Chao Zhao

Ms. Cummings

Annotated Bibliography--GPP SSR Course

9/19/2018

Dinh, L., Krueger, D., & Bengio, Y. (2015). NICE: Non-linear Independent Components Estimation. In *International Conference on Learning Representations, workshop* (Vol. 1, pp. 1–13). https://doi.org/1410.8516

To learn a good representation for the high-dimensional data, Dinh, Krueger, & Bengio (2015) propose a new framework called Non-linear Independent Component Estimation (NICE). This model is a kind of normalizing flow. By transforming a latent variable using a series of non-linear neural networks, it provides a flexible way to approximate the real probability. But different with other flows, its operations are well-designed to make the Jacobian matrix strictly triangular, and therefore allows the efficient computation of the Jacobian determinant. That's why this work is important in my field.

This work provides a new method to construct the normalizing flow and has been adopted in He's work to solve the problem of unsupervised syntactic prediction. Since the poster will introduce this work, it is necessary to introduce the NICE model as a background. In this way, the listeners can better understand the method utilized in He's paper.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using Real NVP. In *International Conference on Learning Representations*. https://doi.org/1605.08803

Most data in real life is high-dimensional. A middle-sized image, for example, contains at least thousands of dimensions. Such high-dimensional space would contain more dependencies and structures compared with the lower-dimensional space and therefore is more difficult to be modeled. For these reasons, most of the statistical methods proposed before have to use the approximate methods to estimate the probability. This paper (Dinh, Sohl-Dickstein, & Bengio, 2017)

focused on the problem of how to enables the exact and tractable log-likelihood estimation of observed data without damaging the model flexibility, which is important in my field.

This paper would provide a theoretical foundation for all the other parts of my poster. It described a general invertible density method for high-dimensional data modeling and can be applied to a large number of applications where the input dimension cannot be well modeled by other approaches. In the final poster, I will describe this method, as well as several of the typical applications in natural language processing.

He, J., Neubig, G., & Berg-Kirkpatrick, T. (2018). Unsupervised Learning of Syntactic Structure with Invertible Neural Projections. In *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Retrieved from http://arxiv.org/abs/1808.09111

Dependent parsing is important for language understanding because it provides the grammatical and semantic relationships between words, which is useful for downstream tasks such as information extraction and coreference resolution. Annotating dependent parsing trees manually for each language, however, has to spend several years and is time-consuming. This work (He, Neubig, & Berg-Kirkpatrick, 2018) focuses on the unsupervised approaches for syntactic parsing, which means the computers can learn the parsing tree by themselves and therefore the annotated data becomes unnecessary. That is why this problem is important in my field.

Besides its quite innovative and unique ideas, this paper contains many new techniques of recent years. Its bibliography is a great resource for me to follow and receive updates in related areas. I therefore prefer to regard this paper as a backbone of my final project, and use other related materials listed in the reference to enrich the poster.

Lin, C.-C., Ammar, W., Dyer, C., & Levin, L. (2015). Unsupervised POS Induction with Word Embeddings. In *Human Language Technologies:*

The 2015 Annual Conference ofthe North American Chapter of the ACL **(pp. 1311–1316). Retrieved from http://arxiv.org/abs/1503.06760**

The task of unsupervised part-of-speech induction tries to assign a pos tag for each token in an unlabeled text. The traditional models utilize the hidden Markov model to solve this problem, where the output is a categorical distribution among the discrete tokens. Lin et al. (2015) first introduced the continuous word embeddings to this task, where can inform the model with the syntactic similarities. They changed the emission probability to a multivariate Gaussian distribution and remains the other parts of HMM unchanged. In this way, we can integrate the word embeddings into the unsupervised HMM structure, and further improve the performance. That's why it is important in my field.

This paper is the first attempt to integrate word embeddings into HMM, and the work we mainly introduced in the poster is directly inspired by this work. In the final poster, I will describe this method as a preliminary background, which can better help readers understand the unsupervised syntactic prediction algorithms.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *International Conference on Learning Representations*. **https://doi.org/10.1051/0004-6361/201527329**

The generative model is a kind of approach in machine learning to calculate the probability of the input data. Since this modeling process does not need the corresponding label, generative models are widely utilized in unsupervised learning. However, if the data has high dimensions (e.g., the images and languages), it can be difficult to provide an explicit expression of its probability distribution, and therefore the estimation of this probability would also be challenging. To address this problem, a common strategy is to first generate a latent variable from a simple distribution (e.g., an isotropic Gaussian) and then use a non-linear encoder to transform it as our observed data. However, calculating or maximizing it directly by marginalizing the latent variable is usually intractable. Variational inference (VI) is an effective technique to handle this problem. VI tries to maximize the probability indirectly by optimizing a lower bound based on the variational principle. By selecting some particular families of posterior (e.g., the mean-field assumption), the optimization of this lower bound can become

tractable using gradient ascendant. It would, however, also impact the model flexibility, which makes the approximation never be the same with the real posterior. To address this problem, Kingma & Welling (2013) proposed the variational auto-encoder, to implement this approximation using neural networks. The flexibility of neural networks makes the tight approximation become possible and therefore obtains better results compared with the traditional methods. This method has wide usage in generative models and therefore important for my research fields.

As another generative model, variational auto-encoder is different with the normalizing flow. But these two methods have subtle connections in another perspective. I believe it is useful to list all of them in the poster.

Rezende, D. J., & Mohamed, S. (2015). Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 3–6). Lille, France. Retrieved from http://arxiv.org/abs/1505.05770

Variational inference (VI) is an important technique to approximate a complex and unknown distribution, and therefore has wide applications in statistics, especially in generative models. However, instead of directly maximizing the likelihood of data, it tunes the parameters by maximizing a lower bound of the data likelihood. Worse still, to make the inference tractable, traditional VI methods have to limit the flexibility of the approximate posteriors, and therefore would inevitably introducing errors. To address this problem, Rezende & Mohamed (2015) propose a special class of posterior distributions called normalizing flow, which is able to cover the real posterior with a series of invertible mapping functions. That's why it is important in my field.

This paper proposed the idea of normalized flow for the first time, and there are many subsequent studies. Therefore, it is necessary to introduce this work in the poster as a theoretical foundation for other similar works.