DISENTANGLING LATENT REPRESENTATIONS WITH CFACTORIZED VARIATIONAL AUTOENCODER

Prince Zizhuang Wang
Department of Computer Science
College of Creative Studies
UC Santa Barbara
zizhuang_wang@umail.ucsb.edu

ABSTRACT

One of the key aspects of humans' learning ability is that humans are able to break down a set of latent factors behind a specific event or a working task. Given an image of human face, humans are able to do a lot of learning and recognition tasks, such as recognizing the identity of the given face, observing the hair color, and summarizing the background behind the face. However, many computer vision systems fail to learn this set of meaningful latent factors that can guide machine to do multiple tasks. Recently, some generative models are proposed to address this problem. In this paper, we will focus on one major branch of these generative models, called variational autoencoder. We will compare the performance of some recently published models, and we propose a new model called CFactorVAE aiming at learning more interpretable latent representations. Finally, we tested the disentanglement of the learned features over MNIST by traversing each latent dimensions. We observed that our model is able to learn disentangled latent features while still produces realistic samples data whose reconstruction accuracy is better than the previously proposed baseline VAE models.

1 INTRODUCTION

Extracting a small set of salient features out of millions of latent factors behind a specific event of a phenomenon is a key element imperative to learning. Many feature learning models are trying to mimic the way that humans learn sets of meaningful features from complex daily life data. Convolutional neural network (Krizhevsky et al., 2012) is known as a good feature extractor, and some variants such as Residual network (He et al., 2016) and Mask RCNN (He et al., 2017) based on CNN backbone are able to handle large scale image classification and segmentation tasks. For unsupervised learning, Generative Adversarial Network (Goodfellow et al., 2014) and Variational Autoencoder (Kingma & Welling, 2013) are two popular generative models which can make use of CNN to produce realistic unseen data samples. One of the key aspects of unsupervised learning is the ability to learn meaningful latent factors behind data without any label information. Such latent factors can then serve as features for a lot of learning tasks. For example, in (Xu et al., 2017), a semi-supervised model is proposed to do classification task by first learning salient features with VAE unsupervisely, and then do the classification based on the learned feature representations. One of the current research trends in this two unsupervised learning approaches is then how to learn more meaningful latent features which are not only useful for a lot of semi-supervised learning tasks and are able to generate realistic unseen dataset, but are also interpretable and able to capture the most salient latent factor of variations behind the given data.

The most important difference between GAN and VAE is that GAN lacks the model to extract features from input data. In GAN perspective, a generator able to produce realistic unseen samples is learned with a discriminator in order to model a posterior distribution of samples given randomly generated feature vectors. Hence, GAN is not able to infer the specific feature correlated with the samples, which makes learning based on latent features really difficult. On the other hand, VAE addresses this problem by imposing an encoder model which learns latent factors from the samples, and unseen samples can then be generated by a decoder. Given VAE's ability to model latent features explicitly, the question then becomes how we can make the learned representations

more interpretable from human perspective, and similarly how we can the learned features better for tasks of generating new samples and tasks of semi-supervised classification.

Recently, a lot of works on learning such interpretable latent features try to define a general metric which can measure how good a learned feature is. There are many debates on what kinds of latent features are good, and what kinds of latent representation should be defined as interpretable, and how we can use this metric to find more meaningful latent representations. Among all these recent works, the task of learning latent representations that are highly disentangled draws a lot of attention. The problem of learning disentangled representations can be stated as a task of learning sets of latent variables that is mutually statistically independent, that is, each unit of latent variables only accounts for one factor of variation of the sample data. For example, when doing the task of facial recognition, humans are good at extracting the most important latent factors describing the given faces and backgrounds. This is because humans are sensitive to the changes of sensory input that are salient to the learning task. For example, the length of hair, the color of eyes, viewpoint of the observer, emotion of the given face, variety of the background, are all latent factors behind the data that are crucial for determining the identity of the given face. The ideal latent features can then be defined as the ones which can be divided into a set of individual dimensions, each of which then accounts for only one latent factor, such as color, viewpoint, and emotion. Given VAE's advantage of encoding input data into learned feature space, a lot of recent works (Higgins et al., 2016) (Burgess et al., 2018) (Kim & Mnih, 2018) proposed variants of VAE models to learn disentangled representations. For example, β-VAE proposed in (Higgins et al., 2016) modified the objective function of the original VAE model to force the learned latent variables to be more statistically independent. Unfortunately, some experimental results showed (Higgins et al., 2016) there exists trade-off between level of disentanglement and reconstruction accuracy. To address this problem, many more works try to modify the VAE objective further in order to disentangle learned latent variable while at the same produce realistic samples that are of the same quality of the ones generated by original VAE model. In (Burgess et al., 2018), the author shows that trade-off between disentanglement and reconstruction accuracy can be relaxed by introducing a controllable coefficient into the β-VAE objective in order to manually control the amount of mutual information between the input samples and the learned features. Further, in (Kim & Mnih, 2018), a new objective is derived with the notion of total correlation among learned latent variables and make use of a discriminate model to help compute this new objective. In this paper, we will briefly introduce the existing models and compare their performance. And then, we will propose our new VAE objective aiming at the task of learning disentangled latent representations. In the experiment sections, we will show that our model outperform the existing VAEs in terms of reconstruction accuracy. The results show that our model is able to disentangle learned representation while at the same time produce realistic samples with good quality.

2 VARIATIONAL AUTOENCODER

A variational autoencoder jointly trains an encoder model $q_\phi(z|x)$ and a decoder model $p_\theta(x|z)$. The objective is to the maximize the marginal likelihood of given data samples,

$$\max_{\phi,\theta} E_{q_\phi(z|x)}[\log p_\theta(x|z)] \tag{1}$$

where ϕ abd θ are the parameters of encoder and decoder respectively. This marginal likelihood can be rewritten as,

$$\log p_\theta(x|z) = D_{KL}(q_\phi(z|x)||p(z)) + \mathcal{L}(\phi,\theta) \tag{2}$$

where $p(z)$ is th prior assumption about latent variables. Therefore, since Kullback Leibler divergence term D_{KL} is always non-negative, as explained in (Kingma & Welling, 2013), maximizing this marginal likelihood is equivalent to maximizing its evidence lower bound(ELBO), which is,

$$\log p_\theta(x|z) \geq \mathcal{L}(\phi,\theta) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)) \tag{3}$$

In experiments, instead of maximizing this ELBO, we usually minimizing the negative of $\mathcal{L}$, that is

$$\min_{\phi,\theta} \mathcal{L}(\phi,\theta) = -E_{q_\phi(z|x)}[\log p_\theta(x|z)] + D_{KL}(q_\phi(z|x)||p(z)) \tag{4}$$

The negative marginal likelihood $-E_{q_\phi(z|x)}[\log p_\theta(x|z)]$ can then be interpreted as the reconstruction error, and is often computed by cross entropy loss.

Since we usually choose isotropic unit Gaussian distribution as our prior assumption towards latent variables, that is, $p(z) = N(0, I)$, the encoder model $q_\phi(z|x)$ is also chosen as a Gaussian distribution $N(u, \sigma)$. By reparameterization trick proposed by (Kingma & Welling, 2013), we can write latent variable z as,

$$z = \mu + \sigma\epsilon$$

where $\epsilon \sim N(0, I)$, as it is like z is sampled from a Gaussian distribution $N(0, I)$. In this way, we are able to differentiate z with respect to parameters ϕ and θ and then send the gradients to update the parametric model $p_\theta(x|z)$ and $q_\phi(z|x)$. Note that this reparameterization trick can also be generalized to other more complex distributions. The only modification would be to introduce more parameters to rewrite latent variable z as a combination of them.

3 DISENTANGLING WITH VAE

In this section, we will briefly review the approach in (Higgins et al., 2016) (Burgess et al., 2018) (Kim & Mnih, 2018) using variational autoencoder to learn disentangled latent representations, and discuss some advantages and potential drawbacks of these models.

3.1 β-VAE

In the original VAE objective, the Kullback Leilber divergence term $D_{KL}(q_\phi(z|x)||p(z))$ serves as a regularizer preventing the model from overfitting by forcing the posterior distribution $q_\phi(z|x)$ that encodes input samples into corresponding latent variables to be as simple as an isotropic unit Gaussian prior $p(z)$. Since the prior is chosen to be an isotropic unit Gaussian whose covariance matrix is diagonal, each unit dimension of latent random variable is also assumed to be statistically independent. Hence, the D_{KL} term also makes the encoder model to learn more disentangled latent variables from sample data.

To force the encoder to be closer to isotropic unit prior, (Higgins et al., 2016) introduced β-VAE with the use of a hyperparameter β that further penalize the Kullback Leiber divergence between the posterior and the prior. The modified objective is,

$$\mathcal{L}(\phi,\theta) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x)||p(z)) \tag{5}$$

When $\beta = 1$, it becomes the original VAE framework. With larger value of β ($\beta > 1$), the above objective imposes stronger pressure for the posterior $q_\phi(z|x)$ to match the factorized unit Gaussian prior $p(z)$, which has been suggested to be crucial for learning disentangled latent representation z.

It is often observed that increasing the value of β will also result in a drop of the quality of reconstructed samples generated by decoder model $p_\theta(x|z)$ of VAE framework. To analyze this trade-off between reconstruction quality and disentanglement of latent representation, we can further decompose the mean of D_{KL} term which we want to penalize into a combination of Kullback Leiber divergence between the marginal and prior of latent representation and the mutual information between input samples and the corresponding latent variables. That is,

$$\begin{aligned} E_{p(x)}[D_{KL}(q_\phi(z|x)||p(z))] &= E_{p(x)}[D_{KL}(q_\phi(z|x)||q(z))] + E_{q(x,z)}[\log \frac{q(z)}{p(z)}] & (6)\\ &= I_q(x;z) + E_{q(z)}[\log \frac{q(z)}{p(z)}] & (7)\\ &= I_q(x;z) + D_{KL}(q(z)||p(z)) & (8) \end{aligned}$$

where $I_q(x;z)$ stands for the mutual information between input samples x and its latent representation z that is encoded by encoder posterior model $q_\phi(z|x)$. Obviously, penalizing

$E_{p(x)}[D_{KL}(q_\phi(z|x)||p(z))]$ more will result in a tremendous drop in mutual information that can be encoded by q_ϕ, which then results in more blurry reconstructed samples to be produced by generative model p_θ, as now the latent representation z that is fed into the generative model can say nothing about the original sample x if there is low mutual information between them.

3.2 CONTROLLABLE β-VAE

To make q_ϕ match to the prior $p(z)$ by penalizing $D_{KL}(q_\phi(z|x)||p(z))$ without sacrificing two much mutual information between samples and latent representations that can be encoded by q_ϕ, (Burgess et al., 2018) introduced another parameter C into β-VAE objective,

$$\mathcal{L}(\phi, \theta) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta|D_{KL}(q_\phi(z|x)||p(z)) - C| \tag{9}$$

where C is the controlled capacity we imposed over the encoder model, and is gradually increased to a certain level during training. The intuitive thougth behind introducing this new parameter C is not we don't want to penalize $D_{KL}(q_\phi(z|x)||p(z))$ too much at the later stage of training, so that we can prevent the encoding capacity of the encoder model from collapsing. Therefore, at the early stage of training, we want $D_{KL}(q_\phi(z|x)||p(z))$ to be small for learning disentangled representations z, and then, when $D_{KL}(q_\phi(z|x)||p(z))$ is small enough at the later stage of training, we want to refrain it from dropping to zero which makes the model learn random representations regardless of input samples.

3.3 FACTORVAE

(Kim & Mnih, 2018) proposed another solution for relaxing the trade-off between reconstruction quality and level of disentanglement for learned representations by further modifying the original β-VAE objective. Instead of penalizing $D_{KL}(q_\phi(z|x)||p(z))$, which eventually results in loss of mutual information, (Kim & Mnih, 2018) propose to add an additional total correlation term into the objective,

$$\mathcal{L}(\phi, \theta) = \frac{1}{N}\sum[E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z))] - \gamma D_{KL}(q(z)||\prod q(z_j)) \tag{10}$$

where the last term stands for the total correlation for random variables z. It is suggested that by minimizing the total correlation directly also result in disentangled latent representations, while still maintain the mutual information between samples and latent representations. To approximate this Total Correlation term, (Kim & Mnih, 2018) used a discriminator to compute the probability for z to be sampled from the non-factorised distribution $q(z)$, and then minimize the following approximation,

$$D_{KL}(q(z)||\prod q(z_j)) = E_{q(z)}[\log \frac{D(z)}{1 - D(z)}] \tag{11}$$

4 OUR APPROACH: CFACTORVAE

We propose a new VAE learning objective which incorporates FactorVAE with Controlled-VAE to take the advantages they have. The new learning objective allows the model to gradually increase its capacity of encoding meaningful latent representations and at the same time factorized the learned representations. We combine FactorVAE and Controlled-VAE to propose the following evidence lower bound to maximize,

$$\mathcal{L}(\phi, \theta) = \frac{1}{N}\sum[E_{q_\phi(z|x)}[\log p_\theta(x|z)] - |D_{KL}(q_\phi(z|x)||p(z)) - C|] - \gamma D_{KL}(q(z)||\prod q(z_j))$$

(12)

where C is gradually increased during training. Like Controlled-VAE, increasing C during training increase the amount of information that can be encoded about samples and latent representations. However, in Controlled-VAE framework, at the later stage of training the model loses its ability to further disentangle learned representation because a sufficiently large C prevent the Kullback

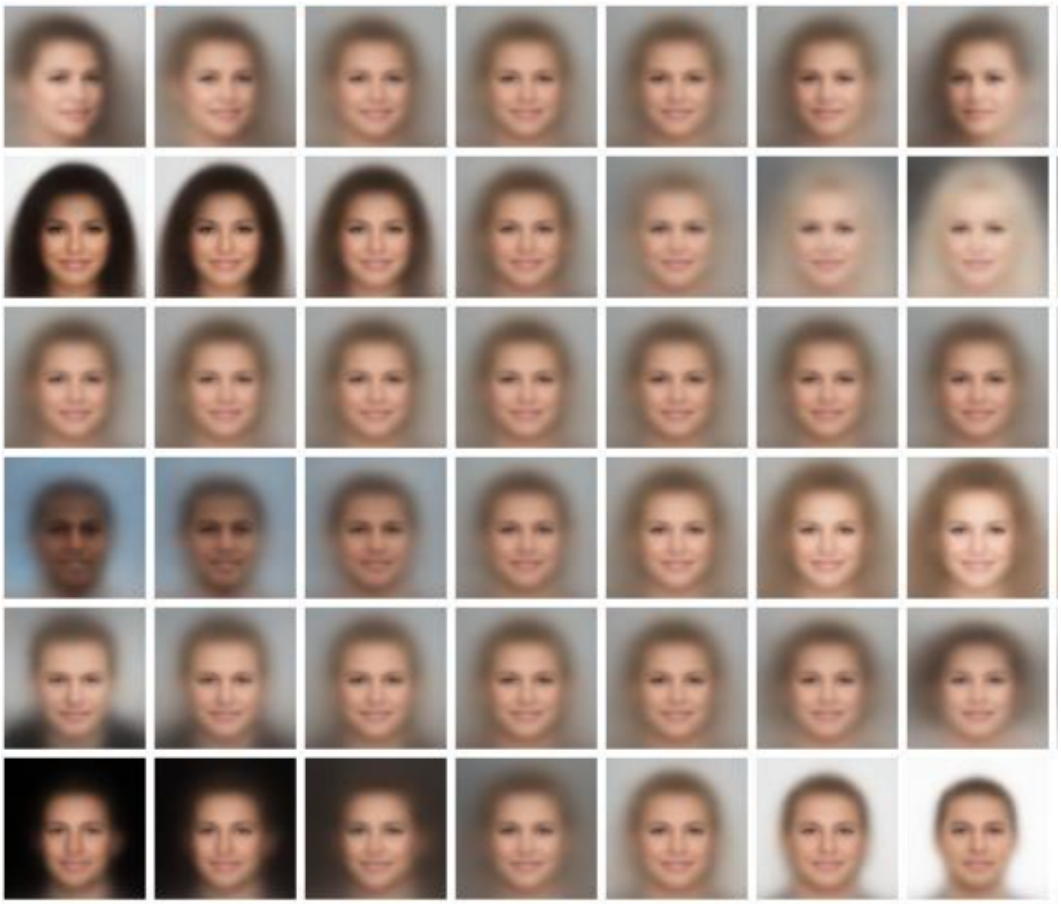


Figure 1: Latent traversal over **celebA** representations learned by FactorVAE
The latent representations is high disentangled and factorized, as each unit of latent dimension corresponds to only one salient factor of variations, such as, background color, hair color, and azimuth.

Leibler divergence between the posterior distribution and the prior distribution of latent representation z. Hence, inspired by FactorVAE, we add an additional Total Correlation term to take the job of factorizing learned latent representations z. Restricting $D_{KL}(q_\phi(z|x)||p(z))$ from going to zero in order for our model to encode more information will then have no effect on disentangling latent representations, as minimizing the Total correlation is sufficiently enough to learn factorized latent variables and it has nothing to do with the controlled capacity C. Like FactorVAE, we use a discriminator to approximate the Total Correlation term, which is learned jointly with VAE framework during training.

To show that our model also learns factorized latent representations, we tested it on MNIST and 3D-Chairs dataset by using latent variable traversals. We observed from the generative samples that each unit variable corresponds to one salient factor of variations behind the data, which then supports our assumption that it can learn disentangled latent representations just like previously proposed VAE models. Then, to measure the quality of reconstructed unseen samples from randomly generative latent variables, that is, to test our model's capacity of encoding salient mutual information between the original training samples and latent representations, we compared our model's reconstruction error with the one produced by all other existing models. We observed that our model outperforms the current state-of-art result in terms of reconstruction quality.

(a) FactorVAE

(b) CFactorVAE

Figure 2: Latent traversal over **chairs** representations learned by FactorVAE and CFactorVAE

5 EXPERIMENTS

5.1 LATENT TRAVERSAL

We first implemented FactorVAE in (Kim & Mnih, 2018). We used convolutional neural network as backbone for both encoder and decoder. For MNIST, we transform the original dataset consisting of $28x28$ images into a set of images with size $32x32$. The encoder has three convolutional layers, each of which has the same kernel size, stride size, and padding, which maps the original $32x32$ digits images into a feature maps of size $(64, 4, 4)$, where there are 64 channels of maps, each of which is of size $(4, 4)$. Then we used two fully connected layers to map the features into latent representations z of 20 dimensions. Note the dimension of latent representations can be changed to other values. For Chairs dataset, we first do a grayscale operation to transform the RGB format images into gray images of only one channel, then we rescale each image to size $(64, 64)$. We add an additional convolutional layer in the encoder to match up the feature dimension. The corresponding decoder has a similar structure.

To get a taste of how good the learned disentangled latent representations are, we did latent traversal over each individual latent variable dimension from range $(-3, 3)$. The idea is that if each latent variable unit is statistically independent, then when we do a latent traversal over one unit dimension, while keeping other unit latent variables fixed, we should expect to observe changes only in one salient factor of variation behind the data.

Table 1: Reconstruction quality

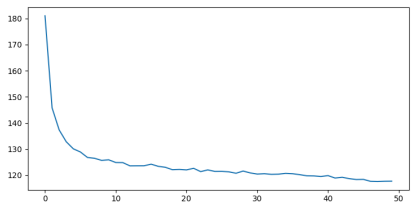
Performance on generating new samples				
Model	Train loss	Train recon loss	Test loss	Test recon error
CVAE	125.75	124.89	21.13	21.0
FVAE	141.12	121.14	23.73	20.48
CFVAE	118.69	118.02	20.01	19.92

We then tested latent traversal on our CFactorVAE model, which is built using the same encoder and decoder structure as in FactorVAE. We see that the learned latent representations by our model is also disentangled.

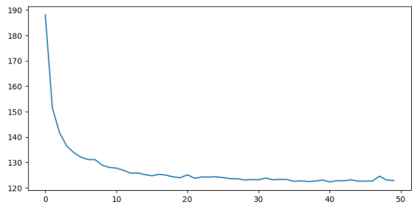
5.2 CFactorVAE: better Reconstruction quality

Learning disentangled latent representations is not the only goal. The learned features are useless not matter how disentangled they, if they cannot produce realistic samples with decent reconstruction quality.

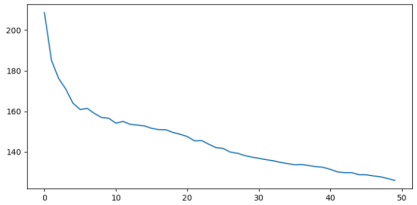
As explained in previous section, β-VAE and its variants are not able to address the trade-off between reconstruction quality and disentanglement for latent representations. FactorVAE, as mentioned, has been suggested to be successful in relaxing this trade-off, and is able to learn very factorized latent variables while producing new unseen data samples with better quality then previous models. To see whether our model can further improve this reconstruction quality, we compare its performance with FactorVAE on MNIST dataset. Based on our results, we see that our model has a tremendous drop not only training loss and reconstruction loss, but also has lower reconstruction error during the evaluation stage in which we generate unseen samples based on randomly generated latent representations. We can therefore conclude that our model can generate more realistic data samples than previous β-VAE like frameworks such as FactorVAE and Controlled-VAE.



(a) cfvae



(b) fvae



(c) cvae

Figure 3: Reconstruction error after training 50 epochs:
(a) our CFactor VAE model (b) FactorVAE (C) Controlled VAE

6 Conclusion

In this paper, we proposed a new VAE learning objective that is designed for learning highly disentangled representations without much sacrifice for reconstruction quality for the task of generating new unseen data samples. We justified that this objective allows us to manually control the amount

Figure 4: Reconstructed samples generated by CFactorVAE from random latent variables

of information between sample and its latent representation which will not be affected by the regularization process of penalizing the marginal distribution of latent representation to be as simple as an isotropic, unit factorized Gaussian distribution. We compared our model's performance with FactorVAE, and we observed that our model is able to learn latent features as disentangled as the ones learned by FactorVAE. Furthermore, we observed that our model can produce new samples with better reconstruction quality then the samples generated by FactorVAE. We thus can conclude that our model can produce realistic samples while at same time learn highly disentangled latent representations.

7 Model Architecture

Table 2: CFactorVAE

Encoder $\mathbf{q}_\phi$	**Decoder $\mathbf{p}_\theta$**
32 Conv 4x4, stride 2, padding 1	Linear latent dimensions $\Rightarrow 256$
32 Conv 4x4, stride 2, padding 1	Linear 256 $\Rightarrow 64 * 4 * 4$
64 conv 4x4, stride 2, padding 1	64 ConvTran 4x4, stride 2, padding 1
Linear 64*4*4 $\Rightarrow 256$	32 ConvTran 4x4, stride 2, padding 1
Linear 256 $\Rightarrow$ latent dimensions	32 ConvTran 4x4, stride 2, padding 1

References

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in $\backslash$ beta $-vae.$ *arXiv preprint arXiv:1804.03599*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2980–2988. IEEE, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. Variational autoencoder for semi-supervised text classification. In *AAAI*, pp. 3358–3364, 2017.