Topic-VQ-VAE: Leveraging Latent Codebooks for Flexible Topic-Guided Document Generation

YoungJoon Yoo[1], **Jongwon Choi**[2]
[1] ImageVision, NAVER Cloud.
[2] Department of Advanced Imaging (GSAIM) and Graduate School of AI, Chung-Ang University.
youngjoon.yoo@navercorp.com, choijw@cau.ac.kr

Abstract

This paper introduces a novel approach for topic modeling utilizing latent codebooks from Vector-Quantized Variational Auto-Encoder (VQ-VAE), discretely encapsulating the rich information of the pre-trained embeddings such as the pre-trained language model. From the novel interpretation of the latent codebooks and embeddings as conceptual bag-of-words, we propose a new generative topic model called Topic-VQ-VAE (TVQ-VAE) which inversely generates the original documents related to the respective latent codebook. The TVQ-VAE can visualize the topics with various generative distributions including the traditional BoW distribution and the autoregressive image generation. Our experimental results on document analysis and image generation demonstrate that TVQ-VAE effectively captures the topic context which reveals the underlying structures of the dataset and supports flexible forms of document generation. Official implementation of the proposed TVQ-VAE is available at https://github.com/clovaai/TVQ-VAE.

Introduction

Topic modeling, the process of extracting thematic structures, called **topic**, represented by coherent word sets and subsequently clustering and generating documents based on these topics, constitutes a foundational challenge in the manipulation of natural language data The initiative Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) and subsequent studies (Teh et al. 2004; Paisley et al. 2014) configure the inference process as a Bayesian framework by defining the probabilistic generation of the word, interpreted as bag-of-words (BoW), by the input word and document distributions. The Bayesian frameworks utilize the co-occurrence of the words in each document and have become a standard for topic models.

Despite the success, topic modeling has also faced demands for the evolution to reflect advances of recent deep generative studies. One main issue is utilizing information from large-scale datasets encapsulated in pre-trained embeddings (Pennington, Socher, and Manning 2014; Devlin et al. 2018; Radford et al. 2021). Many follow-up studies have approached the problem in generative (Dieng, Ruiz, and Blei 2020) or non-generative (Duan et al. 2021; Xu et al. 2022; Grootendorst 2022) directions. Moreover, with the advancements in generation methods, such as autoregressive and diffusion-based generation, there is a growing need for the topic-based generation to evolve beyond the traditional BoW form and become more flexible.

To address the issue, we propose a novel topic-driven generative model using Vector-Quantized (VQ) embeddings from (Van Den Oord, Vinyals et al. 2017), an essential building block for the recent vision-text generative model such as (Ramesh et al. 2021). In contrast to previous approaches in topic modeling (Gupta and Zhang 2021, 2023) that treat VQ embeddings as topics, in our method, each VQ embedding represents the embeddings of conceptually defined words. Through the distinct perspective, we achieve the enhanced flexibility that a corresponding codebook serves as its BoW representation. We further demonstrate that the codebook consisting of VQ embedding itself is an implicit topic learner and can be tuned to achieve exact topic context, with a supporting flexible format of sample generation.

Based on the interpretation, we present a novel generative topic model, Topic-VQ Variational Autoencoder (TVQ-VAE), which applies a VQ-VAE framework (Van Den Oord, Vinyals et al. 2017) incorporating topic extraction to the BoW representation of the VQ-embedding. The TVQ-VAE facilitates the generation of the BoW-style documents and also enables document generation in a general configuration, simultaneously. We demonstrate the efficacy of our proposed methodology in two distinct domains: (1) document clustering coupled with set-of-words style topic extraction, which poses a fundamental and well-established challenge in the field of topic modeling. For the pre-trained information, we utilize codebooks derived from inputs embedded with a Pre-trained Language Model (PLM) (Reimers and Gurevych 2019). Additionally, (2) we delve into the autoregressive image generation, leveraging the VQ-VAE framework with latent codebook sequence generation as delineated in (Van Den Oord, Kalchbrenner, and Kavukcuoglu 2016; Esser, Rombach, and Ommer 2021).

The contributions of the paper are summarized as follows:

- We propose a new generative topic modeling framework called **TVQ-VAE** utilizing codebooks of VQ embeddings and providing a flexible form of sampling. Our proposed model interprets the codebooks as a **conceptual** word and extracts the topic information from them.

- Our proposed model **TVQ-VAE** provides a general form of probabilistic methodology for topic-guided sampling. We demonstrate the application of samplings, from a typical histogram of the word style sample used in the topic model to an autoregressive image sampler.
- From the extensive analysis of two different data domains: (1) document clustering typically tackled by the previous topic models and (2) autoregressive image generation with topic extraction. The results support the proposed strength of the **TVQ-VAE**.

Preliminary

Key Components of Topic Model

We summarize the essence of the topic model where the generative or non-generative approaches commonly share as (1) semantic topic mining for entire documents and (2) document clustering given the discovered topics. Given K number of topics $\beta_k \in \boldsymbol{\beta}, k = 1, ..., K$, the topic model basically assigns the document to one of K topics, which is a clustering process given the topics. This assigning can be deterministic or generatively by defining the topic distribution of each document, as:

$$z_{dn} \sim p(z|\theta_d), \tag{1}$$

where the distribution $p(z|\theta_d)$ draws the indexing variable z_{dn} that denotes the topic index $\beta_{z_{dn}}$ that semantically includes the word w_{dn} in d'th document. In a generative setting, the random variable $\boldsymbol{\theta}$ is typically defined as K dimensional Categorical (Blei, Ng, and Jordan 2003) distribution with Dirichlet prior α or Product of Expert (PoE) (Srivastava and Sutton 2017). The topic β_k is defined as a set of semantically coherent words $w_{kn} \in \beta_k, 1, ..., N_w$ or by a word distribution in a generative manner, as:

$$w_k \sim p(w|\beta_k). \tag{2}$$

Similarly, the $p(w|\beta_k)$ can be defined as categorical (Blei, Ng, and Jordan 2003) like distributions. Classical probabilistic generative topic models (Blei, Ng, and Jordan 2003; Srivastava and Sutton 2017; Miao, Yu, and Blunsom 2016; Zhang et al. 2018; Nan et al. 2019) interpret each document d as BoW $\mathbf{w}_d = \{w_{d1}, ..., w_{dn}\}$ and analysis the joint distribution $p(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{w}_d)$ from equations (1-2), by approximated Bayesian inference methods (Casella and George 1992; Wainwright, Jordan et al. 2008; Kingma and Welling 2013). We note that their probabilistic framework reflects word co-occurrence tendency for each document.

When embedding is applied to the topic modeling frameworks (Dieng, Ruiz, and Blei 2020; Duan et al. 2021; Xu et al. 2022; Meng et al. 2022), some branches of embedded topic models preserve the word generation ability, and hence the word embedding is also included in their probabilistic framework, such as ETM (Dieng, Ruiz, and Blei 2020). The non-generative embedded topic models including recent PLM-based topic models (Sia, Dalmia, and Mielke 2020; Grootendorst 2022; Meng et al. 2022) extract topic embedding directly from distance-based clustering method, bypassing the complicated Bayesian inference approximation.

Vector Quantized Embedding

Different from the typical autoencoders mapping an input x to a continuous latent embedding space $\mathcal{E}$, Vector-Quantized Variational Auto-Encoder (VQ-VAE) (Van Den Oord, Vinyals et al. 2017) configures the embedding space to be discrete by the VQ embeddings $\boldsymbol{\varrho} = \{\rho_n \in \mathcal{R}^{D_\rho}, n = 1, ..., N_\rho\}$. Given the encoder function of the VQ-VAE as $f = Enc(x; W_E)$, the vector quantizer $(c_x, \rho_x) = Q(f)$ calculates the embedding $\rho_x \in \boldsymbol{\varrho}$, which is the closest embedding to f among the set of VQ embedding $\boldsymbol{\varrho}$, and its one-hot encoded codebook $c_x \in \mathcal{R}^{N_\rho}$. The embedding ρ_x and c_x is defined as:

$$\rho_x = c_x \cdot \hat{\boldsymbol{\rho}}, \quad \hat{\boldsymbol{\rho}} = [\rho_1, ..., \rho_{N_\rho}] \in \mathcal{R}^{N_\rho \times D_\rho}, \tag{3}$$

where N_ρ denotes the size of the discrete latent space, which is smaller than the original vocabulary size N_w. D_ρ is the dimensionality of each latent embedding vector. Here, we denote the resultant sets of embedding $\boldsymbol{\rho}$ and codebook $\boldsymbol{c}$ are defined as $\boldsymbol{\rho} = \{\rho_x\}$ and $\boldsymbol{c} = \{c_x\}$. When given an image $x \in \mathcal{R}^{H \times W \times 3}$ as a VQ-VAE input, we collect the sequence of quantized vector $\boldsymbol{\rho}$ and $\boldsymbol{c}$ as:

$$\begin{aligned} \boldsymbol{\rho} &= \{\rho_{ij} \in \boldsymbol{\varrho} | i = 1, ..., h, j = 1, ..., w\}, \\ \boldsymbol{c} &= \{c_{ij} \in \mathcal{R}^{N_\rho} | i = 1, ..., h, j = 1, ..., w\}, \end{aligned} \tag{4}$$

where the embedding ρ_{ij} and the codebook c_{ij} maps the closest encoding of the spatial feature f_{ij} of the latent variable $\boldsymbol{f} = \{f_{ij} | i = 1, ..., h, j = 1, ..., w\}, \boldsymbol{f} = Enc(x; W_E) \in \mathcal{R}^{h \times w \times d}$. The decoder function $\tilde{x} = Dec(\boldsymbol{c}, \boldsymbol{\rho}; W_D)$ then reconstruct the original image x using the VQ embedding $\boldsymbol{\rho}$ and its codebook $\boldsymbol{c}$. In this case, the vector quantizer $Q(\cdot)$ calculates the sequence of codebook $\boldsymbol{c}$ and the corresponding embeddings $\boldsymbol{\rho}$, as $(\boldsymbol{c}.\boldsymbol{\rho}) = Q(\boldsymbol{f})$.

Methodology

We present a new topic-driven generative model, TVQ-VAE, by first introducing a new interpretation to the VQ-VAE output: codebooks $\boldsymbol{c}$ and their embedding $\boldsymbol{\rho}$.

Vector Quantized Embedding as Conceptual Word

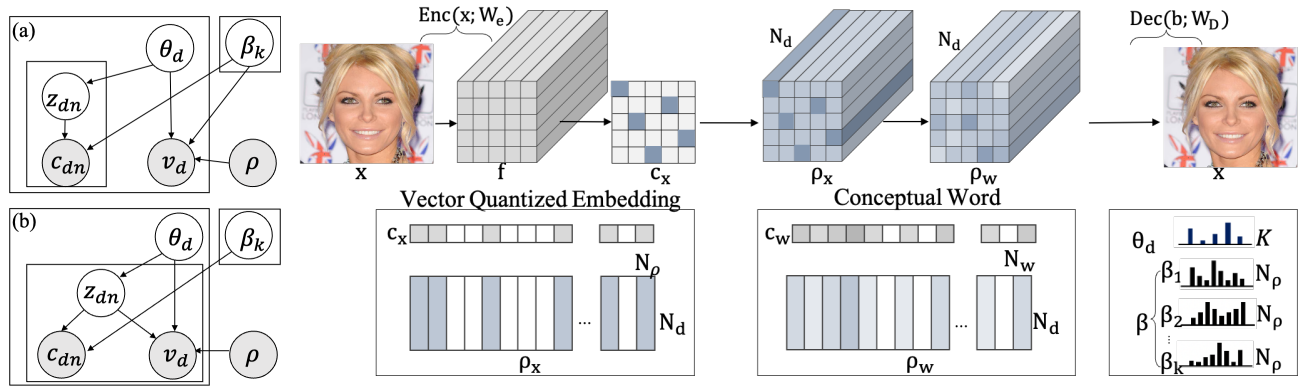
Here, we first propose a new perspective for interpreting a set $\boldsymbol{B}$ including the VQ embedding ρ and its codebook c:

$$\boldsymbol{B} = \{b_i = (c_i, \rho_i) | i = 1, ...N_\rho\}, \tag{5}$$

as **conceptual** word. The conceptual word b_i each consists of VQ embedding ρ_i and its codebook c_i.

One step further, since the typical selection of the number N_ρ is much smaller than the original vocabulary, we modify the set B so that multiple embeddings express the input, where the codebook c in Equation (3) becomes a multi-hot vector. This relaxation lets the codebooks deal with a much larger size of words. Specifically, given word w and its embedding $z_w = Enc(w)$ from the VQ-VAE encoder, we support the expansion from one-hot to multi-hot embedding by using K-nearest embeddings $\rho_1, ..., \rho_k$ from B to represent quantized embedding ρ_w for z_w as:

$$\begin{aligned} c_w &= \sum_k c_k, \\ \rho_w &= c_w \cdot \hat{\boldsymbol{\rho}}, \end{aligned} \tag{6}$$



(a) Graphical representation. (b) Visualized diagram of TVQ-VAE.

Figure 1: Graphical representation of the TVQ-VAE. Diagrams (a) illustrate the TVQ-VAE's graphical representation in both BoW (top) and General forms (bottom), while diagram (b) presents an example of VQ embedding, conceptual word, and output.

where the matrix $\hat{\boldsymbol{\rho}}$ denotes the encoding matrix in Equation (3). Using the expanded codebook c_w and its embedding ρ_w from equation (6), wwe define a expanded Bag-of-Word $\boldsymbol{B}_w$, the final form of the **conceptual** word, as follows:

$$\boldsymbol{B}_w = \{b_w = (c_w, \rho_w) | w = 1, ..., N_w\}. \tag{7}$$

We note that the multi-hot embedding $c_w \in \mathcal{R}^{N_\rho}$ is defined as N_ρ dimensional vector which is $N_w >> N_\rho$. Theoretically, the cardinality of $\boldsymbol{B}_w$ increases to combinatorial order $\binom{N_\rho}{K}$, where the number K called expansion value, denotes the number of assigned embeddings for each input.

Generative Formulation for TVQ-VAE

This section proposes a generative topic model called TVQ-VAE analyzing the **conceptual** words $\boldsymbol{B}_w$ in Equation (7). As illustrated in the graphical model in Figure 1, the TVQ-VAE model follows typical topic modeling structures formed by independent $d = 1, ..., D$ documents, and each document d has independent N_w words $c_w \equiv c_{dn} \in \mathcal{R}^{N_w}$, $n = 1, ..., N_w$. An output sample v_d is matched to a document d. TVQ-VAE provides various output forms for v_d. For the typical set-of-word style output, v_d is defined as a set of word $v_d = \{v_{d1}, ..., v_{dN_w}\}$ (Figure 1a), where the word $v_{dn} \in \mathcal{R}^{N_w}$ denotes the one-hot encoding of the original word w_{dn} corresponding to $c_{dn} \in \mathcal{R}^{N_\rho}$. Also, we can define v_d as an image corresponding to the document d (Figure 1a-(a)).

The joint distribution of the overall random variable $\{\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\rho}\}$ is formulated as:

$$\begin{aligned} &p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\rho}) \\ &= p_{pr} \prod_{d=1}^{D} p(v_d | \theta_d, \boldsymbol{\beta}, \boldsymbol{\rho}) \prod_{n=1}^{N_w} p(c_{dn} | \beta z_{dn}) p(z_{dn} | \theta_d), \end{aligned} \tag{8}$$

where the distribution $p_{pr} = p(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\rho})$ denotes the prior distribution for each independent random variable. The configuration in Equation (8) is a typical formulation for the generative topic model from (Blei, Ng, and Jordan 2003) or (Dieng, Ruiz, and Blei 2020), each defines $p(c|\beta_{z_{dn}})$ and $p(z_{dn}|\theta_d)$ to be **categorical** and **softmax** distribution. The main factor that discriminates the previous topic models to TVQ-VAE here is the generation of the output v_d from $p(v_d|\theta_d, \boldsymbol{\beta}, \boldsymbol{\rho})$.

As mentioned above, TVQ-VAE supports various forms of generation for output v_d. First, for the typical set-of-word style output $v_d = \{v_{d1}, ..., v_{dN_w}\}$, as in Figure 1a, the generation $p_g(v_d) \equiv p(v_d|\theta_d, \boldsymbol{\beta}, \boldsymbol{\rho})$ is defined as:

$$p_g(v_d) = \prod_{n=1}^{N_w} \sum_{z_{dn}=1}^{K} p(v_{dn} | \alpha(\beta_{z_{dn}} \cdot \hat{\boldsymbol{\rho}})) p(z_{dn} | \theta_d), \tag{9}$$

where a trainable fully connected layer $\alpha \in \mathcal{R}^{N_w \times N_\rho}$ connects the topic embedding $\beta_{z_{dn}} \cdot \hat{\boldsymbol{\rho}} \in \mathcal{R}^{N_\rho}$ to the original word dimension. Here, we define $p(v|\cdot)$ and $p(z_{dn}|\cdot)$ as **softmax** distribution, which is a PoE implementation of the topic model in (Srivastava and Sutton 2017). Note that we can marginalize out the indexing variable z_{dn} in equation (9) by computing all the possible samples from $p(z_{dn}|\theta_d)$.

For a more general case (Figure 1a-(b)), we assume the output v_d is generated by a sequence of codebook $\mathbf{c}_d = \{c_{dn} | n = 1, ..N_w\}$ and VQ-VAE decoder $v_d = Dec(\boldsymbol{c}_d, \boldsymbol{\rho}; W_D)$. To generate $\mathbf{c}_d$, we use AR prior $p_{ar}(\cdot)$ including PixelCNN and Transformer (Esser, Rombach, and

---

Algorithm 1: Pseudo-code of TVQ-VAE generation

Require: Given an topics $\boldsymbol{\beta} = \{\beta_1, ..., \beta_K\}$,
1: Sample or define θ_d.
2: **if** document analysis **then**
3: Sample $z_{dn} \sim p(z|\theta_d)$: $p(z|\cdot)$ be the softmax.
4: $v_{dn} \sim p(v|\alpha(\beta_{z_{dn}} \cdot \hat{\boldsymbol{\rho}}))$: $p(v|\cdot)$ be the softmax.
5: **Repeat** $n = 1, ..., N_w$
6: **else if** Image generation **then**
7: $\mathbf{c}' \sim \text{AR}(\boldsymbol{\theta} \cdot \hat{\boldsymbol{\beta}} \cdot \hat{\boldsymbol{\rho}})$.
8: $v = Dec(\mathbf{c}', \boldsymbol{\rho})$, $Dec(\cdot)$ be VQ-VAE decoder.
9: **end if**

Algorithm 2: Pseudo-code of TVQ-VAE training

Require: The batch of the input x_d and the output v_d.
if document analysis **then**
  x_d is the PLM vector from each Sentence.
  v_d be the histogram of the original word.
else if Image generation **then**
  $x_d \in \mathcal{R}^{H \times W \times 3}$ is an image.
end if
Initialize $\boldsymbol{\beta}$, γ_p.
$(\boldsymbol{\rho}, \boldsymbol{c}) = Q(Enc(x; W_E))$. (In equation (3-4) and (6)).
Calculate θ from $q(\theta|\gamma)$ (In equation (11)).
  $(\gamma_m, \log(\gamma_\sigma)) = NN(\mathbf{c}; W_\gamma)$.
  $\theta_d = Reparam(\gamma_m, \log(\gamma_\sigma))$.
if document analysis **then**
  $\boldsymbol{\beta} = \alpha(\theta_d \cdot \hat{\boldsymbol{\beta}} \cdot \hat{\boldsymbol{\rho}})$.
else if Image generation **then**
  $\mathbf{c}'$ = AR$(\theta_d \cdot \hat{\boldsymbol{\beta}} \cdot \hat{\boldsymbol{\rho}}; W_{ar})$.
end if
$l_{KL} = D_{KL}(\log(\gamma_\sigma), \gamma_m, \gamma_p)$.
$l_c = \mathbf{c} * \log(softmax(\theta_d \cdot \hat{\boldsymbol{\beta}}))$.
if document analysis **then**
  $l_v = v_d * \log(\beta)$.
else if Image generation **then**
  $l_v = CE(\mathbf{c}, \mathbf{c}')$.
end if
$l = l_{KL} + l_c + l_v$.

Ommer 2021), as:

$$p(v_d = Dec(\mathbf{c}_d, \boldsymbol{\rho}_d)|\theta_d, \boldsymbol{\beta}, \boldsymbol{\rho}) = P(\mathbf{c}_d|\theta_d \cdot \hat{\boldsymbol{\beta}} \cdot \hat{\boldsymbol{\rho}}) = \prod_{n=1}^{N} p_{ar}(c_{dn}|c_{dn-1}, ..., c_{d1}, \theta_d \cdot \hat{\boldsymbol{\beta}} \cdot \hat{\boldsymbol{\rho}}), \tag{10}$$

where the matrix $\hat{\boldsymbol{\beta}}$ denotes $\hat{\boldsymbol{\beta}} = [\beta_1, ..., \beta_K]$. We note that $Dec(\cdot)$ is a deterministic function, and the AR prior coupled with VQ-VAE decoding provides Negative Log Likelihood (NLL)-based convergence to the general data distributions. A detailed explanation of the algorithm is in Algorithm (1).

Training TVQ-VAE

For the model inference, we leverage autoencoding Variational Bayes (VB) (Kingma and Welling 2013) inference to the distribution in Equation (8) in a manner akin to (Srivastava and Sutton 2017; Dieng, Ruiz, and Blei 2020). In these methods, VB inference defines the variational distribution $q(\boldsymbol{\theta}, \boldsymbol{z}|\gamma, \phi)$ that can break the connection between $\boldsymbol{\theta}$ and $\boldsymbol{z}$, as $q(\boldsymbol{\theta}, \boldsymbol{z}|\gamma, \phi) = q(\boldsymbol{\theta}|\gamma)q(\boldsymbol{z}|\phi)$, of the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{z}|\mathbf{c}, \mathbf{v}, \beta, \rho)$. By the VB, the ELBO here is defined as:

$$L(\gamma) = - D_{KL}[q(\boldsymbol{\theta}|\gamma)||p(\boldsymbol{\theta})] + E_{q(\boldsymbol{\theta}|\gamma)}[\log p(\mathbf{c}, \mathbf{v}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\beta})], \tag{11}$$

where we pre-marginalize out the $\boldsymbol{z}$, similar to equation (9). In the equation, the first term measures the Kullbeck-Leibler (KL) distance between the variational posterior over the real posterior distribution, called **KL** term, and the second term denotes the **reconstruction** term. Followed by (Dieng, Ruiz, and Blei 2020), we define the variational parameter $\gamma = NN(\boldsymbol{c}; W_\gamma)$ as a neural network (NN) function of the input set-of-word $\boldsymbol{c}$, and $\boldsymbol{\theta}$ is drawn by a reparameterization technique given the variable γ.

Different from the previous methods (Srivastava and Sutton 2017; Dieng, Ruiz, and Blei 2020), we also consider the reconstruction of the output samples $\mathbf{v}$, as:

$$E_q[\log p(\mathbf{c}, \mathbf{v}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\beta})] = E_q[\log p(\mathbf{c}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\beta})] + E_q[\log p(\mathbf{v}|\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\beta})], \tag{12}$$

where $E_q[\cdot]$ integrates out $\gamma \sim q_\gamma(\theta)$. Here, $\mathbf{c}$ and $\mathbf{v}$ are conditionally independent given θ, as in Figure 1a. Therefore, the TVQ-VAE model has three loss terms corresponding to KL and the reconstruction terms:

$$l_{tot} = l_{KL}(\theta) + l_{rec}(\mathbf{c}) + l_{rec}(\mathbf{v}). \tag{13}$$

A detailed training process is given in Algorithm (2). See our official implementation for more explanation.

Related Works

Since the initiative generative topic modeling from (Blei, Ng, and Jordan 2003), many subsequent probabilistic methods (Teh et al. 2004; Paisley et al. 2014) have been proposed. After the proposal of autoencoding variational Bayes, a.k.a., variational autoencoder (VAE), from (Kingma and Welling 2013), neural-network-based topic models (NTMs) (Miao, Yu, and Blunsom 2016; Srivastava and Sutton 2017; Zhang et al. 2018; Nan et al. 2019) have been proposed. To reflect the discrete nature of the topic, (Gupta and Zhang 2021, 2023) introduces discrete inference of the topics by VQ-VAE (Van Den Oord, Vinyals et al. 2017). Unlike the above methods that treat each Vector Quantization (VQ) embedding as a distinct topic representation, our method leverages both the VQ embedding and its corresponding codebook as an expanded word feature, enabling extraction of a variable number of topics decoupled from the VQ embedding count.

Topic models with Embedding. Attempts to include word embeddings, mostly using GloVe (Pennington, Socher, and Manning 2014), into generative (Petterson et al. 2010; Dieng, Ruiz, and Blei 2020; Duan et al. 2021) or non-generative (Wang et al. 2022; Xu et al. 2022) topic modeling frameworks have also demonstrated successfully topic modeling performance. Moreover, utilizing pre-trained language models (PLMs) such as BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), and XLNet (Yang et al. 2019) has emerged as a new trend in mining topic models. Many recent studies have enhanced the modeling performance by observing the relation between K-means clusters and topic embeddings (Sia, Dalmia, and Mielke 2020). These studies require post-training steps including TF-IDF (Grootendorst 2022) or modifying of PLM embeddings to lie in a spherical embedding space through autoencoding (Meng et al. 2022) to mitigate the curse-of-dimensionality. Here, our method re-demonstrates the possibility of handling discretized PLM information in a generative manner without post-processing.

Vector Quantized Latent Embedding. Since (Van Den Oord, Vinyals et al. 2017) proposes a discretization

method for latent embedding incorporated with the variational autoencoding framework, this quantization technique has become an important block for generation, especially for visual generation (Razavi, Van den Oord, and Vinyals 2019). Following the study, subsequent studies (Peng et al. 2021; Esser, Rombach, and Ommer 2021; Yu et al. 2021; Hu et al. 2022) including text to image multi-modal connection (Gu et al. 2022; Tang et al. 2022; Esser et al. 2021) incorporated with autoregressive generation.

Empirical Analysis

We analyze the TVQ-VAE performance with two different applications: document analysis and image generation.

Document Analysis

Dataset. We conduct experiments on two datasets: 20 Newsgroups (**20NG**) (Lang 1995), the New York Times-annotated corpus (**NYT**) (Sandhaus 2008), as following the experiments of (Dieng, Ruiz, and Blei 2020). We present the detailed statistics of the datasets in Table 1. While documents in 20NG consist of about 46 words on average, we note that NYT is a much larger dataset to the 20NG dataset, 32K documents with 328 words per document on average.

Baseline Methods. To facilitate a comprehensive comparison, we select four representative topic models to encompass BoW-based, embedding-based, neural network-ignored, and neural-network employed approaches as well as generative and non-generative models, as: (1) **LDA** (Blei, Ng, and Jordan 2003) - a textbook method of BoW-based generative topic model, (2) **ProdLDA** (Srivastava and Sutton 2017) - a BoW-based generative neural topic model (NTM) (3) **ETM** (Dieng, Ruiz, and Blei 2020) - a generative NTM considering Word2Vec embedding (Petterson et al. 2010) as well, (4) **BerTopic** (Grootendorst 2022) - a non-generative PLM-based topic model utilizing sentence-Bert (Reimers and Gurevych 2019) information. We use the implementation from OCTIS (Terragni et al. 2021) for LDA, ProdLDA, and ETM. For ETM, we use Google's pre-trained Word2Vec as its embedding vector. For BerTopic, we use the official author's implementation. For **TVQ-VAE**, we set the embedding number and expansion k to 300 and 5.

Evaluation Metric. We evaluate the model's performance in terms of topic quality (TQ) and document representation, following the established evaluation setup for topic models. TQ is evaluated based on Topic Coherence (TC) and Topic Diversity (TD). TC is estimated by using Normalized Point-wise Mutual Information (NPMI) (Aletras and Stevenson 2013), quantifying the semantic coherence of the main words within each topic. NPMI scores range from -1 to 1, with higher values indicating better interpretability. TD measures word diversity by computing the unique word numbers among the top 25 words across all topics (Dieng, Ruiz, and Blei 2020). TD scores range from 0 to 1, with higher values indicating richer word diversity. TQ is defined as the multiplication of the TC, measured by NPMI, and TD values.

Furthermore, to measure document representation, we report the purity and Normalized Mutual Information (NMI) (Schutze, Manning, and Raghavan 2008). Following (Xu et al. 2022), we cluster the θ_d of every document d and measure the purity and NMI termed as **Km-NMI** and **Km-Purity**. Both values range from 0 to 1, and the higher values indicate better performance.

	# Doc.	# Vocab.	# Total words	# Labels
20NG	16.3K	1.60K	0.75M	20
NYT	32.0K	28.7K	10.5M	10 (9)

Table 1: Statistics of datasets. For 20NG, we follow the OCTIS setting from (Terragni et al. 2021). NYT dataset has two categories corresponding to locations (10) and topics (9).

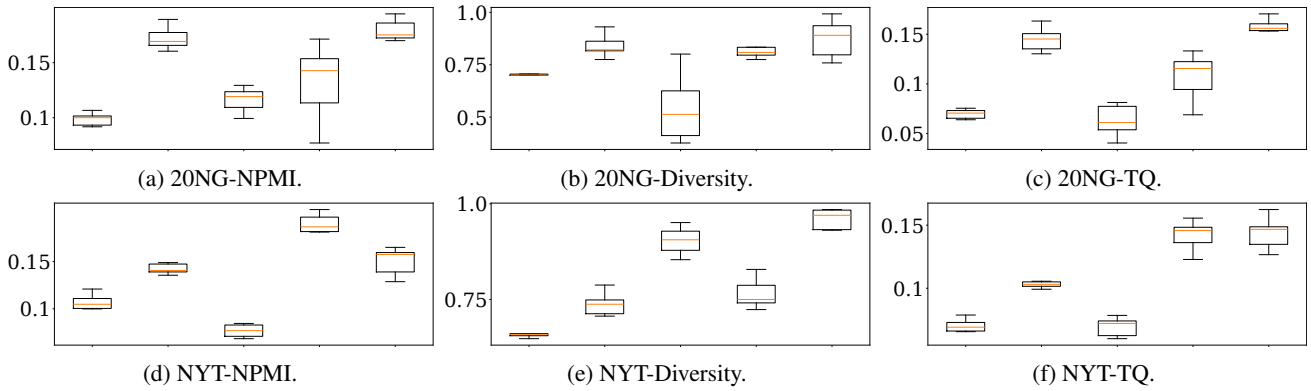
Topic Quality Evaluation. We present the evaluation results for topic quality (TQ), as depicted in Figure 2. From the evaluation settings outlined in (Grootendorst 2022), we infer a range of 10 to 50 topics with a step size of 10 and measure their TC and TD to evaluate TQ.

First, we evaluate the performance of TVQ-VAE on the 20NG dataset, which is widely used in the field of topic modeling. Notably, the TVQ-VAE demonstrates either comparable or superior performance compared to other baselines in terms of TQ measures. It is worth mentioning that the 20NG dataset has a small vocabulary size, which stands at $1.6K$. This scale is considerably smaller considering the number of TVQ-VAE codebook sizes. These results represent that TVQ-VAE effectively extracts topic information for documents with limited size, where BoW-based topic models like ProdLDA have exhibited impressive success.

In the NYT dataset, characterized by a significantly larger vocabulary to 20NG, the TVQ-VAE model achieves competitive topic quality when utilizing only 300 virtual codebooks, which accounts for less than 1% of the original vocabulary size. Among the baselines, BerTopic stands out as it demonstrates exceptional performance, particularly in terms of NPMI, deviating from the results observed in the 20NG dataset. The result verifies BerTopic's claim that PLM-based methods are scalable for larger vocabulary.

Figure 3 presents the ablation study conducted with varying the number of codebooks by $\{100, 200, 300\}$ and the expansion values by $k = \{1, 3, 5\}$. In the case of the 20NG dataset, the evaluation results indicate minimal performance differences across all settings. This presents that the choice of embedding and expansion numbers does not necessarily guarantee performance enhancements. This may happen due to the relatively small vocabulary size of 20NG, Moreover, exceeding certain bounds for the number of codebooks and expansion appears to capture no additional information from the original dataset. Conversely, the evaluation results obtained from the NYT dataset support our analysis. Here, the performance improves with larger codebook sizes and expansion numbers, given the vocabulary size of approximately 20 times that of the 20NG.

Document Representation Evaluation. Table 2 presents the km-NMI and km-Purity scores for each topic model. In the 20NG dataset, characterized by a relatively smaller vo-



(a) 20NG-NPMI. (b) 20NG-Diversity. (c) 20NG-TQ.

(d) NYT-NPMI. (e) NYT-Diversity. (f) NYT-TQ.

Figure 2: The quantitative evaluation of topic quality over two datasets: 20NG and NYT. The methods are listed from left to right: LDA, ProdLDA (PLDA), ETM, BerTopic, and TVQ-VAE.

	LDA	PLDA	ETM	BerTopic	TopClus	TVQ-VAE
20NG	(0.309/0.231)	(0.276/0.184)	(**0.331**/0.207)	(0.133/0.178)	(0.168/0.221)	(0.210/**0.242**)
NYT	(0.144/0.399)	(0.107/0.367)	(0.094/0.346)	(0.155/0.481)	(0.137/0.461)	(**0.184/0.510**)

Table 2: Evaluation on Km-NMI and Km-Purity on 20NG and NYT datasets: (Km-NMI / Km-Purity). We note that BerTopic, TopClus (Meng et al. 2022) and TVQ-VAE both use PLM (Reimers and Gurevych 2019).

cabulary size, the previous BoW-based method exhibited superior NMI scores. However, in the case of the NYT dataset, PLM-based methods like BerTopic and TVQ-VAE demonstrated higher performance. These findings suggest that our TVQ-VAE model exhibits robust document representation capabilities, particularly as the vocabulary size expands.

Image Generation

Dataset. To demonstrate that TVQ-VAE can mine topic information from the visual codebooks from VQ-VAE, we tested our method into two image datasets: CIFAR-10 (Krizhevsky, Hinton et al. 2009) and CelebA (Liu et al. 2015) typically used for supervised and unsupervised image generation, respectively. While CIFAR-10 contains $60K$ 32x32 dimensional images with 10 class objects, CelebA consists of about $200K$ of annotated facial images. We center-crop and resize the images to have 64x64 dimension. We convert the images to a sequence consisting of 64 and 256 codebooks, respectively, i.e., each image is represented as a document having 64 and 256 words.

Baseline Methods. Since the general form of document generation conditioned to a topic is a newly proposed task, it is difficult to directly compare to the previous methods. Quantitatively, therefore, we compare the TVQ-VAE to the baseline VQ-VAE generation guided by PixelCNN prior, which is a typical method of auto-regressive generation.

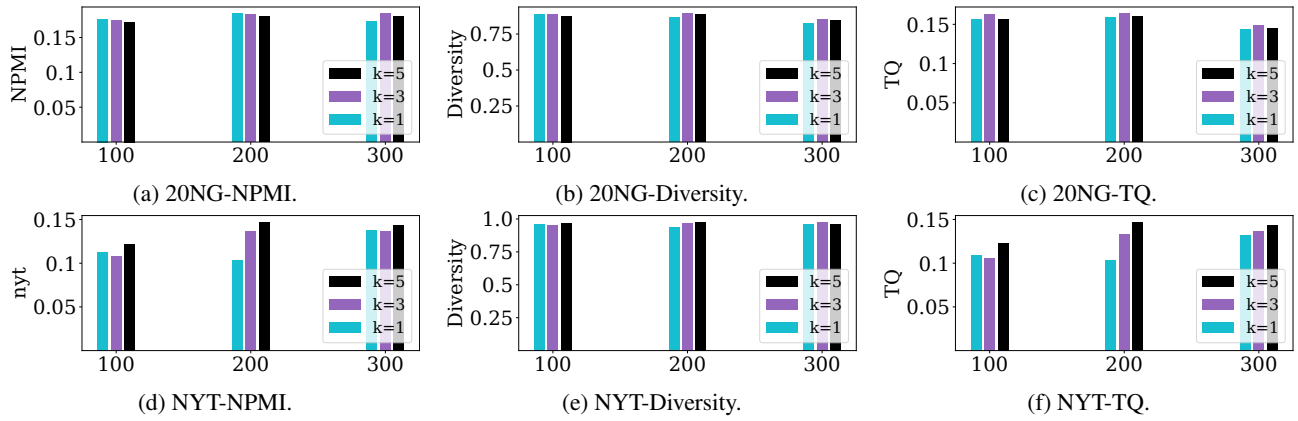
Evaluation. Regarding the quantitative evaluation, we utilize the Negative Log-Likelihood (NLL) metric on the test set, a widely adopted measure in the field of auto-regressive image generation. A lower NLL value means better coverage of the dataset. For the qualitative evaluation, we demonstrate the generated images corresponding to each topic, illustrating the topic's ability to serve as a semantic basis in shaping the generated data. Furthermore, we show image generation examples conditioned on a reference image by leveraging its topic information expressed as θ.

	U	10	20	50	100	S
CelebA	3.10	3.15	3.14	3.14	3.13	-
CIFAR-10	3.29	3.27	3.25	3.22	3.20	3.29

Table 3: NLL evaluation on CelebA and CIFAR-10 dataset. The terms 'U' and 'S' denote unsupervised and supervised generation from the VQ-VAE integrated with PixelCNN prior. The numbers $\{10, 20, 50, 100\}$ denote the number of topics assigned to TVQ-VAE.

Quantitative Evaluation. Table 3 presents the NLL evaluation results for the CelebA and CIFAR-10 datasets. We conjecture that the extraction of the topic variables θ and β helps the easier generation of the samples, quantified by lower NLL, since the topic variables already extract the hidden structures of the dataset which is originally the role of the generation module. The evaluations conducted on the CelebA and CIFAR-10 datasets yield contrasting outcomes. Specifically, in the case of CelebA, the unsupervised baseline exhibits a lower NLL. Conversely, for CIFAR-10, the NLL demonstrates a linear decrease with an increasing number of topics, surpassing the NLL values of both unsupervised and class-label supervised generation methods.

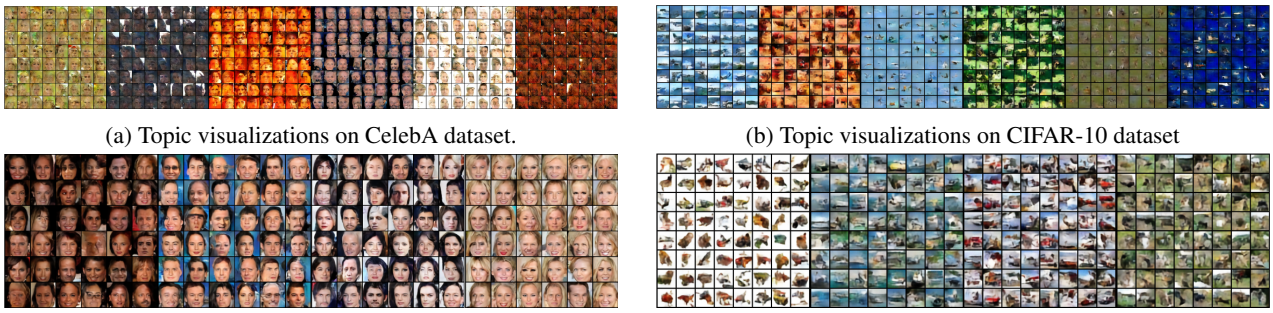
The complexity of the two datasets provides insights into the observed patterns. The CelebA dataset comprises aligned facial images, and the preprocessing step involves center-cropping the facial region to produce cropped im-



(a) 20NG-NPMI. (b) 20NG-Diversity. (c) 20NG-TQ.

(d) NYT-NPMI. (e) NYT-Diversity. (f) NYT-TQ.

Figure 3: Demonstration of the TQ over various numbers of codebook $\{100, 200, 300\}$ and expansion $k = \{1, 3, 5\}$.



(a) Topic visualizations on CelebA dataset. (b) Topic visualizations on CIFAR-10 dataset

(c) Reference-based generation on CelebA dataset. (d) Reference-based generation on CIFAR-10 dataset

Figure 4: Illustrations of visualized topics and reference-based generation for topic number K of 100. Best viewed on color.

ages that specifically include the eyes, nose, and mouth. This preprocessing step effectively reduces the dataset's complexity. In contrast, the CIFAR-10 dataset consists of unaligned images spanning ten distinct categories, resulting in an increased level of complexity. Previous evaluations from the baseline methods (Van Den Oord, Kalchbrenner, and Kavukcuoglu 2016; Van Den Oord, Vinyals et al. 2017) have similarly highlighted the challenging nature of NLL-based generation for CIFAR-10. Therefore, Table 3 supports our conjecture that topic extraction can enhance the model's generation capabilities for complicated datasets.

Qualitative Evaluation. Figure 4 shows visual examples of topics as well as generated samples obtained from reference images. The visualized topic examples in Figures 4a and 4b, arranged in an 8×8 grid, illustrate the generated samples obtained by fixing θ in Equation (10) to a one-hot vector corresponding to the topic indices. Subsequently, the PixelCNN prior $p_{pix}(\cdot|\boldsymbol{\theta} \cdot \hat{\boldsymbol{\beta}} \cdot \hat{\boldsymbol{\rho}})$ generates the codebook sequences by an auto-regressive scheme, conditioned on each k-th topic vector $\rho_{(\beta)} = \beta_k \cdot \hat{\boldsymbol{\rho}}$. The topic visualization clearly demonstrates that each topic exhibits distinct fundamental characteristics, such as color, shape, and contrast.

Furthermore, we demonstrate the generation ability of the TVQ-VAE by first, extracting the topic distribution θ_d of the image x_d, and subsequently generate new images from the extracted θ_d. In this case, we expect the newly generated images to share similar semantics to the original image x, which is called **reference-based generation**. As shown in Figures 4c and 4d, we generate images similar to the reference image, which is on the top-left corners each. The visual illustration for both CIFAR-10 and CelebA clearly demonstrates that TVQ-VAE effectively captures the distinctive attributes of reference images and generates semantically similar samples by leveraging the integrated topical basis.

Conclusion and Future Remark

We introduced TVQ-VAE, a novel generative topic model that utilizes discretized embeddings and codebooks from VQ-VAE, incorporating pre-trained information like PLM. Through comprehensive empirical analysis, we demonstrated the efficacy of TVQ-VAE in extracting topical information from a limited number of embeddings, enabling diverse probabilistic document generation from Bag-of-Words (BoW) style to autoregressively generated images. Experimental findings indicate that TVQ-VAE achieves comparable performance to state-of-the-art topic models while showcasing the potential for a more generalized topic-guided generation. Future research can explore the extension of this approach to recent developments in multi-modal generation.

Acknowledgements

We thank Jiyoon Lee[1] for the helpful discussion, experiments, and developments for the final published version. This research was supported by the Chung-Ang University Research Grants in 2023 and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) (2021-0-01341, Artificial Intelligence Graduate School Program (Chung-Ang Univ.)).

References

Aletras, N.; and Stevenson, M. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers*, 13–22.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Casella, G.; and George, E. I. 1992. Explaining the Gibbs sampler. *The American Statistician*, 46(3): 167–174.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dieng, A. B.; Ruiz, F. J.; and Blei, D. M. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8: 439–453.

Duan, Z.; Wang, D.; Chen, B.; Wang, C.; Chen, W.; Li, Y.; Ren, J.; and Zhou, M. 2021. Sawtooth factorial topic embeddings guided gamma belief network. In *International Conference on Machine Learning*, 2903–2913. PMLR.

Esser, P.; Rombach, R.; Blattmann, A.; and Ommer, B. 2021. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34: 3518–3532.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10696–10706.

Gupta, A.; and Zhang, Z. 2021. Vector-quantization-based topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(3): 1–30.

Gupta, A.; and Zhang, Z. 2023. Neural Topic Modeling via Discrete Variational Inference. *ACM Transactions on Intelligent Systems and Technology*, 14(2): 1–33.

[1]Independent researcher (jiyoon.lee52@gmail.com). The co-research was conducted during her internship at ImageVision, NAVER Cloud, in 2023.

Hu, M.; Wang, Y.; Cham, T.-J.; Yang, J.; and Suganthan, P. N. 2022. Global context with discrete diffusion in vector quantised modelling for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11502–11511.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, 331–339. Elsevier.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Meng, Y.; Zhang, Y.; Huang, J.; Zhang, Y.; and Han, J. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, 3143–3152.

Miao, Y.; Yu, L.; and Blunsom, P. 2016. Neural variational inference for text processing. In *International conference on machine learning*, 1727–1736. PMLR.

Nan, F.; Ding, R.; Nallapati, R.; and Xiang, B. 2019. Topic modeling with wasserstein autoencoders. *arXiv preprint arXiv:1907.12374*.

Paisley, J.; Wang, C.; Blei, D. M.; and Jordan, M. I. 2014. Nested hierarchical Dirichlet processes. *IEEE transactions on pattern analysis and machine intelligence*, 37(2): 256–270.

Peng, J.; Liu, D.; Xu, S.; and Li, H. 2021. Generating diverse structure for image inpainting with hierarchical VQ-VAE. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10775–10784.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Petterson, J.; Buntine, W.; Narayanamurthy, S.; Caetano, T.; and Smola, A. 2010. Word features for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Razavi, A.; Van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sandhaus, E. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12): e26752.

Schutze, H.; Manning, C. D.; and Raghavan, P. 2008. *Introduction to information retrieval*. Cambridge University Press.

Sia, S.; Dalmia, A.; and Mielke, S. J. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914*.

Srivastava, A.; and Sutton, C. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Tang, Z.; Gu, S.; Bao, J.; Chen, D.; and Wen, F. 2022. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*.

Teh, Y.; Jordan, M.; Beal, M.; and Blei, D. 2004. Sharing clusters among related groups: Hierarchical Dirichlet processes. *Advances in neural information processing systems*, 17.

Terragni, S.; Fersini, E.; Galuzzi, B. G.; Tropeano, P.; and Candelieri, A. 2021. Octis: comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 263–270.

Van Den Oord, A.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel recurrent neural networks. In *International conference on machine learning*, 1747–1756. PMLR.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wainwright, M. J.; Jordan, M. I.; et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2): 1–305.

Wang, D.; Guo, D.; Zhao, H.; Zheng, H.; Tanwisuth, K.; Chen, B.; and Zhou, M. 2022. Representing mixtures of word embeddings with mixtures of topic embeddings. *arXiv preprint arXiv:2203.01570*.

Xu, Y.; Wang, D.; Chen, B.; Lu, R.; Duan, Z.; Zhou, M.; et al. 2022. HyperMiner: Topic Taxonomy Mining with Hyperbolic Embedding. *Advances in Neural Information Processing Systems*, 35: 31557–31570.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yu, J.; Li, X.; Koh, J. Y.; Zhang, H.; Pang, R.; Qin, J.; Ku, A.; Xu, Y.; Baldridge, J.; and Wu, Y. 2021. Vector-quantized image modeling with improved VQGAN. *arXiv preprint arXiv:2110.04627*.

Zhang, H.; Chen, B.; Guo, D.; and Zhou, M. 2018. WHAI: Weibull hybrid autoencoding inference for deep topic modeling. *arXiv preprint arXiv:1803.01328*.