Semi-supervised Variational Autoencoders for Sequence Classification

Weidi Xu, Haoze Sun
Computational Intelligence Laboratory of Peking University
wead_hsu@pku.edu.cn

Abstract

Semi-supervised learning becomes one of the most significant problems nowadays since the size of datasets is increasing tremendously while labeled data is limited. We propose a new semi-supervised learning method for sequence classification tasks. Our work is based on both deep generative model for semi-supervised learning (Kingma et al., 2014) and variational auto-encoder for sequence modeling (Bowman et al., 2015). We found the introduction of Sc-LSTM is critical to the success in our method. We have obtained some preliminary experimental results on IMDB sentiment classification dataset, showing that the proposed model improves the classification accuracy comparing to pure supervised classifier.

1 Introduction

Semi-supervised learning makes use of both labeled and unlabeled data for training. Classification using semi-supervised learning is a critical problem to be tackled since the data size nowadays is increasing much more faster than before, while only a limited subset of data samples has their corresponding labels. Hence lots of attention has been drawn from researchers over machine learning and deep learning communities, leading to many semi-supervised learning methods. Previous semi-supervised methods (Hinton et al., 2006; Vincent et al., 2010; Bengio et al., 2007) are often used for weight initialization, i.e. pre-training, followed by normal supervised fine-tuning phase. Unsupervised learning are applied to get better initial weight parameters by trying to extract the hidden features which are beneficial for data reconstruction. However these features may be irrelevant to determine the class label in certain classification tasks. For sequence classification problem, Socher et al. (Socher et al., 2013) proposed a similar model using semi-supervised method. The model jointly optimize both reconstruct error and prediction accuracy.

Recently, variational auto-encoder (Kingma and Welling, 2013) is proposed and it has shown impressive performance in image generation tasks. It has also been applied in semi-supervised learning (Kingma et al., 2014; Maaløe et al., 2016) and achieved the best results in several image semi-supervised learning tasks . We refer to these kind of models as SemiVAEs. Different from traditional semi-supervised learning method, this method consists of two cost functions for both labeled data and unlabelled data. The cost for unlabelled data is the expectation of conditional generation likelihood on classification probability of each category. Hence minimizing this cost optimizes conditional generation model as well as classification model. Adoption of conditional generation models and independent classifier alleviates the problem with previous pre-training methods, i.e. extracting features only relevant to data reconstruct.

Sequence classification tasks take sequence as input to predict category labels, whose research hotspot including sentiment classification problem, text categorization, etc. However, SemiVAE can not be applied to sequence classification tasks directly. Sequence classification is different from common image classification problem in several aspects, a) the length of datasamples are variable b) the data distribution of same category is more diverse and c) the category information is more abstract. We combines the variational auto-encoders for semi-supervised learning (Kingma et al., 2014) and its application in sequence generation (Bowman et al., 2015) for sequence classifi-

cation problem. In our implementation the generation is modelled by a long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997). To model conditional generation (main component necessary in SemiVAE), one obvious option is to take the init hidden state and category feature as the first input and then predict each word sequentially using recurrent neural networks, e.g LSTM. Nonetheless we found this implementation would confuse the model to ignore the label input and hence failed to improve the classification accuracy. To force the conditional generation model to be aware of category feature, we utilize the similar structure with semantically controlled LSTM (Sc-LSTM), which is proprosed in (Wen et al., 2015). Using Sc-LSTM, the category feature is fed to generation model at each step, explicitly informing the generation model to be conditioned on given category. The resulting model is capable of apply this semi-supervised method in sequence domain successfully.

In the experiment we show our results on IMDB dataset, which is a common dataset for sentiment classification task. We compare our model with pure supervised learning models and illustrate the importance of several key components in our method.

The article is organized as follows. In the next section, we introduce several background works. And then our model is presented in details. In section 4, we obtain both quantitive results and qualitative analysis about our models. At last we conclude our paper with a detailed discussion.

2 Backgrounds

2.1 Variational Autoencoder

Recently variaitonal autoencoder (VAE) have drawn a lot of attentions due to its impressive results reported in (Kingma and Welling, 2013) and (Gregor et al., 2015). It is equipped with a top-down generative network θ and a bottom-up recognition network ϕ. Both networks are jointly trained to maximize the variational lower bound of data likelihood. Given dataset $X = \{x_1, x_2, ..., x_N\}$, the variational auto-encoder aims to maximize the loglikelihood of all datapoints $logp_\theta(x_1, x_2, ..., x_N) = \sum_{i=1}^{N} logp_\theta(x_i)$. Different form traditontal auto-encoder architecture, the VAE at first samples latent representation z from a learned posterior distribution $p_\phi(z|x)$, using reparameterization tricks (Kingma and Welling, 2013). Then x is reconstructed from the latent space of z by generative model $p_\theta(x|z)$. VAEs encode each datapoint x into a small region (but not a point) in latent space. With this trait, the model has a more smooth latent space than previous autoencoders. The variational lower bound is:

$$\begin{aligned} logp_\theta(x) \geq & E_{q_\phi}[logp_\theta(x|z)] \\ & - KL[q_\phi(z|x)||p(z)] \end{aligned} \tag{1}$$

Generally the first term approximate the reconstruction probability. The distribution of $p(x|z)$ takes the form of Gaussian or Bernoulli distribution, for continuous or binary valued data respectively. The second term acts as the regulariser by minimizing the difference between the prior $p(z)$ and the learned posterior $p_\phi(z|x)$, usually both distributions are diagonal Gaussian. The model is specified as follows:

$$\begin{aligned} logp_\theta(x) \geq & E_{q_\phi}[logp_\theta(x|z)] \\ & - KL[q_\phi(z|x)||p(z)] \end{aligned} \tag{2}$$

$$p(z) = N(Z|0, I) \tag{3}$$

$$\begin{aligned} p_\theta(x|z) = & \mathcal{N}(x|\mu_\theta(z), \delta_\theta(z)) \\ or \quad & \mathcal{B}(x|\mu_\theta(z)) \end{aligned} \tag{4}$$

where $\mu_\phi(x), \delta_\phi(x), \mu_\theta(z), \delta_\theta(z)$ are represented as multi-layer perceptron networks.

2.2 Semi-supervised Learning Using VAEs

Conditional generative models can generate data according to certain attributions of given label. Kingma et al.(Kingma et al., 2014) firstly proposed a conditional variational auto-encoder (CVAE) to successfully separate the image style and content information. Several other works (Maaløe et al., 2016; Yan et al., 2015) also proved CVAE to be powerful in image generation tasks. In addition, semi-supervised learning (Kingma et al., 2014; Maaløe et al., 2016) can help improving the discriminative results by employing unlabeled data. Given labeled dataset $X,Y = \{(x_1, y_1), (x_2, y_2)....., (x_N, y_N)\}$ the lower bound of CVAE is:

$$\begin{aligned} logp_\theta(x, y) \geq & E_{q_\phi(z|x,y)}[logp_\theta(x|y, z) + logp_\theta(y) \\ & + logp(z) - logq_\phi(z|x, y)] \\ & = -L(x, y) \end{aligned} \tag{5}$$

For the unlabeled dataset, the unobserved label y is predicted form the recognition model with a learned classifier $q_\phi(y|x)$. The lower bound is hence:

$$\begin{aligned} logp_\theta(x) &\geq \sum_y q_\phi(y|x)(-L(x,y)) + H(q_\phi(y|x)) \\ &= -U(x) \end{aligned} \tag{6}$$

This equation is the key for semi-supervised learning. Note that the first term is the expection of labeled lower bound on $q_\phi(y|x)$ and intuitively we can explain the procedure as follows: 1) once we can roughly infer the log likelihood of a certain sample x with ground-true label y_p, i.e. $p_\theta(x|y_p, z)$ and wrong label y_n. $p_\theta(x|y_p, z)$ is likely to be greater than $p_\theta(x|y_n, z)$. And hence the gradient computed for $q_\phi(y_p|x)$ has a greater coefficient for $q_\phi(y_p|x)$, and such reinforce the classifier. 2) if the classifier can predict the right label for most time, the generation model will be strengthened in a similar way. This is a positive feedback in the training. We will show how this mechanism can be applied in this sequence classification tasks in our method.

Actually the classification loss term is also added to the objective function, so that the classifier $q_\phi(y|x)$ can be learned from labeled dateset. Hence the overall objective for entire dataset is now:

$$\begin{aligned} J = &\sum_{(x,y)\in S_l} L(x,y) + \sum_{x\in S_u} U(x) \\ &+ \alpha E_{(x,y)\in S_l}[-logq_\phi(y|x)] \end{aligned} \tag{7}$$

2.3 Variational Autoencoder for Sentence Generation

Recurrent neural networks (RNNs) are the most successful methods for sequence generation tasks such as machine translation and image captioning. However, one-by-one generation mechanism in RNNs can't extract high level features like topic, style and sentiment properties. To solve the problem, variational recurrent autoencoders (VRAEs) have been employed in modeling global features for sequential data like sentences(Bowman et al., 2015) and music(Fabius and van Amersfoort, 2014). Similar with the aforementioned models, VRAEs use encoder-decoder structure. In recognitive model, sequence x is processed by encoder RNNs to extract the global information, while in generative model latent variables z are used to initialize the hidden state for decoder RNNs.

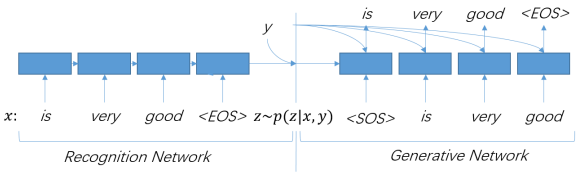

Figure 1: This is the sketch of our conditional variational autoencoder for sequence, which is the key component of our method. **Left**: The sequence is modelled by a recurrent neural network and **Right**: followed by another recurrent neural network for generation. Additional condition given input is imported to this both parts. A sample z from prior $p_\theta(z|x,y)$ is passed to generative nework, which is used to recover the sequential data x.

3 Model

In this section, we introduce our work in details. The model is based on the aforementioned networks and is sketched in figure 1.

Specifically the inference model of our method is described as following equation.

$$q_\phi(z|y,x) = \mathcal{N}(z|\mu(x,y), diag(\sigma^2(x,y))) \tag{8}$$

$$\hat{x} = LSTM(x) \tag{9}$$

$$\mu(x,y) = W([\hat{x};y]) \tag{10}$$

$$log\sigma^2(x,y) = softplus(W([\hat{x};y])) \tag{11}$$

where sequence x is encoded by LSTM network, and the output is concatenated with y for setting this diagonal Gaussian distribution, y is represented as one-hot vectors. Here we use the notation $b = W(a)$ to denote a linear weight matrix with bias from vector a to vector b for simplicity .

3.1 Sc-LSTM

To model the sequence generation conditioned on both hidden state z and class label y, we originally simply concatenate y and z as the initial state for LSTM. However during experiments we found this simple implementation failed to improve the classification performance since the model figures out that ignoring the class feature and minimizing the generation likelihood according to language model (i.e. predicting next word according to a small context windows) is the best strat-

egy to optimize the objective function. This is because the category information is passed to generation network only at first time-step and the conditional generation probability is maximized over all kinds of categories (each multiplied with a coefficient produced by classifier $q_\phi(y|x)$). If the conditional generation model can not distinguish between different category features, the model will be incapable to take advantage of positive feedback mechanism described in section 2.2, and hence fail to boost the performance using this semi-supervised learning method.

Considering this, we force the conditional generation model to generate sequence at each time-step to be aware of given labels. Simplified semantically controlled LSTM (Sc-LSTM) network proposed in (Wen et al., 2015) is adopted to achieve this goal. The Sc-LSTM is defined by the following equations:

$$i_t = \sigma(W_{wi}w_t + W_{hi}h_{t-1}) \tag{12}$$

$$f_t = \sigma(W_{wf}w_t + W_{hf}h_{t-1}) \tag{13}$$

$$o_t = \sigma(W_{wo}w_t + W_{ho}h_{t-1}) \tag{14}$$

$$\hat{c}_t = tanh(W_{wc}w_t + W_{hi}h_{t-1}) \tag{15}$$

$$h_t = o_t \otimes tanh(c_t) \tag{16}$$

$$c_t = f_t \otimes c_{t-1} + i_t \circ \hat{c}_t + tanh(W_{yc}y) \tag{17}$$

where the first six equations are the same as in standard LSTM networks and the last equation has an extra term about y. Therefore the model is explicitly informed of the existence of category feature all the steps. By using this setting the conditional generation model has to distinguish between the different labels. The details analysis will be made in section 4.

3.2 Cost Annealing

Cost annealing is a training trick introduced in (Bowman et al., 2015; Huszár, 2015), which gradually increasing the weight of KL cost from zero to one. Without this the model tend to ignore the input x and most training likely yields models consistently set $q(z|x)$ approximate to $q(z)$. This technique is also considered in our implementation.

4 Experimental Results

This section will show several preliminary experimental results in Large Movie Review Dataset, sometimes know as IMDB dataset. The dataset consists of 25K labelled datasamples for training, 25K labelled datasamples for testing and 50K unlabelled datasamples used for semi-supervised learning. The dataset has adequate samples to training classifier as well as generation model.

methods		5K	10K	15K
LSTM	dev	84.43%	86.56%	87.56%
	test	84.33%	86.51%	87.70%
SemiSeq	dev	86.88%	88.48%	88.84%
	test	**86.84%**	**88.28%**	**88.96%**

Table 1: The results of methods on IMDB dataset

For device set, we randomly sample 5K samples from training set. We adopt a common implementation for classification model, i.e. using the averaged hidden state generated by standard LSTM network to predict the label. Averaging the hidden states is more easier for training reported in (Hong and Fang,). Although there are some other more sophisticated methods for sequence classification, we use this standard model for simplicity in following experiments. Our results may not supass the state-of-art results for IMDB dataset yet due to the limitation of classifier, but still has demonstrated the sufficiency of our semi-supervised method for sequential data.

We compare our method (denoted as **SemiSeq**) with pure-supervised LSTM classifier, with different size of labelled datasamples. To split the labelled data for each experiment we sample from training set with two categories balanced and leave the others unlabelled. The results is listed in table 4. We draw the results of testset with best results on devset in each single experiment.

In all experiments our method is able to improve the classification accuracy for about 1$\sim$2%. Note that the accuracy using 25K samples and carefully chosen hyper-parameters is about 89.1%, which can be regarded as the performance limit of LSTM network for IMDB sentiment classification task. With less labelled samples more improvement can be obtained.

In the future other datasets may be considered.

4.1 Implementation Details

4.1.1 Length Sampling

When using datasets consists of long sequence samples, it is not efficient to use the whole sequence for generation. Assuming sub-sequence has enough information to infer the category, trun-

cated sequence will suffice to differ the generation likelihood of various class feature. On the other hand, the recurrent neural network, even LSTM, is incapable of model very long sequence generation. This trick alleviate the difficulty for conditional generation model.

In the implementation for IMDB dataset where average sequence length is approximately 300, we randomly draw sub-sequence uniformly from each data sample to speed up training. In experiments we found that this trick works well for the purpose of fast training. The training speed is about 3 times faster comparing to using whole sequence.

4.1.2 Dropout and Word-dropout

To improve generalization ability word dropout method is utilized. In conditional sequence generation model, we randomly drop out some words and replace with blanks. This method introduces noise into networks and help the network to be more general. However we found it is inappropriate to use the word-dropout mechanism in equation 6. Since the word-dropout make the generation probability unstable and hence influence the gradient computed for sequence classifier. Hence in our implementation, we use this mechanism only for labelled data, i.e. equation 5, but not in equation 6.

4.2 Qualitative Analysis

Here we examine the capability of generation model. Even though the classification performance is improved using our method, the generation power still remains unknown. To check it out we randomly sample several sentences using trained conditional generation model. The sampled sentences are shown in table 4.2. We show several cases when using the same initial state z and different labels.

From the samples we can tell that both of sentence generated by same z share the similar sentence structure and words, but the sentimental implication is totally different with each other. On the other hand the model is still unable to capture the high level meaning of sentiment but tried to remember the frequency of words for each category.

Negative	Positive
i remember seeing the old godzilla from the valley of the living dead , and i thought it was one of the worst films i have ever seen ...	i remember seeing the out-takes at the end of the film , and when the credits rolled , i was surprised to see how the grinch stole the film ...
suffice to say that the movie is about a group of people who want to see this movie , but this is the only reason why this movie was made in the united states ...	suffice to say that this is one of those movies that will appeal to children and adults alike , but this is one of the best movies i have ever seen .
and most of the characters are so UNK that they have no idea what they were doing in the movie , and they were UNK in the movie ...	and of course the characters are very well done , and they have to be the best of his life , and he has to be the best of his career ...
it reminds me of a lot of people who want to see this movie to see what they are doing ...	it reminds me of the old UNK UNK who can 't afford to be a UNK , but they don 't have a lot of money to do it ...

Table 2: Generated sentences conditioned on different categories and same hidden state z.

5 Conclusion

In this paper we present a new semi-supervised learning method for sequence classification. We explained the reason for using Sc-LSTM as conditional generative model. The resulting model is

capable of improve the classification performance remarkably.

Acknowledgments

This work was supported by the Natural Science Foundation of China (NSFC) under grant no. 61375119 and the Beijing Natural Science Foundation under grant no. 4162029, and partially supported by National Key Basic Research Development Plan (973 Plan) Project of China under grant no. 2015CB352302.

References

[Bengio et al.2007] Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, et al. 2007. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153.

[Bowman et al.2015] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

[Fabius and van Amersfoort2014] Otto Fabius and Joost R van Amersfoort. 2014. Variational recurrent auto-encoders. *arXiv preprint arXiv:1412.6581*.

[Gregor et al.2015] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*.

[Hinton et al.2006] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Hong and Fang] James Hong and Michael Fang. Sentiment analysis with deeply learned distributed representations of variable length texts.

[Huszár2015] Ferenc Huszár. 2015. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*.

[Kingma and Welling2013] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Kingma et al.2014] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

[Maaløe et al.2016] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. 2016. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*.

[Socher et al.2013] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

[Vincent et al.2010] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408.

[Wen et al.2015] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

[Yan et al.2015] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2015. Attribute2image: Conditional image generation from visual attributes. *arXiv preprint arXiv:1512.00570*.