Learning Discrete Structured Representations by Adversarially Maximizing Mutual Information

Karl Stratos[1]

Joint work with Sam Wiseman[2]

[1]Rutgers University

[2]Toyota Technological Institute at Chicago

Maximal Mutual Information (MMI)

- Maximizing mutual information is an effective objective for unsupervised representation learning.
  - Brown clustering (Brown et al., 1992)
  - Information bottleneck (Tishby et al., 2000)
  - Neural extensions: VIB (Alemi et al., 2017), MINE (Belghazi et al., 2018), CPC (van den Oord et al., 2018), DIM (Hjelm et al., 2019), ...
- Success so far limited to
  - **Continuous** representations
  - Discrete representations, but with **small mutual information** (McAllester, 2017; Stratos, 2018)
  - **Lower bounds** on mutual information that suffer from fundamental statistical limitations (McAllester and Stratos, 2018)

This Work

- We present **AMMI**: an adversarial approach to MMI
  - A new objective for learning **discrete structured** representations
  - Allows for **large mutual information**
  - The objective is **adversarial**, neither an upper bound nor a lower bound on mutual information ($\approx$ GANs).
- A concrete model: **structured bit string encoder**
  - State-of-the-art performance on document hashing

Outline

MMI

AMMI

- Structured bit string encoder

Experiments on document hashing

Conventional Approach to Representation Learning

Unknown joint distribution $\mathbf{pop}_{XY}$ over random variables (X, Y)

$$X = \text{"past" signal}$$
$$Y = \text{"future" signal}$$

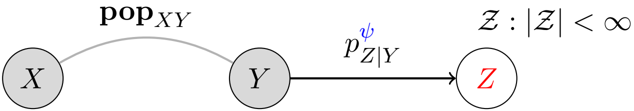
Representation learning by **density estimation**: learn ψ such that

$$\mathbf{pop}_{Y|X} \approx p^{\psi}_{Y|X} \qquad \text{("self-supervised", e.g., BERT)}$$
$$\mathbf{pop}_{Y} \approx p^{\psi}_{Y} \qquad \text{(autoencoding, e.g., VAEs)}$$

Limitations

- **Wasteful**: the model fits noise
- **Uninterpretable**: continuous representations implied by ψ

MMI Predictive Coding



$$\max_{\psi} \; I_{\psi}(X, Z)$$

- **No decoder**: never estimates density over raw signals
- Representation explicitly in a **finite discrete** codebook $\mathcal{Z}$

The log bottleneck problem. We are limited by

$$\begin{aligned} I_{\psi}(X, Z) &= H_{\psi}(Z) - H_{\psi}(Z|X) \\ &\leq H_{\psi}(Z) \\ &\leq \log |\mathcal{Z}| \end{aligned}$$

Game Plan

- We will make $\mathcal{Z}$ exponentially large (e.g., $\{0,1\}^m$).
- For such $\mathcal{Z}$, we will derive a tractable objective based on **adversarial** optimization.

Outline

MMI

AMMI

- Structured bit string encoder

Experiments on document hashing

Mutual Information as a Difference of Entropies

Objective. Find parameters of encoder $p^{\psi}_{Z|Y}$ by

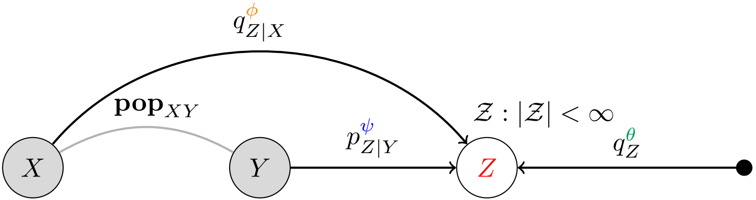
$$\max_{\psi} \; H_{\psi}(Z) - H_{\psi}(Z|X)$$

While directly estimating entropy is difficult, effective **upper bounds** are available:

$$H_{\psi}(Z) = \min_{\theta} \; H^{+}_{\psi,\theta}(Z)$$
$$H_{\psi}(Z|X) = \min_{\phi} \; H^{+}_{\psi,\phi}(Z|X)$$

where we introduce **variational models** q^{θ}_{Z} estimating the marginal of Z, $q^{\phi}_{Z|X}$ estimating the marginal of Z given X

Adversarial MMI (AMMI)



Models. Encoder $p^{\psi}_{Z|Y}$, variational q^{θ}_{Z}, $q^{\phi}_{Z|X}$

Objective. Given $(x_1, y_1) \dots (x_N, y_N) \sim \mathbf{pop}_{XY}$, optimize

$$\max_{\phi,\psi} \; \min_{\theta} \; \underbrace{\frac{1}{N} \sum_{i=1}^{N} \sum_{z \in \mathcal{Z}} p^{\psi}_{Z|Y}(z|y_i) \log \frac{q^{\phi}_{Z|X}(z|x_i)}{q^{\theta}_{Z}(z)}}_{\text{empirical estimate of } H^{+}_{\psi,\theta}(Z) - H^{+}_{\psi,\phi}(Z|X)}$$

Outline

MMI

AMMI

- Structured bit string encoder

Experiments on document hashing

Model

Encoder ψ. Markov distribution over $\mathcal{Z} = \{0,1\}^m$ of order $o \geq 0$

$$p^{\psi}_{Z|Y}(z|y) = \prod_{i=1}^{m} p^{\psi}_{Z_i|YZ_{<i}}(z_i|y, i, \underbrace{z_{i-o} \ldots z_{i-1}}_{o \text{ previous bits}})$$

Variational models θ, ϕ. Markov distributions of order $r, h \geq o$

Cross entropies.

$$-\sum_{z \in \{0,1\}^m} p^{\psi}_{Z|Y}(z|y) \log q^{\theta}_{Z}(z)$$

$$-\sum_{z \in \{0,1\}^m} p^{\psi}_{Z|Y}(z|y) \log q^{\phi}_{Z|X}(z|x)$$

Computable in time <u>linear in m</u> by the **forward algorithm**!

Summary of Training and Inference

- Parameterize Markov distributions $p^{\psi}_{Z|Y}, q^{\theta}_{Z}, q^{\phi}_{Z|X}$ over $\{0,1\}^m$ of orders o, r, h ($r, h \geq o$) with neural networks
- At each minibatch of samples from $\mathbf{pop}_{XY}$
  1. Take G gradient steps to minimize $H^{+}_{\psi,\theta}(Z)$ with respect to θ.
  2. Take 1 gradient step to maximize $H^{+}_{\psi,\theta}(Z) - H^{+}_{\psi,\phi}(Z|X)$ with respect to ψ, ϕ.
- **Inference.** Given new $y \sim \mathbf{pop}_Y$ compute

$$\arg\max_{z \in \{0,1\}^m} p^{\psi}_{Z|Y}(z|y) \qquad \text{(Viterbi)}$$

Outline

MMI

AMMI

- Structured bit string encoder

Experiments on document hashing

Unsupervised Document Hashing

- **Task**: encode a document into a binary vector such that nearest neighbors (in Hamming distance) share same labels
  - Labels are only used for evaluation: nearest-100 label precision

- **Autoencoding baselines**
  - NASH: discrete VAE, Bernoulli prior (Shen et al., 2018)
  - BMSH: discrete VAE, Bernoulli-mixture prior (Dong et al., 2019)
  - DVQ: vector-quantized VAE (van den Oord et al., 2017) with decomposition (Kaiser et al., 2018)

- **AMMI**: single-variable version ($\mathbf{pop}_Y$): learn $p^{\psi}_{Z|Y}$ and q^{θ}_Z by

$$\max_{\psi} \min_{\theta} H^{+}_{\psi,\theta}(Z) - H_{\psi}(Z|Y)$$

Results

Data	TMC				NG20				Reuters				Avg
	16b	32b	64b	128b	16b	32b	64b	128b	16b	32b	64b	128b	
bow	50.86				9.22				57.62				39.23
LSH	43.93	45.14	45.53	47.73	5.97	6.66	7.70	9.49	32.15	38.62	46.67	51.94	31.79
S-RBM	51.08	51.66	51.90	51.37	6.04	5.33	6.23	6.42	57.40	61.54	61.77	64.52	39.61
SpH	60.55	62.81	61.43	58.91	32.00	37.09	31.96	27.16	63.40	65.13	62.90	60.45	51.98
STH	39.47	41.05	41.81	41.23	52.37	58.60	58.06	54.33	73.51	75.54	73.50	69.86	56.61
VDSH	68.53	71.08	44.10	58.47	39.04	43.27	17.31	5.22	71.65	77.53	74.56	73.18	53.66
NASH	65.73	69.21	65.48	59.98	51.08	56.71	50.71	46.64	76.24	79.93	78.12	75.59	64.62
GMSH	67.36	70.24	70.86	72.37	48.55	53.81	58.69	55.83	76.72	81.83	82.12	78.46	68.07
DVQ	**71.47**	73.27	75.17	**76.24**	47.23	54.45	58.77	62.10	79.57	83.43	83.73	**86.27**	70.98
BMSH	70.62	**74.81**	**75.19**	74.50	**58.12**	**61.00**	60.08	58.02	79.54	82.86	82.26	79.41	71.37
AMMI	71.17	73.67	75.05	**76.24**	55.49	59.58	**63.80**	**65.74**	**82.62**	**83.39**	**85.18**	86.16	**73.17**
brute-force	70.52	✗	✗	✗	49.74	✗	✗	✗	79.97	✗	✗	✗	✗

Please see the paper for additional experiments on *predictive* document hashing: $(X, Y) =$ related news articles

Conclusions

- We presented **AMMI**: an adversarial approach to MMI
  - A new objective for learning <u>discrete structured</u> representations with <u>large mutual information</u>
  - Competitive with discrete VAEs on document hashing
- Future work includes
  - Extensions to other discrete structures (e.g., trees)
  - Better optimization

EXTRA SLIDES

Cross Entropy Upper Bounds with Variational Models

Variational models. q_Z^θ estimating the marginal of Z, $q_{Z|X}^\phi$ estimating the marginal of Z given X

$$H_\psi(Z) \leq \quad H^+_{\psi,\theta}(Z) = \underset{\substack{(x,y)\sim\mathbf{pop}_{XY} \\ z\sim p^\psi_{Z|Y}(\cdot|y)}}{\mathsf{E}} \left[-\log q_Z^\theta(z)\right]$$

$$H_\psi(Z|X) \leq \quad H^+_{\psi,\phi}(Z|X) = \underset{\substack{(x,y)\sim\mathbf{pop}_{XY} \\ z\sim p^\psi_{Z|Y}(\cdot|y)}}{\mathsf{E}} \left[-\log q_{Z|X}^\phi(z|x)\right]$$

Assuming a sufficiently expressive class of models for θ and ϕ,

$$H_\psi(Z) = \min_\theta \; H^+_{\psi,\theta}(Z)$$

$$H_\psi(Z|X) = \min_\phi \; H^+_{\psi,\phi}(Z|X)$$

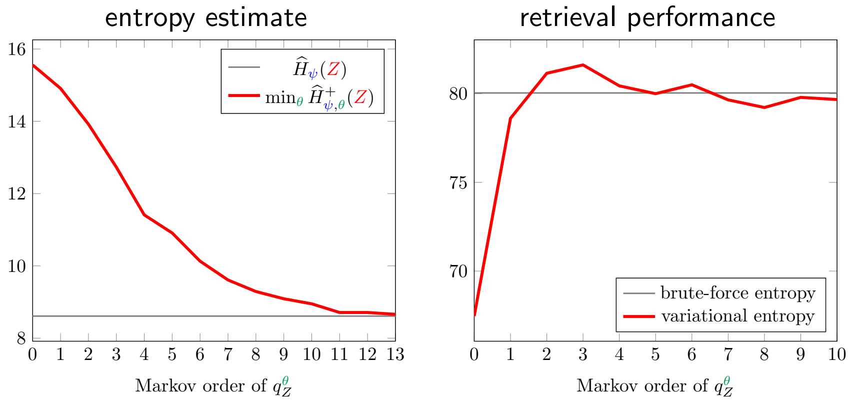
AMMI for Document Hashing

Given documents $y_1 \ldots y_N \sim \mathbf{pop}_Y$ optimize

$$\max_{\psi} \min_{\theta} \underbrace{\frac{1}{N} \sum_{i=1}^{N} \sum_{z \in \{0,1\}^m} p^{\psi}_{Z|Y}(z|y_i) \log \frac{p^{\psi}_{Z|Y}(z|y_i)}{q^{\theta}_{Z}(z)}}_{\text{empirical estimate of } H^{+}_{\psi,\theta}(Z) - H_{\psi}(Z|Y)}$$

- $p^{\psi}_{Z|Y}$ and q^{θ}_{Z}: Markov distributions over $\{0,1\}^m$
- Markov orders = hyperparameters

Importance of the Markov Order of the Variational Prior



Variational prior q_Z^{θ} needs enough "capacity" to model the <u>marginal</u> of Z under $p_{Z|Y}^{\psi}$!

Predictive Document Hashing

(X, Y): related news article pairs

$x =$ NYT article on 12/19/06 on a case against Yoko Ono's chauffeur

$y =$ AFP article on 12/20/06 on a case against Yoko Ono's chauffeur

$z =$ 0001010000000010000001111000100000000010100011000

Retrieval performance

	Dim	# Distinct Codes	Precision
BOW	20000	208808	26.66
BMSH	128	208004	75.77
DVQ	128	208655	76.80
AMMI	128	**153123**	**79.14**

Qualitative Analysis

Nearest neighbors in Hamming distance

Distance	Document
0	**O.J. Simpson lashed out at the family of the late Ronald Goldman, a day after they won the rights to Simpson's canceled "If I Did It" book about the slayings of Goldman**
1	News Corp. on Monday announced that it will cancel the release of a new book by former American football star O.J. Simpson and a related exclusive television interview
5	Phil Spector's lawyers have asked the judge to tell jurors they must find the record producer either guilty or not guilty of murder with no option to find lesser offenses
10	Sen. Ted Stevens' defense lawyer bore in on the prosecution's chief witness on Tuesday, portraying him to a jury as someone who betrayed a longtime friend to protect his fortune.
20	Words that cannot be said on American television are not often uttered at the U.S. Supreme Court, at least not by high-priced lawyers and the justices themselves.
50	Cols 1-6: Sending a strong message that the faltering economy will be his top focus, President-elect Barack Obama on Friday urged Congress to pass an economic stimulus package
90	President Hu Jintao's upcoming visits to Latin America and Greece would boost bilateral relations and deepen cooperation