Mutual Information Maximization for Simple and Accurate Part-Of-Speech Induction

Karl Stratos

Toyota Technological Institute at Chicago

Mutual Information for NLP and Speech

- Maximizing mutual information is a hugely successful approach to unsupervised learning.
  - Brown clustering (Brown et al., 1992)
  - Estimation of HMMs for speech recognition (Bahl et al., 1986)
  - The information bottleneck method (Tishby et al., 2000)
  - Deep representation learning: MINE (Belghazi et al., 2018), CPC (van den Oord et al., 2018), DIM (Hjelm et al., 2019)
- Mutual information is difficult to work with.
  - Theoretical problem: measurement is intractable (McAllester and Stratos, 2018).
  - Practical problem: optimization is difficult.
  - Past methods rely on problem-specific assumptions (e.g., the Brown clustering algorithm).

This Work

- Neural parameterizations of the mutual information objective
  1. A generalization of Brown clustering
  2. A variational approximation (McAllester, 2017)
- State-of-the-art results on part-of-speech induction
  - Simple architecture: no feature engineering or expensive structured computation

Outline

Maximal Mutual Information (MMI) Predictive Coding

Variational Approximation

Experiments

Conventional Approach to Representation Learning

- Unknown joint distribution p_{XY} over random variables (X, Y)

$$X = \text{“past” signal}$$
$$Y = \text{“future” signal}$$

- We draw a sample (x, y) by masking a part of observation

$$x = \left(\texttt{had these } \underbrace{\texttt{?}} \texttt{ in my}\right) \qquad y = \texttt{keys}$$

- Conventional approach: **conditional density estimation**
  - Given $(x_1, y_1) \dots (x_N, y_N) \sim p_{XY}$, estimate $p_{Y|X}$.
  - Examples: word2vec, ELMo, BERT, GPT/GPT-2
  - Often **uninterpretable** (continuous vectors), **wasteful** (noise in raw signals)

Desiderata

Goal: learn interpretable representations without modeling noise.

1. Explicitly define appropriate **discrete** encodings

$$Z' = \text{discrete encoding of "past" signal } X$$
$$Z = \text{discrete encoding of "future" signal } Y$$

2. Directly estimate **distributions over Z' and Z**
   - Never estimate distributions over raw signals!

Mutual Information Between Random Variables

Strength of statistical dependencies between (X, Y)

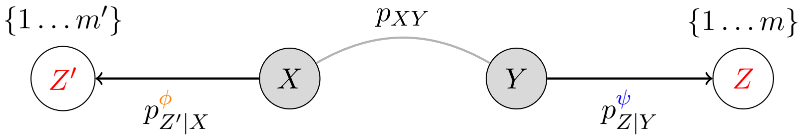
$$I(X, Y) = \sum_{x,y} p_{XY}(x, y) \log \frac{p_{XY}(x, y)}{p_X(x) p_Y(y)} \geq 0$$

- $I(X, Y) = 0$ iff (X, Y) are independent
- Largest when one variable determines the other

Data processing inequality: for any $p^{\phi}_{Z'|X}$ and $p^{\psi}_{Z|Y}$

$$I(X, Y) \geq I_{\phi,\psi}(Z', Z)$$

Maximal Mutual Information (MMI) Predictive Coding



Data: N samples $(x_1, y_1) \dots (x_N, y_N) \sim p_{XY}$

Objective: find parameters ϕ, ψ that maximize the empirical mutual information between discrete encodings

$$\max_{\phi,\psi} \; \underbrace{\frac{1}{N} \sum_{i=1}^{N} \sum_{z',z} p^{\phi}_{Z'|X}(z'|x_i) p^{\psi}_{Z|Y}(z|y_i) \log \frac{N \sum_{i=1}^{N} p^{\phi}_{Z'|X}(z'|x_i) p^{\psi}_{Z|Y}(z|y_i)}{\sum_{i=1}^{N} p^{\phi}_{Z'|X}(z'|x_i) \sum_{i=1}^{N} p^{\psi}_{Z|Y}(z|y_i)}}_{\text{estimate of a lower bound on } I(X,Y)}$$

Outline

Maximal Mutual Information (MMI) Predictive Coding

Variational Approximation

Experiments

Problem with Stochastic Optimization

- The previous objective is not amenable to SGD
  - Nonlinear function of N samples
  - SGD is ineffective
- Empirical success with a simpler lower bound on mutual information

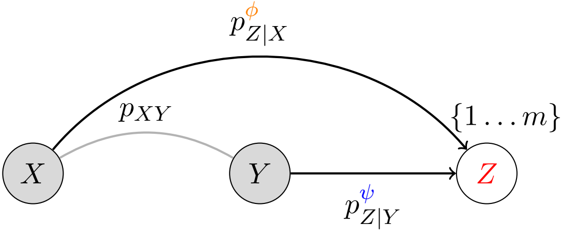
Variational Lower Bound on Mutual Information

$$I(X,Y) = \sum_{x,y} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}$$

Data processing inequality: for any $p^{\psi}_{Z|Y}$

$$\begin{aligned} I(X,Y) &\geq I_{\psi}(X,Z) \\ &= H_{\psi}(Z) - H_{\psi}(Z|X) \\ &\geq H_{\psi}(Z) - H^{+}_{\psi,\phi}(Z|X) \quad \forall p^{\phi}_{Z|X} \end{aligned}$$

Information Theoretic Co-Training (McAllester, 2017)



$$\max_{\psi,\phi} \underbrace{\frac{1}{N}\sum_{i=1}^{N}\sum_{z} p^{\psi}_{Z|Y}(z|y_i)\log\frac{N p^{\phi}_{Z|X}(z|x_i)}{\sum_{j=1}^{N} p^{\psi}_{Z|Y}(z|y_j)}}_{\text{estimate of a lower bound on } I(X,Y)}$$

Outline

Maximal Mutual Information (MMI) Predictive Coding

Variational Approximation

Experiments

Evaluation: Part-Of-Speech (POS) Induction

- **Task**: given unlabeled text, infer the POS tags of words

DET	NOUN	VERB	ADJ
a	cat	run	hot
an	dog	walk	cold
this	car	do	new
that	book	eat	long
⋮	⋮	⋮	⋮

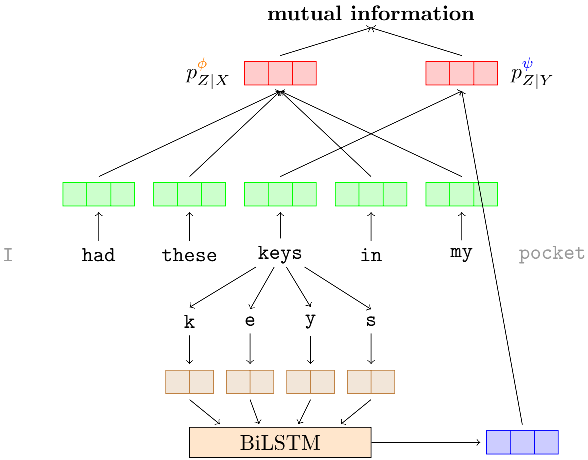
- **Evaluation metric**: many-to-one accuracy
  - Number of labels m: always fixed to true number of POS tags

- **Baselines**
  - **HMM**: standard HMM trained with EM (Baum-Welch)
  - **Brown**: Brown clusters (Brown et al., 1992)
  - **A-HMM**: anchor HMM (Stratos et al., 2016)
  - **F-HMM**: featurized HMM (Berg-Kirkpatrick et al., 2010)
  - **CRF-AUTO**: CRF autoencoder (Ammar et al., 2014)

Architecture

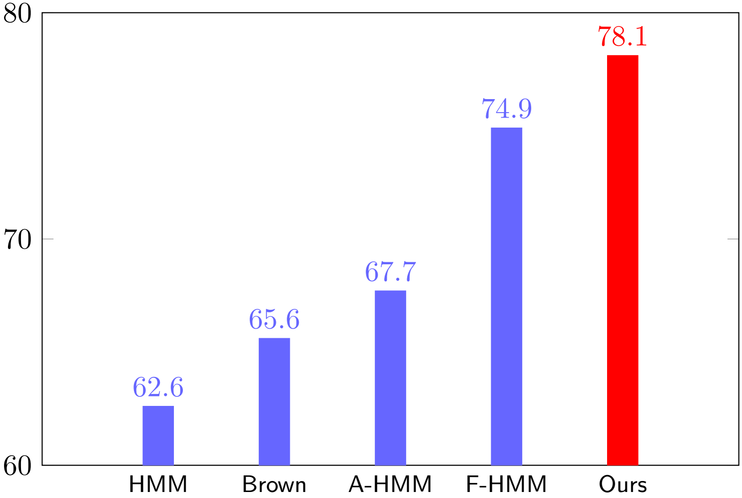
$x = (\texttt{had these}, \texttt{in my})$

$y = \texttt{keys}$



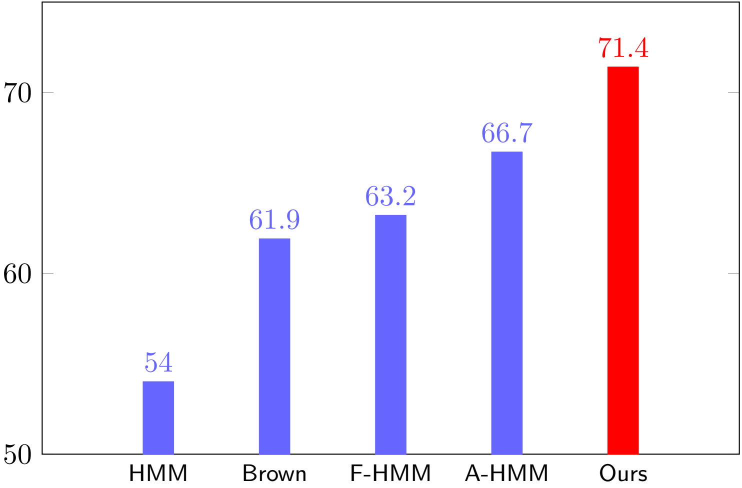
Result on Penn Treebank ($m = 45$ Tags)

Averaged over 10 random restarts



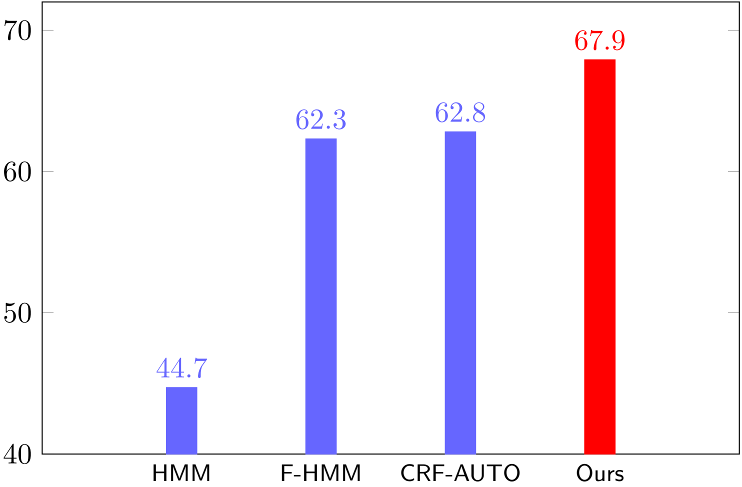
Result on Universal Treebank ($m = 12$ Tags)

Tuned on Penn Treebank, averaged over 10 languages



Comparison with CRF Autoencoders

Same setup: 8 languages from CoNLL with 12 tags (Ammar et al., 2014), model tuned on Penn Treebank



Summary

- We identified an effective neural parameterization of the mutual information objective for MMI predictive coding.
  - Excellent POS induction results with a very simple architecture
- Future work includes
  - Structured label induction
  - Extrinsic evaluation of the induced representations