Semi Supervised Learning using Deep Hybrid Models

Abhinav Garg
14013

Gopichand Kotana
14249

Jayant Agrawal
14282

1 Introduction

Most of the machine learning techniques require large amounts of labeled training data. However, obtaining these large labeled datasets is infeasible for many real world problems. Semi-Supervised Learning as a paradigm, focuses on such problems where unlabeled data, often available in abundance, is used to improve the performance of machine learning methods. Generative and Discriminative are the two broad categories, SSL methods are generally grouped in.

Generative methods offers better regularization and can deal effectively with missing data, while discriminative methods boast of higher predictive accuracy. We aim to bring these two together in a unified framework to get advantages from both of them to tackle problems in a semi-supervised setting.

In this project, we propose two novel models. One, we propose a novel hybrid model based on gaussian mixture prior(Section 3) for semi-supervised learning. We show that our model is almost as competitive as the state-of-the-art for the semi-supervised learning for MNIST.

Second, we propose a deep hybrid model to do semi-supervised text classification. For this, we studied various models like VAEs for text, deep models for text classification and, we propose a model which combines a state of the art architecture for text classification and state of the art VAE architecture for text to achieve our objective. We describe our model in Section 4.

2 Previous Work

Hybrid Models were first proposed in [1] and [2], where they simply trained their model by using a convex combination of generative and discriminative log likelihoods as their objective function. *Multi-Conditional Likelihood*, proposed in [3]. also uses a weighted combination of posterior over y given x, and the marginal over x as:

$$\alpha \log p(y|x, \theta) + \beta \log p(x, \theta)$$

where θ is the model parameter and α, β are scalar weights. Section 2.1 and Section 2.2 cover hybrid models briefly. We then describe Gaussian Mixture VAEs in Section 2.3, which serves as an inspiration for one of our approaches**??** to extend the work on Deep Hybrid Models. We also plan to use Deep hybrid models and analyze the feasibility of these models for Semi Supervised Learning for text. Some recent works on Semi Supervised Learning for text have been briefly described in Section 2.4 and Section 2.5.

2.1 Principled Hybrids [6]

Lasserre et.al in [6], uses an extra set of parameters, to come up with a joint Bayesian model where the generative and discriminative likelihoods are conditioned on different parameters as:

$$p(x, y, \theta_d, \theta_g) = p_{\theta_d}(y|x)p_{\theta_g}(x)p(\theta_d, \theta_g)$$

where $p(\theta_d, \theta_g)$ is a parameter coupling prior. The coupling prior is used to interpolate between generative and discriminative methods to extract advantages from both approaches.

2.2 Deep Hybrid Models [7]

Deep Models are capable of modeling very complex functions, thus making them the default choice for many complex machine learning problems. *Kuleshov et.al* in [7] use deep models such as variational auto-encoder[4] and generative adversarial networks[8] for modeling $p(x, z)$, along with CNNs to model $p(y|x, z)$, both sharing the same latent z. They use z as the common link between the generative and discriminative approaches as shown in Figure 1.

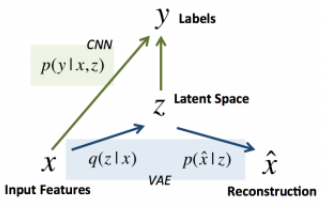

Figure 1: Deep Hybrid Model, Image Credits: [7]

2.3 GMVAE [9]

Generally, regular VAEs use the normal distribution as a prior over the latent variables. Though this serves as a good way to obtain structured and regularized latent embeddings, this choice of prior is rather too simple and does not allow for more complex representations. As an extension to the normal prior, *Nat et al.* [9] use mixture of Gaussians instead.

The model used in [9] is the following, where y is generated from x, w and z as in Figure 2. They train the model using the ELBO objective in Figure 3. The objective has four parts:

$$\begin{aligned}
\boldsymbol{w} &\sim \mathcal{N}(0, \boldsymbol{I}) \\
\boldsymbol{z} &\sim Mult(\boldsymbol{\pi}) \\
\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{w} &\sim \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_{z_k}(\boldsymbol{w}; \beta), diag\left(\boldsymbol{\sigma}^2_{z_k}(\boldsymbol{w}; \beta)\right)\right)^{z_k} \\
\boldsymbol{y}|\boldsymbol{x} &\sim \mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{x}; \theta), diag\left(\boldsymbol{\sigma}^2(\boldsymbol{x}; \theta)\right)\right) \text{ or } \mathcal{B}(\boldsymbol{\mu}(\boldsymbol{x}; \theta))
\end{aligned}$$

Figure 2: Generative Process, Figure: [9]

$$\begin{aligned}
\mathcal{L}_{ELBO} = \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{y})}\left[\log p_\theta(\boldsymbol{y}|\boldsymbol{x})\right] - \mathbb{E}_{q(\boldsymbol{w}|\boldsymbol{y})p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{w})}\left[KL(q_{\phi_x}(\boldsymbol{x}|\boldsymbol{y})||p_\beta(\boldsymbol{x}|\boldsymbol{w}, \boldsymbol{z}))\right] \\
- KL(q_{\phi_w}(\boldsymbol{w}|\boldsymbol{y})||p(\boldsymbol{w})) - \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{w}|\boldsymbol{y})}\left[KL(p_\beta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{w})||p(\boldsymbol{z}))\right].
\end{aligned}$$

Figure 3: ELBO Objective, Figure: [9]

reconstruction error, conditional prior, w-prior and z-prior.

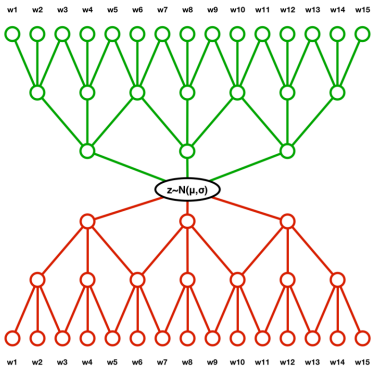

Figure 4: Feed forward component [15]

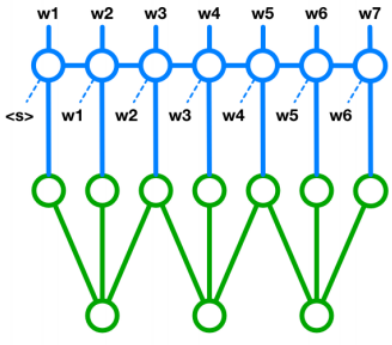

Figure 5: LSTM decoder [15]

2.4 Semi-supervised learning algorithms for text

Recently, it has been shown that we can use virtual adversarial training in the text domain[11]. The method proposed in [11] gives state of the art results on various standard data sets.
Virtual adversarial training allows us to do adversarial training in semi-supervised and unsupervised settings. [12] extends it for text classification; During adversarial training, they achieve this by making small perturbations to the word embeddings and not to the input words. This method is used as a regularization mechanism only. It doesn't not serve as a defense mechanism against an adversary as the perturbed word embeddings do not represent any word.

2.5 Semi-Supervised Learning for Text

In [10], *Johnson et al.* use one-hot representation based LSTM for the purpose of semi-supervised classification of text. Generative training was done using LSTM and CNN based approaches using *tv-embedding*. Discriminative training was done using one-hot LSTM followed by pooling and a one layer classifier.

2.6 VAEs for text

Variational Autoencoder(VAE) integrates stochastic latent variables into the autoencoder architecture. They have been shown to perform well in the image and speech domains. Training a VAE architecture in the text domain needs additional care[13]. In [13], the authors propose a VAE architecture for text and training methods to overcome the difficulties in training such models.

Recurrent neural network language models(RNNLMs) are the state of the art architectures for generative modeling of sentences in unsupervised manner[14]. In [13], the authors propose a natural extension to this model by using the VAE architecture. They use LSTM rnns for both encoder and decoder and, the hidden node is acted upon by the gaussian prior. The decoder is conditioned on the value of the hidden node, so when the hidden node, i.e, the latent variable, doesn't encode any meaningful information, this model becomes a standard RNNLM.

Recently, a fully feed-forward convolutional and deconvolutional encoder-decoder architecture for a VAE has been proposed for text generation [15] (Figure 4). A recurrent language model is attached to the deconvolutional decoder (for reconstruction) whose input is the activations of the decoder concatenated with the previous output (Figure 5). This is shown to converge better and handle long sequences well. We use this network in the hybrid model we propose for text classification.

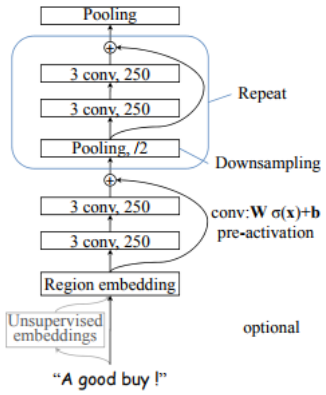

Figure 6: DPCNN architecture [16]

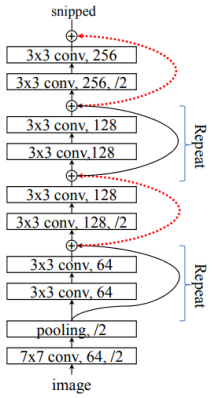

Figure 7: ResNet with skip connections (cf. ResNet for image [HZRS15])

2.7 Deep Pyramid Convolutional Neural Networks (DPCNN) for Text Categorization

This is a word-level deep convolutional neural network (CNN) architecture recently proposed for text categorization (Figure 6) with skip connections similar to the ones introduced in ResNet for images(figure 7). It is shown to represent long range sequences in text[16]. We use features obtained from the last layers of DPCNN, along with the latent features we obtain from the convolutional VAE to do the classification. More on this in section 4, where we describe the hybrid model.

3 Gaussian Mixture VAE with Supervision

Deep Hybrid Models[7](Section 2.2) by *Kuleshov et. al* used the technique of multi-task learning for deep models to integrate the generative and discriminative objectives (see Figure 1). Multi-task learning uses multi-task loss and optimizes the deep learning model through shared layer parameters.

The model uses a VAE for the generative branch, where the latent variable(z) is sampled from a single normal distribution with identity variance. However, it might help if the prior is instead sampled from a mixture of gaussians where each mixture can be mapped to a class. This can help regularization as each example of a particular class comes from a single gaussian and at the same time, providing more flexibility since all the classes do not have to generate examples from the same gaussian. This idea is implemented by *Nat et. al* in [9] for unsupervised clustering.

3.1 Hybrid GMVAE

Taking inspiration from Deep Hybrid Model[7], we add a discriminative branch to GMVAE[9] to introduce supervision in the generative model(see Figure 8). We argue that accuracy of both the branches improve as they help each other during training.

Our model is shown in figure 8. The latent variable x is the encoding for input y. w and z are the additional latent variables which model the means and variances of gaussians, and the probability choosing vector respectively(see Section 2.3). The configuration used for the neural networks which model the above latent variables are kept the same as in [9]. We add the discriminative branch variable y_label, which is connected to the latent x through a fully connected layer.

3.2 GMVAE-Feedback for SSL

Although the model described in Section 3.1 is a good hybrid model, it has one drawback. The discriminative branch and generative branch help each other only through the shared layers which

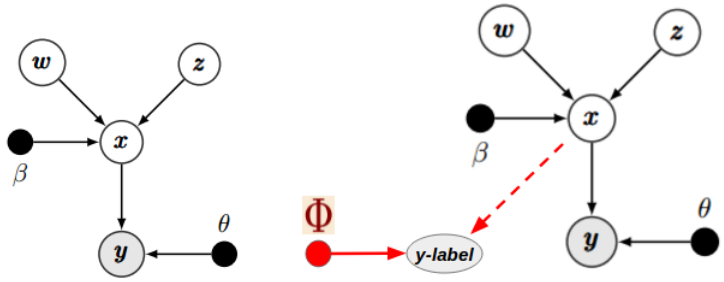

Figure 8: GMVAE(left), Hybrid GMVAE(right): Input-y; Latent Variables - w,z,x; Input Label - y_label

model the latent x. We can provide a more direct feedback if provide supervision to the conditional prior term itself.

As we already have the label of the input, we can sample x from that gaussian directly(see Figure 9). This model has one more advantage over Hybrid GMVAE I, which is very useful for using the model in Semi-Supervised Setup. This model does not have the additional parameters that are added when we add the discriminative branch to GMVAE. Those additional parameters earlier depended solely on the labeled examples. Less number of parameters to train in this model, implies a better performance for semi-supervised learning.

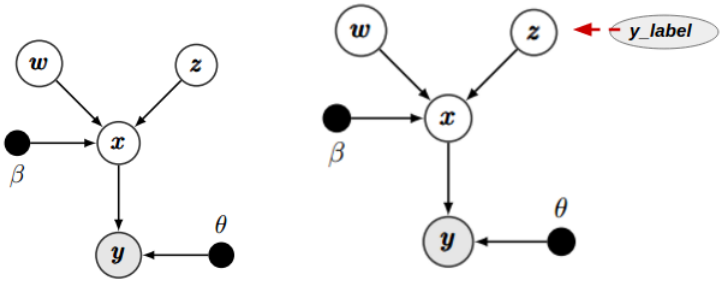

Figure 9: GMVAE(left), GMVAE-Feedback(right): Input-y; Latent Variables - w,z,x; Input Label - y_label

Apart from the above two models(Hybrid GMVAE and GMVAE Feedback), we also experiment with Hybrid GMVAE II, which is a combination of the two and has both the discriminative branches and the feedback connection.

4 Hybrid model for text classification

In this section we describe the models we propose for text classification and explain the reasoning behind the prescribed combination.

Figure 11 shows the model we propose to do text classification. Figures 10 and 11 describe how the VAE described in [15] is combined with deep models for text classification (DPCNN).

The model described in figure 11 directly follows from the deep hybrid models described in [7] (Figure 1). The training of the model is done jointly and this allows the discriminative and generative branches present in the model to help each other, improving each other mutually. We do not use the latent variable generated through the VAE as input to DPCNN as it has already passed through feed-forward convolutional network during it's creation. We use the last layer of a DPCNN concatenated with the latent variable generated through the VAE to perform classification of text

using a network of fully connected layers.

This architecture naturally allows us to do semi-supervised training. When a new sentence

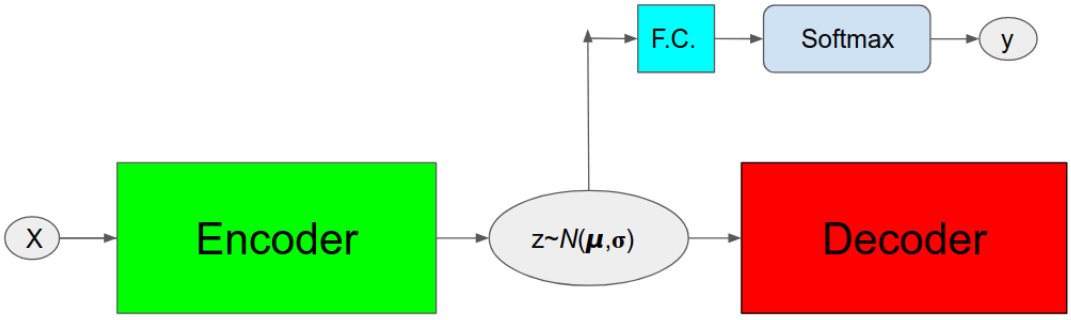

Figure 10: Hybrid model-I for text classification

is given, we use (x, z) and pass them through the fully connected layers to do the classification. We expect the latent variable to learn high level semantic features and the loss used in [15] ensures that latent vector produces features that can be used for historyless reconstruction. So we expect the model to make a good prediction even though the discriminative branch has been trained on a few examples.

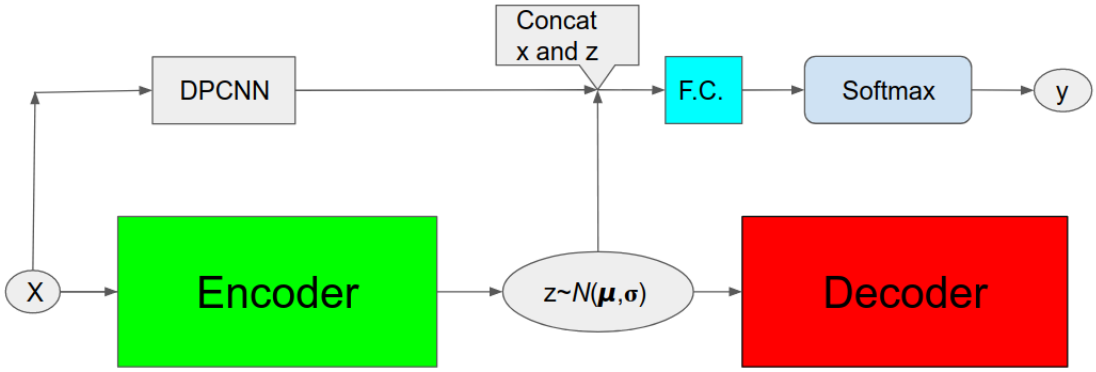

Figure 11: Hybrid model-II for text classification

5 Experiments

5.1 Datasets Used

We used MNIST for experimenting with Hybrid GMVAE and GMVAE-Feedback Models. MNIST has 60000 train examples, 10000 test examples. We also plan to run experiments for SVHN (Street View House Numbers) for a more robust validation of our model.

5.2 Implementation

We have used *torch-lua* for implementation of GMVAE-Feedback models and *theano* for implementing Hybrid Models for text. Training was done on NVIDIA GTX860M with 4GB memory.

5.3 Results

Figure 12 has the results for our models described in Section 3.1 and Section 3.2. The best accuracy of GMVAE as reported in [9] is 96.92 with 10 monte carlo samples and that with 1 monte carlo sample is 89.17. Both our models Hybrid GMVAE II and GMVAE-Feedback outperform the generative accuracy for one monte carlo sample. Also, for 10 monte carlo samples, our model outperforms GMVAE's best run by almost 2%. This was expected because supervision would obviously improve things.

Deep Hybrid Models [7] have an accuracy of 99.03% with 100 labeled examples. Our model(GMVAE-Feedback) has an accuracy of 98.52% which is competitive. We plan to run our experiments on SVHN to validate.

MODEL	Generative Accuracy	Discriminative Accuracy
Generative Branch(DHM)	97.5	-
Deep Hybrid Model(DHM)	98.4	99.4
DHM (SSL = 100)	-	99.03
GMVAE (M=1)	89.17	-
GMVAE (M=10)	96.92	-
Hybrid GMVAE I (M = 1)	81.92	96.58
Hybrid GMVAE II (M = 1)	89.85	96.78
Hybrid GMVAE I (M = 1, SSL = 100)	86.56	49.55
Hybrid GMVAE II (M = 1, SSL = 100)	77.27	51.4
GMVAE Feedback (M = 10)	-	98.8
GMVAE Feedback (M = 10, SSL = 100)	-	98.52

Figure 12: Hybrid GMVAE and GMVAE-Feedback Results, Blue: Baselines, SSL: Number of labeled examples, M: Number of Monte Carlo Examples

Figure 13 shows the samples generated from clusters in both the models GMVAE and GMVAE-Feedback. It can be clearly seen that the clusters on the left are not well defined as some clusters also generate samples from a different class. For example, the fifth row in GMVAE clusters has 4,8 and 9. But, after we provide feedback in GMVAE-feedback every cluster is well defined and generates samples from that class only.

6 Future Work

We plan to to do the following in the near future:

- GMVAE-Feedback experiments on SVHN.
- Hybrid Text Classification Experiments on Amazon Review Dataset.

Figure 13: GMVAE(left), GMVAE-Feedback(right); Samples generated from different Clusters

Apart from the above, the feedback model with GMM prior can be used for text classification also. As we did in the case of image data, we can use gaussian cluster for the prior and argue that each sentiment has a different cluster for the text sentiment classification task. We can also get rid of the discriminative branch and thus lose the extra parameters for the text case also, helping performances in semi-supervised learning. We already show these results for the image data in this report.

References

[1] *The trade-off between generative and discriminative classifiers* G. Bouchard and B. Triggs, In IASC 16th International Symposium on Computational Statistics, pages 721728, Prague, Czech Republic, august 2004.

[2] *A discriminative framework for modelling object classes* A. Holub and P. Perona. , In IEEE Conference on Computer Vision and Pattern Recognition, San Diego (California), USA, june 2005. IEEE Computer Society

[3] *Multi-conditional learning: Generative/discriminative training for clustering and classification* McCallum, Andrew, Pal, Chris, Druck, Greg, and Wang, Xuerui, In Proceedings of AAAI 06: American Association for Artificial Intelligence National Conference on Artifi- cial Intelligence, pp. 433439, 2006.

[4] *Auto-Encoding Variational Bayes* D.P. Kingma, M. Welling, The International Conference on Learning Representations (ICLR), Banff, 2014 [arXiv preprint].

[5] *Semi-Supervised Generation with Cluster-aware Generative Models* Lars Maale Marco Fraccaro Ole Winther arXiv:1704.00637 $[stat.ML]$

[6] *Principled Hybrids of Generative and Discriminative Models* Julia A.Lasserre, Christopher M. Bishop, Thomas P. Mink, Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR06) 0-7695-2597-0/06

[7] *Deep Hybrid Models: Bridging Discriminative and Generative Approaches* Volodymyr Kuleshov, Stefano Ermon, UAI-17. In Proc. 33rd Conference on Uncertainty in Artificial Intelligence, August 2017.

[8] *Generative adversarial nets* Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron C., and Bengio, Yoshua, In Ghahramani et al. (2014), pp. 26722680

[9] *Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders* Nat Dilokthanakul, Pedro A.M. Mediano, Marta Garnelo, Matthew C.H. Lee, Hugh Salimbeni, Kai Arulkumaran, Murray Shanahan, arXiv:1611.02648 [cs.LG]

[10] *Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings* Rie Johnson, Tong Zhang Proceedings of the 33 rd International Conference on Machine Learning, New York, NY, USA, 2016

[11] *Distributional smoothing with virtual adversarial training* Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii , In ICLR, 2016.

[12] *Adversarial training methods for semi-supervised text classification* Takeru Miyato1,Andrew M Dai, Ian Goodfellow, In ICLR, 2017

[13] *Generating Sentences from a Continuous Space* Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. 2016 In CONLL. pages 1021

[14] *Extensions of recurrent neural network language model* Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Honza Cernock'y, and Sanjeev Khudanpur. 2011. In Proc. ICASSP.

[15] *A Hybrid Convolutional Variational Autoencoder for Text Generation 2017* Semeniuta, S.; Severyn, A.; and Barth, E. arXiv preprint arXiv:1702.02390

[16] *Deep Pyramid Convolutional Neural Networks for Text Categorization*; Rie Johnson, Tong Zhang ACL 2017