TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Mutual Information Co-Training

Language is Situated

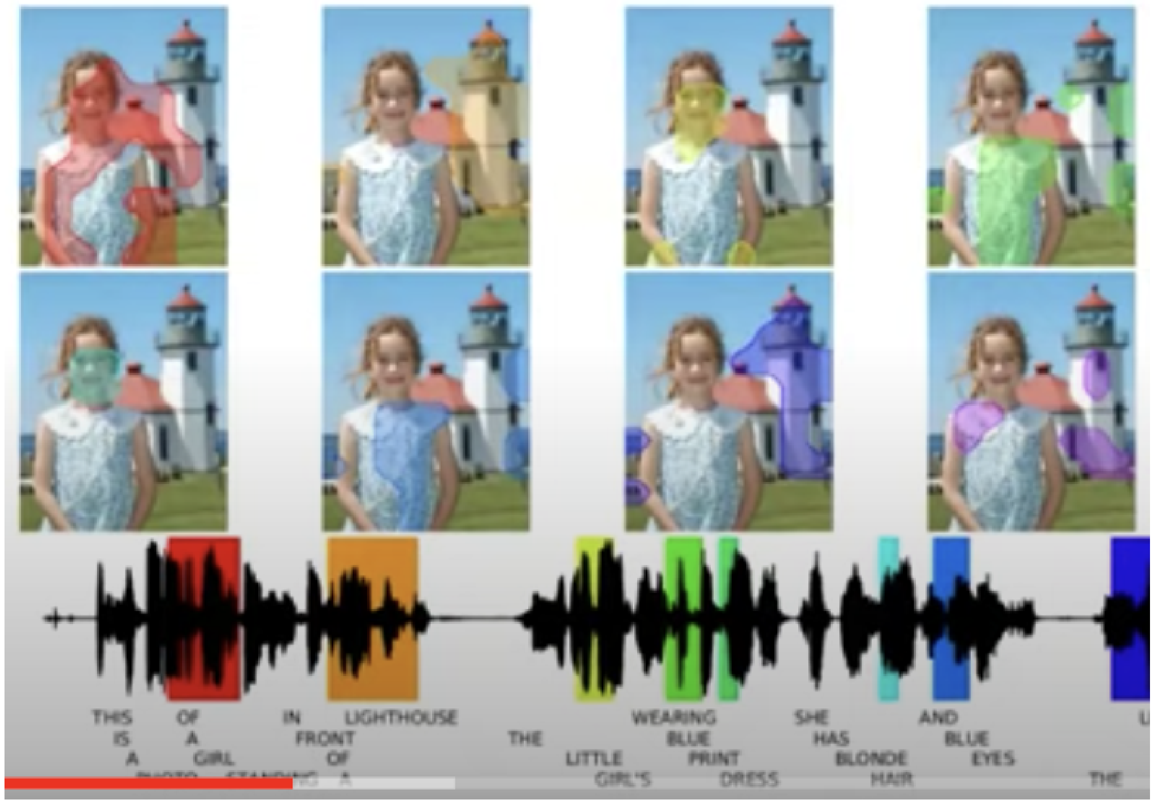
When a child hears an utterance the language is often referring to things present in the immediate physical situation.

We can model this with a collection of images paired with recordings of descriptions of the image.

We can then try to automatically find correspondences between parts of the sound and parts of the image.

Harwath et al., Unsupervised Learning of Spoken Language with Visual Context, NeurIPS 2016.

Language is Situated



Harwath

Mutual Information Co-Training

Here I will formulate mutual information co-training.

This gives an abstract objective for multi-modal learning of latent variables.

This objective was not followed (explicitly) in Harwath et al.

Mutual Information Co-Training

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 Co-Training

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 $P_\Phi(z_y|y)$.

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

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 Co-Training

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

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 Co-Training

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

$$= \operatorname*{argmax}_{\Phi} \left\{ \begin{array}{l} \;\frac{1}{2}(H_{\mathrm{Pop},\Phi}(z_x) - H_{\mathrm{Pop},\Phi}(z_x|z_y)) \\ + \frac{1}{2}(H_{\mathrm{Pop},\Phi}(z_y) - H_{\mathrm{Pop},\Phi}(z_y|z_x)) \\ - \beta(H_{\mathrm{Pop},\Phi}(z_x) + H_{\mathrm{Pop},\Phi}(z_y)) \end{array} \right.$$

$$= \operatorname*{argmax}_{\Phi} \left\{ \begin{array}{l} (1-2\beta)(H_{\mathrm{Pop},\Phi}(z_x) + H_{\mathrm{Pop},\Phi}(z_y)) \\ - H_{\mathrm{Pop},\Phi}(z_x|z_y) - H_{\mathrm{Pop},\Phi}(z_y|z_x) \end{array} \right.$$

Mutual Information Co-Training

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}} \left\{ \begin{array}{l} (1-2\beta)(H_{\mathrm{Pop},\Phi}(z_x) + H_{\mathrm{Pop},\Phi}(z_y)) \\ \\ -\ H_{\mathrm{Pop},\Phi}(z_x|z_y) - H_{\mathrm{Pop},\Phi}(z_y|z_x) \end{array} \right.$$

The above entropies and conditional entropies are defined in terms of the population distribution Pop and the models $P_\phi(z_x|x)$ and $P_\Phi(z_y|y)$.

Since the population distribution is unknown, we cannot optimize this directly.

Mutual Information Co-Training

$$\Phi^* = \operatorname*{argmax}_{\Phi} \left\{ \begin{array}{l} (1-2\beta)(H_{\mathrm{Pop},\Phi}(z_x) + H_{\mathrm{Pop},\Phi}(z_y)) \\ \\ -\ H_{\mathrm{Pop},\Phi}(z_x|z_y) - H_{\mathrm{Pop},\Phi}(z_y|z_x) \end{array} \right.$$

We would like to maximize a lower bound on this expression. Entropies are upper bounded by cross entropies.

$$\Phi^* = \operatorname*{argmax}_{\Phi} \left\{ \begin{array}{l} (1-2\beta)(H_{\mathrm{Pop},\Phi}(z_x) + H_{\mathrm{Pop},\Phi}(z_y)) \\ \\ -\ \hat{H}_{\Phi}(z_x|z_y) - \hat{H}_{\Phi}(z_y|z_x) \end{array} \right.$$

$$\hat{H}_{\Phi}(z_x|z_y) = E_{(x,y)\sim\mathrm{Pop},\ z_x\sim P_\Phi(z_x|x),\ z_y\sim P_\Phi(z_y|y)}\ \ -\ln P_\Phi(z_x|z_y)$$

Mutual Information Co-Training

To do the optimization we can replace all entropies with cross-entropies.

$$\Phi^* = \operatorname*{argmax}_{\Phi} \left\{ \begin{array}{l} (1-2\beta)(\hat{H}_\Phi(z_x) + \hat{H}_\Phi(z_y)) \\ \\ -\hat{H}_\Phi(z_x|z_y) - \hat{H}_\Phi(z_y|z_x) \end{array} \right.$$

While this is no longer a lower bound on the desired mutual information objective, it might still be useful in practice.

This is called a **difference of entropies** objective.

END