)$ is computable.

$$\begin{aligned}
I_\Psi(x, y) &\geq H_\Psi(z) - H^+_{\Psi,\Phi}(z|x) \\
H^+_{\Psi,\Phi}(z|x) &= E_{(x,y)\sim\mathcal{P},\; z\sim P_\Psi(z|y)} \; -\ln P_\Phi(z|x) \\
&= H_\Psi(z|x) + KL(P_\Psi(z|y), P_\Phi(z|x)) \\
&\geq H_\Psi(z|x)
\end{aligned}$$

A first information theoretic co-training objective is then defined by

$$\Psi^*\Phi^* = \operatorname*{argmax}_{\Psi,\Phi} \; H_\Psi(z) - H^+_{\Psi,\Phi}(z|x). \tag{1}$$

It is perhaps useful to rewrite (1) as

$$\begin{aligned}
\Psi^* &= \operatorname*{argmax}_{\Psi} H_\Psi(z) - \left(\min_{\Phi} \; H^+_{\Psi,\Phi}(z|x)\right) \\
\Phi^* &= \operatorname*{argmin}_{\Phi} H^+_{\Psi^*,\Phi}(z|x) \\
&= \operatorname*{argmin}_{\Phi} E_{(x,y)\sim\mathcal{P},\; z\sim P_{\Psi^*}(z|y)} -\ln P_\Phi(z|x).
\end{aligned} \tag{2}$$

It should be noted that the objective (2) is the standard objective for training on labeled data. In (2) z replaces y as a label for x. The term $H_\Psi(z)$ in objective (1) encourages Ψ to extract as much factual information from the future sensation as possible while still making the extracted factual information predictable from the past. Here Φ and Ψ cooperate to find agreement on a "language" (a semantics for symbols) grounded in sensation.

If z is allowed to be a structured object, such as a sequence of symbols, then $H_\Psi(z)$ becomes difficult to measure. However, again following stadanrd practice, we can bound $H_\Psi(z)$ by a cross-entropy $H^+_{\Psi,\Theta}(z)$. We then have the information theoretic co-training objective

$$\begin{aligned} \Psi^* &= \operatorname*{argmax}_{\Psi} \left(\min_{\Theta} H^+_{\Psi,\Theta}(z)\right) - \left(\min_{\Phi}\ H^+_{\Psi,\Phi}(z|x)\right) \\ H^+_{\Psi,\Theta}(z) &= E_{(x,y)\sim\mathcal{P}}\ E_{z\sim P_\Psi(z|y)}\ -\ln P_\Theta(z). \end{aligned} \tag{3}$$

Here Θ is adversarial to Ψ and Φ. Even in the case where z is a structured object such as a string, it may be useful in practice to bound the amount of information in z by, for example, bounding the size of the alphabet and the length of the string. This will make $H_\Psi(z)$ and $H_\Psi(z|x)$ smaller which should improve the numerical stability of the measured difference $H^+_{\Psi,\Theta^*}(z) - H^+_{\Psi,\Phi^*}(z|x)$.

2 Related Learning Models

Co-Training. Information theoretic co-training is closely related to classical co-training (Blum & Mitchell (1998); Dasgupta et al. (2002)). Classical co-training assumes the same three spaces $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ but takes the population $\mathcal{P}$ to be a distribution on triples (x, z, y) where z is not observed in the training data. The goal is to learn rules for predicting z by training on the pairs (x, y). For this to be possible we need additional assumptions such as that x and y are independent given z (in the population) and that $H_\mathcal{P}(z)$ is large. In information theoretic co-training, on the other hand, the population is assumed to be a distribution on (x, y) only and the goal is to measure the mutual information between x and y.

Although the assumptions and theoretical analyses are different, the learning algorithms of information theoretic co-training and classical co-training are very similar. The goal in classical co-training is to find hard (non-stochastic) classifiers $f : \mathcal{X} \to \mathcal{Z}$ and $g : \mathcal{Y} \to \mathcal{Z}$ so as to maximize the probability over the draw of (x, y) that $f(x) = h(y)$ and, at the same time, to require that the values of $f(x)$ and $g(x)$ are diverse. Information theoretic co-training makes the classifiers soft and makes the training objective information theoretic.

The Information Bottleneck. Like information theoretic co-training, Tishby's information bottleneck Tishby et al. (1999) assumes the spaces $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ and assumes a population distribution on the pairs (x, y). The objective is to train a model $P_\Psi(z|y)$ defining a distribution on triples (x, z, y) using the training objective

$$\Psi^* = \operatorname*{argmax}_{\Psi}\ I_\Psi(z, x) - \beta I_\Psi(z, y). \tag{4}$$

In information theoretic co-training the second term is dropped and we retain only $I_\Psi(z, x)$. One might immediately object that the choice of $z = y$ maximizes $I_\Psi(z, x)$ so the objective is trivial if we drop the second term. But the goal of information theoretic co-training is not to maximize mutual information but rather to measure it. Note that setting $z = y$ eliminates Ψ from the information theoretic co-training objective and we are left with setting Φ so as to minimize $H^+_{\mathcal{P},\Phi}(y|x)$. This is the standard training objective for labeled data where we treat y as a label. This can also be viewed as conditional density estimation. Conditional density estimation must be addressed to measure mutual information. Setting $z = y$ is expected to yield a poor measurement of mutual information for two somewhat related reasons. First, the probabilistic modeling of raw sense data is difficult. Second $H^+_{\mathcal{P},\Theta}(y)$ and $H^+_{\mathcal{P},\Phi}(y|x)$ are both typically much larger than $H^+_{\Psi,\Theta}(z)$ and $H^+_{\Psi,\Phi}(z|x)$. So taking $z = y$ exposes one to numerical instability in taking the difference $H^+_{\mathcal{P},\Theta}(y) - H^+_{\mathcal{P},\Phi}(y|x)$.

Density Estimation. Many approaches to unsupervised learning can be viewed as some form of density estimation. Density estimation is the problem of modeling a probability distribution given the ability to draw samples. A paradigmatic example is language modeling. In general we assume a population distribution $\mathcal{P}$ over some set $\mathcal{Y}$ and a model $P_\Psi(y)$ assigning a probability to each $y \in \mathcal{Y}$. The density estimation objective is

$$\begin{aligned} \Psi^* &= \operatorname*{argmin}_{\Psi} \; H^+_{\mathcal{P},\Psi}(y) \\ H^+_{\mathcal{P},\Psi}(y) &= E_{y\sim\mathcal{P}} \; -\ln P_\Psi(y) \end{aligned} \tag{5}$$

The cross-entropy $H^+_{\mathcal{P},\Psi}(y)$ is an upper bound on the unknown, and typically unknowable, true entropy $H_{\mathcal{P}}(y)$.

Expectation maximization (EM) (Dempster et al. (1977)) and variational autoecoders (VAEs) (Kingma & Welling (2014)) optimize (5) for the case where $P_\Psi(y)$ is a marginal distribution over a latent variable z.

$$P_\Psi(y) = \sum_z \; P_\Psi(z, y) \tag{6}$$

Here $P_\Psi(z, y)$ is typically a generative model where y is generatively derived from z.

Data compression algorithms also implicity optimize (5). By Shannon's source coding theorem the most efficient code for instances drawn from a given population uses a number of bits equal to the entropy of the population distribution. The training objective (5) can be interpreted as optimizing the compressed bits per sample when drawing from the population but coding for the model.

Information theoretic co-training as defined by (1) and (3) differs from density estimation as defined by (5) in that information theoretic co-training uses only probability models for the "facts" z — in information theoretic co-training there is no attempt to model distributions on the sensations.

GANs. Generative adversarial networks (GANs) (Schmidhuber (1992); Goodfellow et al. (2014)) are similar to variational autoencoders in that they define a generative model $P_\Psi(z, y)$ where y is generated from z and where we are interested in the marginal distribution (6). However, in GANs there is no attempt to optimize, or even measure, a cross-entropy (5). Instead one define a distribution Q_Ψ on pairs (y, ℓ) by drawing y with equal probability either from the population distribution $\mathcal{P}$ or the model distribution P_Ψ and setting $\ell = 1$ is y is drawn from $\mathcal{P}$ and $\ell = -1$ if y is drawn from P_Ψ. A discriminator model

$P_\Phi(\ell|y)$ must predict which distribution y was drawn from. The GAN objective is

$$
\begin{aligned}
\Psi^* &= \underset{\Psi}{\mathrm{argmax}} \; \min_{\Theta} \; H^+_{\Psi,\Theta}(\ell|y) \qquad (7) \\
H^+_{\Psi,\Theta}(\ell|y) &= E_{(y,\ell)\sim Q_\Psi} \; -\ln P_\Theta(\ell|y)
\end{aligned}
$$

In (7) the generator Ψ is trying to generate values such that discriminator Θ cannot distinguish values generated from P_Ψ from values drawn from $\mathcal{P}$. InfoGANS (Chen et al. (2016)) add a term to the GAN objective to increase the mutual information between certain components of the latent variable z and the generated variable y. This encourages the model to use independent components of the information in z.

A major issue with GANs is the lack of an objective measure of performance. The ability to fool a particular discriminator architecture does not imply low cross-entropy as defined by (5). It is quite plausible that large modes of the population density are omitted from the generator distribution (the problem of mode dropping). In contrast, information theoretic co-training provides a quantitative performance measure.

3 Summary

Information theory already plays a central role in the training objectives typically used in deep learning. Information theoretic co-training introduces a novel information theoretic objective for unsupervised learning in which one can avoid any attempt to measure the entropy, or conditional entropy, of raw sense data. Information theoretic co-training can also be viewed as a way of measuring mutual information by developing a "language" for carrying that information where the entropy of the facts stated in that language is small compared to the entropy of raw sense data.

References

A Blum and T Mitchell. Combining labeled and unlabeled data with co-training. In *COLT-98*, 1998.

Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29*, 2016.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Gated feedback recurrent neural networks. arXive1502.02367, 2015.

Sanjoy Dasgupta, Michael Littman, and David McAllester. Pac generalization bounds for co-training. In *NIPS*, 2002.

A. P. Dempster, N. M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B.*, 1977.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. Timit: Acoustic-phonetic continuous speech corpus ldc93s1, 1993.

Ian Goodfellow, Jean Pouget-Abadie, Hehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. arXiv:1406.2661, 2014.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

D.P. Kingma and M. Welling. Auto-encoding variational bayes. In *The International Conference on Learning Representations (ICLR)*, 2014.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 1992.

Naftali Tishby, Fernando Pereira, and William Bialek. The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, 1999.