
Learning Discrete Structured Representations by Adversarially Maximizing Mutual Information

Karl Stratos¹ Sam Wiseman²

Abstract

We propose learning discrete structured representations from unlabeled data by maximizing the mutual information between a structured latent variable and a target variable. Calculating mutual information is intractable in this setting. Our key technical contribution is an adversarial objective that can be used to tractably estimate mutual information assuming only the feasibility of cross entropy calculation. We develop a concrete realization of this general formulation with Markov distributions over binary encodings. We report critical and unexpected findings on practical aspects of the objective such as the choice of variational priors. We apply our model on document hashing and show that it outperforms current best baselines based on discrete and vector quantized variational autoencoders. It also yields highly compressed interpretable representations.

1. Introduction

Unsupervised learning of discrete representations is appealing because they correspond to natural symbolic representations in many domains (e.g., phonemes in speech signals, topics in text, and objects in images). However, working with discrete variables comes with technical challenges such as non-differentiability and nontrivial combinatorial optimization. Standard methods approach the problem within the framework of variational autoencoding (Kingma & Welling, 2014; Rezende et al., 2014) and bypass these challenges by adopting some form of gradient approximation and possibly strong independence assumptions (Bengio et al., 2013; van den Oord et al., 2017).

In this paper we are instead interested in a promising alternative framework based on maximal mutual information (MMI). Unlike autoencoding, MMI estimates a distribution

over latent variables without modeling raw signals by maximizing the mutual information between the latent and a target variable (Brown et al., 1992; Bell & Sejnowski, 1995; Tishby et al., 1999). It is well motivated as a principled approach to learning representations that retain only predictive information and drop noise. Its neural extensions have recently been quite successful in learning useful continuous representations across domains (Oord et al., 2018; Belghazi et al., 2018; Hjelm et al., 2019; Bachman et al., 2019).

We depart from these existing works on MMI in two important ways. First, we learn discrete structured representations. There are previous works on learning discrete representations with neural MMI (McAllester, 2018; Stratos, 2019), but they only consider unstructured representations which can transmit at most the log of the number of labels bits of information. Breaking the log bottleneck in the discrete regime requires making encodings structured, but it also makes exact computation intractable. Thus the feasibility of optimizing mutual information effectively in this setting remains unclear. We develop a tractable formulation that only requires tractable cross entropy by a combination of mild structural assumptions and an appropriate loss function (see below).

Second, we consider a new mutual information estimator based on the difference of entropies for learning representations. This is a crucial departure from existing works that optimize variational lower bounds (Poole et al., 2019). Estimators of a lower bound on mutual information have been shown to suffer fundamental limitations (McAllester & Stratos, 2020), suggesting a need to investigate alternative estimators. Our estimator is neither a lower bound nor an upper bound, yet it can be optimized adversarially as in generative adversarial networks (GANs) (Goodfellow et al., 2014). We show for the first time that such adversarial optimization of mutual information is a viable option for learning meaningful representations.

Our proposed discrete structured MMI is novel and largely uncharted in the literature. An important contribution of this paper is charting practical considerations for this alien approach by developing a concrete realization based on a structured model over binary encodings. More specifically, the model encodes an observation into a zero-one vector

¹Rutgers University ²Toyota Technological Institute at Chicago. Correspondence to: Karl Stratos <karlstratos@gmail.com>.

of length m , resulting in 2^m possible encodings. We show how mutual information can be estimated efficiently by adversarial dynamic programming with controllable Markov assumptions. One critical and unexpected finding is that the expressiveness of a variational prior needs to be strictly greater than that of the model (i.e., it has to be higher-order Markovian).

To demonstrate the utility of our model in a real-world problem, we apply it to unsupervised document hashing (Dong et al., 2019; Shen et al., 2018; Chaidaroon et al., 2018). The task is to compress an article into a drastically smaller discrete encoding that preserves semantics. Our model outperforms current state-of-the-art baselines based on discrete VAEs (Kingma & Welling, 2014; Maddison et al., 2017; Jang et al., 2017) and VQ-VAEs (van den Oord et al., 2017) with Bernoulli priors. We additionally design a predictive version of document hashing in which the model is tasked with encoding a future article with the knowledge of a past article. We find that our model achieves favorable performance with highly compressed and interpretable representations.

2. Related Work

Many successful approaches to unsupervised representation learning are based on density estimation. For instance, it is now very common in natural language processing to make use of continuous representations that are learned in the process of modeling a conditional distribution $p_{Y|X}$, such as the conditional distribution of a word Y given a context window X (Mikolov et al., 2013; Peters et al., 2018; Devlin et al., 2019). In this case the input pair (X, Y) is easily sampled from unlabeled data, by masking an observed word. There is also much work that identifies representations, often continuous, with the latent variables in an unconditional density model of Y (Kingma & Welling, 2014; Rezende et al., 2014; Higgins et al., 2017).

Learning representations through density estimation, however, suffers from certain limitations. First, it may be unnecessary to fully model the density of noisy, raw data when we are only interested in learning representations. Second, many standard approaches to learning discrete-valued latent representations in the context of density estimation require the use of either biased gradient estimators (Bengio et al., 2013; van den Oord et al., 2017) or high variance ones (Mnih & Gregor, 2014; Mnih & Rezende, 2016).

Maximal mutual information (MMI) is a refreshingly different approach to unsupervised representation learning in which we estimate a conditional distribution over latent representations by maximizing mutual information under these distributions. In contrast with density estimation, there is no issue of modeling noise since the model never estimates a

distribution over raw signals (i.e., there is no decoder). The mutual information objective has been shown to produce state-of-the-art representations of images, speech, and text (Bachman et al., 2019; Oord et al., 2018).

The focus with MMI so far has been largely limited to learning continuous representations. Existing works on learning discrete representations with MMI only consider unstructured one-of- m -labels representations (McAllester, 2018; Stratos, 2019) due to computational reasons. Our main contribution is a tractable formulation for discrete structured MMI. This involves an adversarial objective reminiscent of GANs (Goodfellow et al., 2014) and radically different from existing MMI objectives based on variational lower bounds of mutual information (Poole et al., 2019). Other than tractability reasons, the choice of the objective can be theoretically motivated as avoiding statistical limitations of estimating lower bounds on mutual information (McAllester & Stratos, 2020).

3. Discrete Structured MMI

Let p_{XY} denote an unknown but samplable joint distribution over raw signals (X, Y) . We assume discrete (X, Y) for simplicity and relevance to our experimental setting (document hashing), but the formulation can be easily adapted to the continuous case. We introduce an encoder $p_{Z|Y}^\psi$ that defines a conditional distribution over a discrete latent variable Z representing the encoding of Y and aim to maximize $I_\psi(X, Z)$: the mutual information between X and Z under p_{XY} and $p_{Z|Y}^\psi$. By the data processing inequality, the objective is a lower bound on $I(X, Y)$ and can be viewed as distilling the predictive information of Y about X into Z .

This formulation alone is meaningless since it admits the trivial solution $Z = Y$. In order to achieve compression, there are various options. In the information bottleneck method (Tishby et al., 1999), we additionally regularize the information rate of Z by simultaneously minimizing $I_\psi(Y, Z)$. Here we advocate a more direct approach by giving an explicit budget H_{\max} on the entropy of Z .

$$\max_{\psi: H_\psi(Z) \leq H_{\max}} I_\psi(X, Z)$$

Equivalently, we can maximize $I_\psi(X, Z)$ with a finite encoding space \mathcal{Z} such that $|\mathcal{Z}| \leq 2^{H_{\max}}$. Even with small \mathcal{Z} the objective is intractable because it involves marginalization over Y . Using that $I_\psi(X, Z) = H_\psi(Z) - H_\psi(Z|X)$, we introduce a variational model $q_{Z|X}^\phi$ and optimize

$$\max_{\psi, \phi} H_\psi(Z) - H_{\psi, \phi}^+(Z|X) \quad (1)$$

where $H_{\psi, \phi}^+(Z|X)$ denotes the cross entropy between $p_{Z|X}^\psi$ (i.e., the distribution over Z given X defined under $p_{Z|Y}^\psi$

and p_{XY}) and $q_{Z|X}^\phi$. By the usual property of cross entropy, $H_{\psi,\phi}^+(Z|X)$ is an upper bound on $H_\psi(Z|X)$, hence the objective (1) is a lower bound on $I_\psi(X, Z)$.

Unfortunately, the applicability of this variational lower bound is critically limited to settings in which the entropy of Z is tractable (McAllester, 2018; Stratos, 2019) or constant with respect to learnable parameters (Chen et al., 2016; Alemi et al., 2017). When Z is small, $H_\psi(Z)$ can be easily estimated from N samples $(x_1, y_1) \dots (x_N, y_N) \sim p_{XY}$ as

$$\hat{H}_\psi(Z) = \sum_{z \in \mathcal{Z}} \left(\frac{\sum_l p_{Z|Y}^\psi(z|y_l)}{N} \right) \log \left(\frac{N}{\sum_l p_{Z|Y}^\psi(z|y_l)} \right) \quad (2)$$

But this explicit calculation is clearly infeasible for large \mathcal{Z} . Furthermore, the log bottleneck on the budget $H_{\max} \leq \log |\mathcal{Z}|$ then implies that it is infeasible to achieve a large information rate using this naive formulation (e.g., even if we specify $|\mathcal{Z}|$ to be a trillion we have $H_{\max} \leq 40$).

3.1. Tractable Formulation

To allow for large H_{\max} , we propose to make Z structured. A simplest example of structured Z is a binary vector of length m which yields $|\mathcal{Z}| = 2^m$ so that H_{\max} can be as large as m . More generally, Z can be any structure whose size is exponential in some controllable integer m .

3.1.1. TRACTABLE CROSS ENTROPY

The first key ingredient in deriving a tractable formulation is the tractability of estimating the cross entropy between $p_{Z|X}^\psi$ and $q_{Z|X}^\phi$ from samples.

Assumption 3.1. *The cross entropy $H_{\psi,\phi}^+(Z|X)$ between $p_{Z|X}^\psi$ and $q_{Z|X}^\phi$ estimated from N iid samples of (X, Y)*

$$\hat{H}_{\psi,\phi}^+(Z|X) = -\frac{1}{N} \sum_{l=1}^N \left(\sum_{z \in \mathcal{Z}} p_{Z|Y}^\psi(z|y_l) \log q_{Z|X}^\phi(z|x_l) \right)$$

can be computed in time polynomial in m where $|\mathcal{Z}| = O(2^m)$.

There is a class of structured probabilistic models with standard conditional independence assumptions such that Assumption 3.1 holds. For instance, in the case of $Z \in \{0, 1\}^m$, we may impose Markov assumptions and define (with the convention $z_i = 0$ for $i < 1$)

$$p_{Z|Y}^\psi(z|y) = \prod_{i=1}^m p_{Z_i|YZ_{<i}}^\psi(z_i|y, i, z_{i-o:i-1})$$

$$q_{Z|X}^\phi(z|x) = \prod_{i=1}^m q_{Z_i|XZ_{<i}}^\phi(z_i|x, i, z_{i-h:i-1})$$

Algorithm 1 CrossEntropy

Input: $p(z_i|i, z_{i-o:i-1})$ for $z_{i-o:i} \in \{0, 1\}^{o+1}$ and $i \in [m]$; $q(z_i|i, z_{i-o':i-1})$ for $z_{i-o':i} \in \{0, 1\}^{o'+1}$ and $i \in [m]$ where $o' \geq o$

Subroutine: Forward(p) that computes π in $O(m2^o)$ time such that $\pi(z_{i-o:i-1}|i)$ is the marginalized probability of $\bar{z} \in \{0, 1\}^{i-1}$ ending in $z_{i-o:i-1}$ under p (given in the supplementary material)

Output: Cross entropy between p and q

$$H(p, q) = - \sum_{z \in \{0,1\}^m} p(z) \log q(z)$$

Runtime: $O(m2^{o'})$

Forward computation: $\pi \leftarrow \text{Forward}(p)$

Marginals: For $i = 1 \dots m$, for $z_{i-o':i} \in \{0, 1\}^{o'+1}$,

$$\mu(z_{i-o':i}|i) \leftarrow \pi(z_{i-o':i-o'+o-1}|i - o' + o) \times \left(\prod_{j=i-o'+o}^i p(z_j|j, z_{j-o:j-1}) \right)$$

where we overwrite $\pi(z_{i-o':i-o'+o-1}|i - o' + o) = p(z_{i-o'}|i - o')$ if $o = 0$

Cross entropy: Set $H(p, q)$ as the following scalar

$$- \sum_{i=1}^m \sum_{z_{i-o':i} \in \{0,1\}^{o'+1}} \mu(z_{i-o':i}|i) \log q(z_i|z_{i-o':i-1})$$

where $z_{i:j} = (z_i \dots z_j)$ and $o \leq h$ are the Markov orders of $p_{Z|Y}^\psi$ and $q_{Z|X}^\phi$. It can be easily verified that the estimate of $H_{\psi,\phi}^+(Z|X)$ based on a single sample $(x, y) \sim p_{XY}$ is

$$- \sum_{i=1}^m \sum_{z_{i-h:i}} \mu(z_{i-h:i}|i, y) \log q_{Z_i|XZ_{<i}}^\phi(z_i|x, i, z_{i-h:i-1})$$

where $\mu(z_{i-h:i}|i, y)$ is the marginal probability of the length- $(h+1)$ sequence $z_{i-h:i} \in \{0, 1\}^{h+1}$ ending at position i under the conditional distribution $p_{Z|Y}^\psi(\cdot|y)$. These marginals can be computed by applying a variant of the forward algorithm (Rabiner, 1989) (see the supplementary material). We give a general algorithm that computes the cross entropy between any distributions over $\mathcal{Z} = \{0, 1\}^m$ with Markov orders $o \leq o'$ in $O(m2^{o'})$ time in Algorithm 1.

While we focus on the choice $Z \in \{0, 1\}^m$ for concreteness, we emphasize that similar structural assumptions can be made to consider others. For instance, the conditional entropy of tree-structured Z can be computed using a variant of the inside algorithm (Hwa, 2000). We leave exploring

other types of structure as future work.

3.1.2. ADVERSARIAL OPTIMIZATION

Assumption 3.1 allows us to estimate the second term (cross entropy) of the objective (1) $H_\psi(Z) - H_{\psi,\phi}^+(Z|X)$, but it is still insufficient for estimating the objective since the first term (entropy) remains intractable. Assumption 3.1 only imposes conditional independence: conditioning on the input y , the probability of the i -th value of z is independent of $z_{<i-o}$ under $p_{Z|Y}^\psi$. This independence breaks for the unconditional distribution $p_Z^\psi(z) = \mathbb{E}_{y \sim p_Y}[p_{Z|Y}^\psi(z|y)]$. Consequently the entropy term $H_\psi(Z)$ does not decompose. Note that while the conditional entropy $H_\psi(Z|Y)$ remains tractable, we cannot use it as a meaningful approximation of $H_\psi(Z)$ since the error is exactly $I_\psi(Y, Z)$ which is zero iff Z is independent of Y (i.e., vacuous encoding).

Thus we propose to introduce an additional variational model q_Z^θ to estimate the intractable distribution p_Z^ψ . We would like to make the resulting variational approximation a lower bound on $H_\psi(Z)$ so that the objective remains maximization over all models. Unfortunately, when entropy is large (which is our setting) meaningful lower bounds are impossible (McAllester & Stratos, 2020). This motivates us to again consider the cross-entropy upper bound $H_{\psi,\theta}^+(Z) \geq H_\psi(Z)$ with the following assumption.

Assumption 3.2. *The cross entropy $H_{\psi,\theta}^+(Z)$ between p_Z^ψ and q_Z^θ estimated from N iid samples of Y*

$$\hat{H}_{\psi,\theta}^+(Z) = -\frac{1}{N} \sum_{l=1}^N \left(\sum_{z \in \mathcal{Z}} p_{Z|Y}^\psi(z|y_l) \log q_Z^\theta(z) \right)$$

can be computed in time polynomial in m where $|\mathcal{Z}| = O(2^m)$.

In the binary vector setting $Z \in \{0, 1\}^m$, we can define q_Z^θ to be a Markov model of order $r \geq o$

$$q_Z^\theta(z) = \prod_{i=1}^m q_{Z_i|Z_{<i}}^\theta(z_i|i, z_{i-r:i-1})$$

Then Algorithm 1 can be used to estimate $H_{\psi,\theta}^+(Z)$ in $O(m2^r)$ time.

This gives our final objective

$$\max_{\psi,\phi} \min_{\theta} H_{\psi,\theta}^+(Z) - H_{\psi,\phi}^+(Z|X) \quad (3)$$

which is tractable by Assumption 3.1 and 3.2. Note that for any choice of ψ , exact optimization over ϕ and θ recovers the original objective $I_\psi(X, Z)$. It can be interpreted as a simultaneously collaborative and adversarial game. The second term (cross entropy minimization) encourages ψ and ϕ to agree on the encoding Z of Y . The first term (entropy

Algorithm 2 AdversarialMMI

Input: Models $p_{Z|Y}^\psi, q_{Z|X}^\phi, q_Z^\theta$ satisfying Assumption 3.1 and 3.2; samplable p_{XY} ; gradient update function Step; validation task T

Hyperparameters: Initialization range α , batch size N , number of adversarial gradient steps G , adversarial learning rate η' , learning rate η , entropy weight β

$\psi, \phi, \theta \leftarrow \text{Unif}(-\alpha, \alpha)$

repeat

for $S \sim p_{XY}^N$ **do**

 For G times: $\theta \leftarrow \text{Step}(\hat{H}_{\psi,\theta}^+(Z), \theta, \eta')$

$\psi, \phi \leftarrow \text{Step}(\hat{H}_{\psi,\phi}^+(Z|X) - \beta \hat{H}_{\psi,\theta}^+(Z), \{\psi, \phi\}, \eta)$

end for

until $T(\psi, \phi, \theta)$ stops improving

maximization) encourages ψ to diversify its prediction of Z but also use information from Y to thwart the opponent θ who does not have access to Y .

A notable aspect of the objective is that it is neither an upper bound nor a lower bound on $I_\psi(X, Z)$; we cannot guarantee that $H_{\psi,\theta}^+(Z) - H_{\psi,\phi}^+(Z|X)$ estimated from N samples is larger or smaller than $I_\psi(X, Z)$. While we lack guarantees, theoretical and empirical evidence that this bypasses limitations of lower bounds on mutual information has been shown in McAllester & Stratos (2020).

Inference At test time, given input y we calculate

$$z^* \in \arg \max_{z \in \mathcal{Z}} p_{Z|Y}^\psi(z|y)$$

and use z^* as a discrete structured representation of y . In the current setting in which $p_{Z|Y}^\psi$ is an order- o Markov distribution over $\mathcal{Z} = \{0, 1\}^m$, we can calculate z^* in $O(m2^o)$ time using a variant of the Viterbi algorithm (Viterbi, 1967).

3.2. Practical Issues

We give details of the proposed adversarial MMI training procedure in Algorithm 2. As input we assume model definitions $p_{Z|Y}^\psi, q_{Z|X}^\phi$, and q_Z^θ that satisfy Assumption 3.1 and 3.2, samplable p_{XY} , gradient update function Step (we use Adam for all our experiments (Kingma & Ba, 2014)), and a validation task T . The validation task evaluates the quality of the representation predicted by $p_{Z|Y}^\psi$ and is particularly needed since the running estimate of the adversarial objective (3) may not reflect actual progress. We delineate certain practical issues that are important in making Algorithm 2 effective.

Expressive variational prior We find that it is critical to make the variational prior q_Z^θ strictly more expressive than

the posterior $p_{Z|Y}^\psi$. That is, there exist distributions over \mathcal{Z} that can be modeled by q_Z^θ but not by $p_{Z|Y}^\psi$ (conditioning on any y). In the context of Markov models over binary vectors $Z \in \{0, 1\}^m$, this means the Markov order r of q_Z^θ is strictly greater than the Markov order o of $p_{Z|Y}^\psi$ (Algorithm 1 allows for any $r \geq o$). Recall that $Z_1 \dots Z_m$ are conditionally independent under $p_{Z|Y}^\psi$ but not independent under p_Z^ψ (Section 3.1.2). Thus we must model p_Z^ψ using a distribution q_Z^θ that is strictly more powerful than $p_{Z|Y}^\psi$. The benefit of this explicit joint entropy maximization is suggested in experiments later in which we show that our approach is more effective at learning representations than discrete VAEs or VQ-VAEs (which do not explicitly maximize entropy) as m becomes larger.

We also find it helpful to overparameterize q_Z^θ . We use a feedforward network with $\gg m2^r$ parameters and a tunable number of ReLU layers to define the distribution.

Aggressive inner-loop optimization The adversarial objective (3) reduces to the non-adversarial objective (1) if the inner minimization over θ is solved exactly. We find it important to mimic this by taking multiple (G) gradient steps for θ with large learning rate η' before taking a gradient step for $\{\psi, \phi\}$. Aggressive inner-loop optimization has been shown helpful in other contexts such as VAEs (He et al., 2019). Note that θ is still carried across batches and not learned from scratch at every batch.

Entropy weight Finally, we find it useful to introduce a tunable weight $\beta \geq 1$ for the entropy term, akin to the weight for KL divergence in β -VAEs (Higgins et al., 2017). Note that optimizing this weighted objective is equivalent to optimizing $I_\psi(X, Z) + (\beta - 1)H_\psi(Z)$. The weight can be used to determine a task-specific trade-off between predictiveness and diversity in Z .

4. Experiments

We now study empirical aspects of our proposed adversarial MMI approach (henceforth AMMI) with extensive experiments.¹ We consider unsupervised document hashing (Chaidaroon et al., 2018) as a main testbed for evaluating the quality of learned representations. The task is to compress an article into a drastically smaller discrete encoding that preserves semantics and formulated as an autoencoding problem. To study methods in a predictive setting, we also develop a variant of this task in which the representation of an article is learned to be predictive of the encoding of a related article.

¹Code: <https://github.com/karlstratos/ammi>

4.1. Unsupervised Document Hashing

Let Y be a random variable corresponding to a document. The goal is to learn a document encoder $q_{Z|Y}$ that defines a conditional distribution over binary hashes $Z \in \{0, 1\}^m$. The quality of document encodings is evaluated by the average top-100 precision. Specifically, given a document at test time, we retrieve 100 nearest neighbors from the training set under the encoding measured by the Hamming distance and check how many of the neighbors have overlapping topic labels (thus we assume annotation only for evaluation).

In the literature this is typically approached as an autoencoding problem in which $q_{Z|Y}$ is estimated by maximizing the evidence lower bound (ELBO) on the marginalized log likelihood of training documents

$$\max_{q_{Z|Y}, p_{Y|Z}} \mathbb{E}_{\substack{y \sim p_Y \\ z \sim q_{Z|Y}}} [\log p_{Y|Z}(y|z)] - D_{\text{KL}}(q_{Z|Y} || p_Z) \quad (4)$$

where p_Z is a fixed prior suitable for the task. For example, the current state-of-the-art model (BMSH) defines p_Z as a mixture of Bernoulli distributions (Dong et al., 2019). Here $q_{Z|Y}$ is treated as a variational model that estimates the intractable posterior $p_{Z|Y}$ under the model $p_{Y|Z}(y, z) = p_{Y|Z}(y|z)p_Z(z)$.

In contrast, we propose to learn a document encoder $p_{Z|Y}^\psi$ by the following adversarial formulation of the mutual information between Y and Z :

$$\max_{\psi} \min_{\theta} H_{\psi, \theta}^+(Z) - H_{\psi}(Z|Y) \quad (5)$$

where q_Z^θ is a variational model that estimates the intractable prior p_Z^ψ under the model $p_{Y|Z}^\psi(y, z) = p_{Z|Y}^\psi(z|y)p_Y(y)$. This can be seen as a single-variable variant of the more general objective (3) in which $X = Y$ and ϕ is tied with ψ . Note the absence of a decoder (Section 2).

4.1.1. MODELS

BMSH We follow the standard setting in BMSH for all our models (Dong et al., 2019). The raw document representation y is a high-dimensional TFIDF vector computed from preprocessed corpora (TMC, NG20, and Reuters) provided by Chaidaroon et al. (2018). We aim to learn an m -dimensional binary vector representation $z \in \{0, 1\}^m$ where we vary the number of bits $m = 16, 32, 64, 128$. All VAE-based models compute a continuous embedding \tilde{z} by feeding y through a feedforward layer (FF) and apply some discretization operation on \tilde{z} to obtain z . BMSH computes $\tilde{z} = \sigma(\text{FF}(y)) \in \mathbb{R}^m$ and samples $z \sim \text{Bernoulli}(\tilde{z})$ from which y is reconstructed.² The ELBO objective (4) is opti-

²We refer to Dong et al. (2019) for details of the decoder and the Bernoulli mixture prior since they are not needed for AMMI.

mized by straight-through estimation (Bengio et al., 2013).³

DVQ We are interested in comparing AMMI to mainstream discrete representation learning techniques. Thus we additionally consider vector quantized VAEs (VQ-VAEs) which substitute sampling with a nearest neighbor lookup and are shown to be useful in many tasks (van den Oord et al., 2017; Razavi et al., 2019). In particular, we adopt the decomposed vector quantization (DVQ) proposed in Kaiser et al. (2018). The model learns m codebooks $C_1 \dots C_m \in \mathbb{R}^{2 \times D}$ where the row index in C_i corresponds to $z_i \in \{0, 1\}$. The encoder computes $\tilde{z} = \text{FF}(y) \in \mathbb{R}^{mD}$ and quantizes the i -th segments in \tilde{z} against C_i (implicitly yielding z) from which the decoder reconstructs y . DVQ is trained by minimizing the reconstruction loss and the vector quantization loss. It can be seen as optimizing the ELBO objective (4) with a uniform prior over Z and a point-mass posterior $q_{Z|Y}$.

AMMI Our model consists of an encoder $p_{Z|Y}^\psi$ and a variational prior q_Z^θ , which are respectively parameterized by order- o and order- r Markov models over $\{0, 1\}^m$. The encoder computes $p_{Z|Y}^\psi(\cdot|y) = \sigma(\text{FF}(y)) \in \mathbb{R}^{m2^o}$ which gives the model’s probability of $z_i = 1$ conditioning on each value of $z_{i-o:i-1}$ for every $i \in [m]$. Similarly, the prior computes $q_Z^\theta = \sigma(\text{FF}(\Theta)) \in \mathbb{R}^{m2^r}$ where $\Theta \in \mathbb{R}^{m \times H}$ is a learnable embedding dictionary with dimension H . We can then use Algorithm 1 to calculate

$$\begin{aligned} \widehat{H}_{\psi, \theta}^+(Z) &= \text{CrossEntropy} \left(p_{Z|Y}^\psi(\cdot|y), q_Z^\theta \right) \\ \widehat{H}_{\psi}^+(Z|Y) &= \text{CrossEntropy} \left(p_{Z|Y}^\psi(\cdot|y), p_{Z|Y}^\psi(\cdot|y) \right) \end{aligned}$$

where in practice we use a batch of samples $y_1 \dots y_N$ to estimate these quantities. The algorithm can be batched efficiently.

4.1.2. HYPERPARAMETER TUNING

All hyperparameters for AMMI are shown in Algorithm 2. We perform random grid search on the validation portion of each dataset. We find that an effective range of values is: initialization $\alpha = 0.1$, batch size $N \in \{16, 32, 64, 128\}$, adversarial step $G \in \{1, 2, 4\}$, adversarial learning rate $\eta' \in \{0.03, 0.01, 0.003, 0.001\}$, learning rate $\eta \in \{0.03, 0.01, 0.003, 0.001, 0.0003, 0.0001\}$, and entropy weight $\beta \in \{1, 1.5, 2, 2.5, 3, 3.5\}$. We similarly perform random grid search on all hyperparameters of BMSH and DVQ. We use an NVIDIA Quadro RTX 6000 with 24GB memory.

The Markov orders o and r of the encoder and the prior are also controllable hyperparameters in AMMI. We find

³That is, the decoder receives $(z - \tilde{z}).\text{detach}() + \tilde{z}$ to back-propagate gradients directly to \tilde{z} .

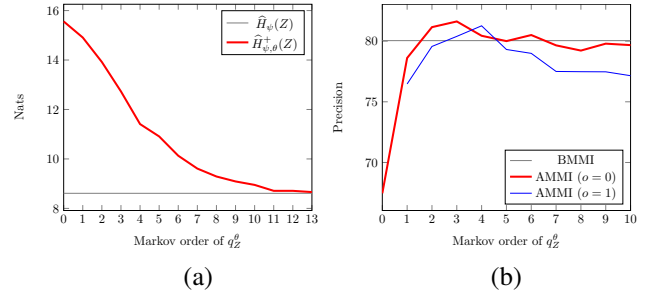


Figure 1. Behavior of the model using different Markov orders r for the variational prior q_Z^θ ; we use $m = 16$ bits and $o = 0$ as the Markov order of $p_{Z|Y}^\psi$. For each choice of r all hyperparameters of the variational model are fully optimized. (a) shows the cross-entropy upper bound estimated on a fixed batch of samples. (b) shows the validation precision on Reuters. The blue line additionally shows the result using $o = 1$. The gray line corresponds to the setting in which we calculate the entropy by brute-force.

that setting $o = 0$ is sufficient for this task (i.e., bits are independent conditioning on the document). But the choice of r is crucial, as we show below.

4.1.3. IMPORTANCE OF THE MARKOV ORDER OF THE VARIATIONAL PRIOR

We first examine the feasibility of variational approximation of the prior. To this end, we consider the small-bit setting $m = 16$ in which we can explicitly enumerate 2^{16} values of Z to estimate the entropy $H_\psi(Z)$ using equation (2). In this case the objective becomes non-adversarial. We refer to this model as BMMI (brute-force MMI) which only consists of $p_{Z|Y}^\psi$ trained by $\max_\psi H_\psi(Z) - H_\psi(Z|Y)$.

Figure 1 shows two experiments that illustrate the importance of the Markov order r of the variational prior q_Z^θ . First, we fix a BMMI with order $o = 0$ partially trained on Reuters and a random batch of samples to calculate the empirical entropy $\widehat{H}_\psi(Z)$ by brute-force. Then for each choice of $r = 0, 1, \dots, 13$, we minimize the empirical cross entropy $\widehat{H}_{\psi, \theta}^+(Z)$ between $p_{Z|Y}^\psi$ and q_Z^θ over θ with full hyperparameter tuning. Figure 1(a) shows that we need $r \gg o$ to achieve a realistic estimate of the empirical entropy. The necessary value of r clearly depends on $p_{Z|Y}^\psi$: $r > 10$ yields an accurate estimate for the partially trained BMMI used in this experiment.

Next, we examine the best achievable validation precision on Reuters across different values of r . We perform full hyperparameter tuning for BMMI ($o = 0$) and for AMMI with $r = 0, 1, \dots, 10$. We see that the performance is poor for $r = o = 0$, but it quickly becomes competitive as $r > 0$ and even surpasses the performance of BMMI. We hypothesize that the adversarial formulation has beneficial regularization effects in addition to making the objective

Table 1. Results on unsupervised document hashing. For each dataset we show test precisions of the top 100 retrieved documents using {16, 32, 64, 128}-bit binary vector encoding. See the main text for task and model descriptions.

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	70.96	74.16	75.22	76.27	55.18	59.56	63.98	66.18	81.73	84.46	85.06	86.02	73.23
BMMI	70.52				49.74				79.97				

tractable for large m : we leave a deeper investigation of this phenomenon as future work. We use $r = 3$ for all our experiments.

4.1.4. RESULTS

Table 1 shows top-100 precisions on the test portion of each dataset TMC, NG20, and Reuters using $m = 16, 32, 64, 128$ bits. Baselines include locality sensitive hashing (LSH) (Datar et al., 2004), stack restricted Boltzmann machines (S-RBMs) (Hinton, 2012), spectral hashing (SpH) (Weiss et al., 2009), self-taught hashing (STH) (Zhang et al., 2010), variational deep semantic hashing (VDSH) (Chaidaroon et al., 2018), and neural architecture for semantic hashing (NASH) (Shen et al., 2018), as well as BMSH (Dong et al., 2019) and DVQ described in Section 4.1.1. The naive baseline BOW refers to the bag-of-words representation $\text{BOW}(y) \in \{0, 1\}^{|V|}$ that indicates presence of words in document y .⁴

We see that AMMI performs favorably to current state-of-the-art methods, yielding the best average precision across datasets and settings. In particular, the precision of AMMI is significantly higher than the best previous result given by BMSH when m is large. With 128 bits, AMMI achieves 76.27 vs 74.50, 66.18 vs 58.02, and 86.02 vs 79.41 on TMC, NG20, and Reuters. We hypothesize that this is partly due to the explicit entropy maximization in AMMI that considers all bits jointly via dynamic programming. While this is implicitly enforced in the mixture prior in BMSH, direct entropy maximization seems to be more effective.

In the case of 16 bits, we also report precisions of BMMI

⁴The vocabulary size $|V|$ is 20000, 9988, and 7164 for TMC, NG20, and Reuters.

that estimates entropy exactly by brute-force (it is computationally intractable to train BMMI with larger than 16 bits). We see that AMMI is again able to achieve better results potentially due to regularization effects. Finally, we observe that the newly proposed DVQ baseline is quite competitive with BMSH and also achieves higher precision when m is large; we suspect that the decomposed encoding allows the model to make better use of multiple codebooks as reported in Kaiser et al. (2018).

4.2. Predictive Document Hashing

Unsupervised document hashing only considers a single variable Y and does not test the conditional formulation $Y|X$. Hence we introduce a new task, predictive document hashing, in which (X, Y) represent distinct articles that discuss the same event. It is clear that $I(X, Y)$ is large: the large uncertainty of a random article is dramatically reduced given a related article.

We construct such article pairs from the Who-Did-What dataset (Onishi et al., 2016). We remove all overlapping articles so that each article appears only once in the entire training/validation/test data containing 104404/8928/7326 document pairs. We give more details of the dataset in the supplementary material. Similarly as before, we use 20000-dimensional TFIDF representations as raw input and consider the task of compressing them into $m = 128$ binary bits. The quality of the binary encodings is measured by top-100 matching precision: given a test article y , we check if the correct corresponding article x is included in 100 nearest neighbors of y under the encoding based on the Hamming distance.

Table 2. Qualitative analysis of AMMI document representations learned by predictive document hashing. For each considered document (top row), we show documents with Hamming distance at least 1, 5, 10, 20, 50, and 90 under the representations to examine their semantic differences.

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
0	Ukrainian President Leonid Kuchma had a meeting on Monday evening with Polish President Alexander Kwasniewski and Lithuanian President Valdas Adamkus
1	Radical Ukrainian opposition figure Yulia Tymoshenko Wednesday ventured into the hostile eastern mining bastion of Prime Minister Viktor Yanukovich
5	Ukrainian President Viktor Yushchenko was forced into an emergency landing Thursday and seized the aircraft of his bitter political foe, Prime Minister Yulia Tymoshenko
10	On a clearing in this disputed city, where enemy homes were bulldozed after the conflict in August, Mayor Yuri M. Luzhkov promised this month to build a new neighborhood
20	Barack Obama is the "American Gorbachev" who will ultimately destroy the United States, militant Russian nationalist Vladimir Zhirinovskiy said Tuesday.
50	Ministers from Pacific Rim nations warned Thursday that imposing trade barriers in reaction to the global economic downturn would only deepen the crisis.
90	We shall move the following graphics: US IRAQ QAEDA Graphic with portraits of Osama bin Laden and Colin Powell, examining US claims that the latest bin Laden tape reinforces
0	NASCAR has a new rivalry: Carl Edwards vs. Kyle Busch. Edwards called the latest installment payback and Busch promised that retribution will come down the road.
1	Penske Racing teammates Ryan Briscoe and Helio Castroneves filled the front-row Friday for Edmonton Indy, repeating their 1-2 finish in last week’s IndyCar race
5	Brazilian race-car driver Helio Castroneves upset Spice Girl Melanie Brown to capture the fifth "Dancing With the Stars" mirrorball trophy in the television dance competition.
10	Audi overcame the challenge of two Peugeot cars and wet racing conditions Sunday to win the 24 Hours of Le Mans for the fourth straight year.
20	The president of cycling’s governing body Monday insists the doping problems in his sport do not threaten its place in the Olympics.
50	Robert Pattinson, who stars as the vampire heartthrob Edward Cullen in the forthcoming movie "Twilight," stepped onto a riser at the King of Prussia Mall
90	Australian Prime Minister John Howard reshuffled his cabinet Tuesday, appointing Education Minister Brendan Nelson to the defence portfolio.
0	How did that Van Halen song go? "I found the simple life ain’t so simple ... " The iconic Los Angeles metal band is set to be inducted Monday night into
8	The boys from Van Halen, most notably mercurial guitarist Eddie Van Halen, showed up as promised at a news conference Monday to announce their fall tour with original singer
20	The PG-13-rated thriller gave 20th Century Fox its first No. 1 launch in seven months. The opening-night crowd was heavily male and young, matching the video-game
50	Due to Wednesday night’s victory, a mathematician and avid Vasco soccer fan calculated on Thursday that the team’s chances of being dropped into the second division fell by
90	A top Iranian minister who admitted to faking his university degree will face a motion of no confidence on Tuesday on charges that he tried to bribe members of Parliament

Table 3. Results on predictive document hashing. For each model we show the representation dimension, number of distinct codes (i.e., clusters) induced on 208808 training documents, and top-100 precision on the test set.

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

4.2.1. MODELS

AMMI now consists of a pair of encoders $p_{Z|Y}^{\psi}$ and $p_{Z|X}^{\phi}$ as well as a variational prior q_Z^{θ} , which are respectively parameterized by order- o , order- h , and order- r Markov models over $\{0, 1\}^m$. Based on our findings in the previous section we use $o = h = 0$ and $r = 3$. We train the model with Algorithm 2.

To compare with VAEs, we consider a conditional variant which optimizes the ELBO objective under the joint distribution $p_{XYZ}(x, y, z) = p_Y(y)p_{Z|Y}(z|y)p_{X|Z}(x|z)$ with an approximation of the true posterior $q_{Z|Y}(z|y) \approx p_{Z|XY}(z|x, y)$ which can be used as a document encoder. In this setting, training for BMSH remains unchanged except that the the model predicts X instead of Y for reconstruction and uses conditional prior $p_{Z|Y}$ in the KL regularization term. The DVQ model likewise simply predicts X instead of Y but loses its ELBO interpretation. We tune hyperpa-

rameters of all models similarly as before.

4.2.2. RESULTS AND QUALITATIVE ANALYSIS

Table 1 shows top-100 precisions on the test portion. We see that AMMI again achieves the best performance in comparison to BMSH and DVQ. We also report the number of distinct values of z induced on the 208808 training articles (union of X and Y). We see that AMMI learns the most compact clustering which nonetheless generalizes best.

We conduct qualitative analysis of the document encodings by examining articles with increasing Hamming distance (i.e., semantic drift). Table 2 shows illustrative examples. The article about the O.J. Simpson trial drifts to the Phil Spector trial, the Ted Stevens trial, and eventually other unrelated subjects in politics and economy. The article about NASCAR drifts to other racing reports, cycling, movie stars and politics.

5. Conclusions

We have presented AMMI, an approach to learning discrete structured representations by adversarially maximizing mutual information. It obviates the intractability of entropy estimation by making mild structural assumptions that apply to a wide class of models and optimizing the difference of cross-entropy upper bounds. We have derived a concrete instance of the approach based on Markov models and identified important practical issues such as the expressiveness of the variational prior. We have demonstrated its utility on

unsupervised document hashing by outperforming current best results. We have also proposed the predictive document hashing task and showed that AMMI yields high-quality semantic representations. Future work includes extending AMMI to other structured models and extending it to cases in which even cross entropy calculation is intractable.

Acknowledgements

We thank David McAllester for many insightful discussions.

REFERENCES

- Alemi, A., Fischer, I., Dillon, J., and Murphy, K. Deep variational information bottleneck. In *ICLR*, 2017. URL <https://arxiv.org/abs/1612.00410>.
- Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15509–15519, 2019.
- Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540, Stockholm, Sweden, 10–15 Jul 2018. PMLR.
- Bell, A. J. and Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.
- Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.
- Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. Class-based n -gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.
- Chaidaroon, S., Ebesu, T., and Fang, Y. Deep semantic text hashing with weak supervision. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1109–1112, 2018.
- Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.
- Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S. Locality-sensitive hashing scheme based on p -stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pp. 253–262, 2004.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.
- Dong, W., Su, Q., Shen, D., and Chen, C. Document hashing with mixture-prior generative models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5229–5238, 2019.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- He, J., Spokoyny, D., Neubig, G., and Berg-Kirkpatrick, T. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, May 2019. URL <https://openreview.net/pdf?id=rylDfnCqF7>.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.
- Hinton, G. E. A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pp. 599–619. Springer, 2012.
- Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bklr3j0cKX>.
- Hwa, R. Sample selection for statistical grammar induction. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pp. 45–52. Association for Computational Linguistics, 2000.

- Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. URL <https://arxiv.org/abs/1611.01144>.
- Kaiser, L., Bengio, S., Roy, A., Vaswani, A., Parmar, N., Uszkoreit, J., and Shazeer, N. Fast decoding in sequence models using discrete latent variables. In *International Conference on Machine Learning*, pp. 2390–2399, 2018.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2014. URL <https://arxiv.org/abs/1312.6114>.
- Kudo, T. and Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, 2018.
- Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. URL <https://arxiv.org/abs/1611.00712>.
- McAllester, D. Information theoretic co-training. *arXiv preprint arXiv:1802.07572*, 2018.
- McAllester, D. and Stratos, K. Formal limitations on the measurement of mutual information. In *Artificial Intelligence and Statistics*, 2020.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- Mnih, A. and Gregor, K. Neural Variational Inference and Learning in Belief Networks. In *Proceedings of ICML*, 2014.
- Mnih, A. and Rezende, D. J. Variational Inference for Monte Carlo Objectives. In *Proceedings of ICML*, 2016.
- Onishi, T., Wang, H., Bansal, M., Gimpel, K., and McAllester, D. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2230–2235, 2016.
- Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, 2018.
- Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180, 2019.
- Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- Razavi, A., van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pp. 14837–14847, 2019.
- Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 1278–1286. PMLR, 2014.
- Shen, D., Su, Q., Chapfuwa, P., Wang, W., Wang, G., Henao, R., and Carin, L. Nash: Toward end-to-end neural architecture for generative semantic hashing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2041–2050, 2018.
- Stratos, K. Mutual information maximization for simple and accurate part-of-speech induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019.
- Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.
- van den Oord, A., Vinyals, O., et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pp. 6306–6315, 2017.
- Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.
- Weiss, Y., Torralba, A., and Fergus, R. Spectral hashing. In *Advances in neural information processing systems*, pp. 1753–1760, 2009.

Algorithm 3 Forward

Input: $p(z_i|i, z_{i-o:i-1})$ for $z_{i-o:i} \in \{0, 1\}^{o+1}$ and $i \in [m]$

Output: For $z_{i-o:i-1} \in \{0, 1\}^o$ and $i \leq m$

$$\pi(z_{i-o:i-1}|i) = \sum_{\bar{z} \in \{0,1\}^{i-1}: \bar{z}_{-j}=z_{i-j} \forall j \in [o]} p(\bar{z})$$

Runtime: $O(m2^o)$

Base: $\pi(z_{-o:-1}|i) \leftarrow [[z_{-o:-1} = 0^o]]$ for $i \leq 1$

Main: For $i = 2 \dots m$, for $z_{i-o:i-1} \in \{0, 1\}^o$,

$$\begin{aligned} \pi(z_{i-o:i-1}|i) &\leftarrow p(z_{i-1}|i-1, z^{(0)}) \times \pi(z^{(0)}|i-1) \\ &\quad + p(z_{i-1}|i-1, z^{(1)}) \times \pi(z^{(1)}|i-1) \end{aligned}$$

where $z^{(b)} = (b, z_{i-o}, \dots, z_{i-2}) \in \{0, 1\}^o$

Zhang, D., Wang, J., Cai, D., and Lu, J. Self-taught hashing for fast similarity search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 18–25, 2010.

A. Forward Algorithm

The forward algorithm is shown in Algorithm 3. We write $[m]$ to denote the set of m integers $\{1, \dots, m\}$ and $[[A]] = 1$ if A is true and 0 otherwise.

B. Dataset Construction for Predictive Document Hashing

We take pairs from the Who-Did-What dataset (Onishi et al., 2016). The pairs in this dataset were constructed by drawing articles from the LDC Gigaword newswire corpus. A first article is drawn at random and then a list of candidate second articles is drawn using the first sentence of the first article as an information retrieval query. A second article is selected from the candidates using criteria described in Onishi et al. (2016), the most significant of which is that the second article must have occurred within a two week time interval of the first. We filtered article pairs so that each article is distinct in all data. The resulting dataset has 104404, 8928, and 7326 article pairs for training, validation, and evaluation.

To follow the standard setting in unsupervised document hashing, we represent each article as a TFIDF vector using a vocabulary of size 20000. We use SentencePiece (Kudo & Richardson, 2018) to learn an optimal text tokenization for the target vocabulary size.