DEBUGGING LARGE SCALE DEEP RECOMMENDER SYSTEMS USING UNCERTAINTY ESTIMATIONS AND ATTENTION

Inbar Naor, Ofer Alper, Dan Friedman, Gil Chamiel
Taboola
{inbar.n, ofer.a, dan.f, gil.c}@taboola.com

ABSTRACT

While DL provides many advantages to recommender systems, it comes at the cost of interpretability. In this talk we will present two methods to interpret DL models: 1. Computing saliency maps based on differences in model uncertainty, as estimated using Bernoulli Dropout, when marginalizing out a categorical feature. 2. Attention based interpretability. We will show examples from our large scale recommender system that serves billions of recommendations per day.

Deep Learning methods are gaining an increasing attention in the Recommender Systems community, replacing traditional methods in many industry applications [1-6]. DL has many attractive properties for this domain: it is able to effectively consider non-linear interactions, even from different data sources (e.g text, user browsing history), thus capturing non trivial user-item relationships. It also allows to train a model with multiple objectives, a useful ability for many business cases. This power comes with an interpretability cost, as DL models are often perceived as black boxes. However, interpretability in Recommender Systems is important both for practitioners, enabling informed decisions about how to improve the models, and for explainable recommendations, which were shown to improve user satisfaction and engagement.

In this talk we will share some of the lessons we have learned while developing and debugging a large scale Deep Learning Recommender System. We will focus on two methods: uncertainty estimations and attention modeling.

Uncertainty can be categorized into three types: data uncertainty, model uncertainty and measurement uncertainty [7]. A number of methods were developed in recent years to estimate the uncertainty of a model in its prediction [8-12]; We will introduce a deep unified framework for explicitly modeling and estimating all three types of uncertainty and show different examples of how we can use uncertainty estimations to debug our models. We focus on interpretation by borrowing the idea of saliency maps [13], where a relevance score is computed for each input dimension (a categorical feature, in our case) given a specific sample and classification result. [14] defined a saliency score for input components as the difference in classification output when marginalizing that component out, effectively treating it as unobserved. We suggest a similar approach but instead of considering only the classification difference, we consider the difference in model uncertainty, estimated using variational Bernoulli dropout [8]. Uncertainty allows us to not only consider the accuracy of the model but also to understand its confidence in its predictions and where this confidence (or lack thereof) comes from. We will discuss different methods to marginalize out a categorical feature, such as replacing its embedding vector with an average feature vector or with an Out of Vocabulary embedding whose weights are learnt during the regular training of the model in order to represent rare values of the feature. This allows us to understand which features increase the model's confidence in its predictions.

Another approach to incorporating interpretability into our model is adding attention layers. Attention mechanisms provide weights that allow a network to focus on some input or internal signals based on contextual information, and have achieved significant success in various tasks [15-17]. In recommender systems attention can be used over features like user browsing history, in which we need to summarize a sequence of articles into one vector representation, thus assigning higher weights to articles that the user read and are relevant in a specific context. We can also use attention over different feature interactions, to find the ones that are more important in a specific context. Attention units, especially when trained explicitly to create a network with a desired explainable behavior [18, 19], can produce a map of how and which information propagates through the network, providing valuable interpretability information.

REFERENCES

1. Van den Oord, Aaron, Sander Dieleman, and Benjamin Schrauwen. ”Deep content-based music recommendation.” Advances in neural information processing systems. 2013.

2. Cheng, Heng-Tze, et al. ”Wide deep learning for recommender systems.” Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. ACM, 2016.

3,Covington, Paul, Jay Adams, and Emre Sargin. ”Deep neural networks for youtube recommendations.” Proceedings of the 10th ACM conference on recommender systems. ACM, 2016.

4. Zeldes, Yoel, et al. ”Deep density networks and uncertainty in recommender systems.” arXiv preprint arXiv:1711.02487(2017).

5. Wang, Ruoxi, et al. ”Deep cross network for ad click predictions.” Proceedings of the ADKDD’17. ACM, 2017.

6. Chen, Minmin, et al. ”Top-K Off-Policy Correction for a REINFORCE Recommender System.” Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM, 2019.

7. Kendall, Alex, and Yarin Gal. ”What uncertainties do we need in bayesian deep learning for computer vision?.” Advances in neural information processing systems. 2017.

8. Gal, Y. and Z. Ghahramani (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning, pp. 10501

9. Blundell, Charles, et al. ”Weight uncertainty in neural networks.” arXiv preprint arXiv:1505.05424 (2015).

10. Zhu, Lingxue, and Nikolay Laptev. ”Deep and confident prediction for time series at uber.” 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017.

11. Neumann, Lukas, Andrew Zisserman, and Andrea Vedaldi. ”Relaxed Softmax: Efficient Confidence Auto-Calibration for Safe Pedestrian Detection.” (2018).

12. DeVries, Terrance, and Graham W. Taylor. ”Learning confidence for out-of-distribution detection in neural networks.” arXiv preprint arXiv:1802.04865 (2018).

13. Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. ”Deep inside convolutional networks: Visualising image classification models and saliency maps.” arXiv preprint arXiv:1312.6034 (2013).

14. Zintgraf, Luisa M., et al. ”Visualizing deep neural network decisions: Prediction difference analysis.” arXiv preprint arXiv:1702.04595 (2017).

15. Vaswani, Ashish, et al. ”Attention is all you need.” Advances in Neural Information Processing Systems. 2017.

16. Xiao, Tianjun, et al. ”The application of two-level attention models in deep convolutional neural network for fine-grained image classification.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

17. Lu, Jiasen, et al. ”Hierarchical question-image co-attention for visual question answering.” Advances In Neural Information Processing Systems. 2016.

18. Gilpin, Leilani H., et al. ”Explaining Explanations: An Overview of Interpretability of Machine Learning.” 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018.

19. Ross, Andrew Slavin, Michael C. Hughes, and Finale Doshi-Velez. ”Right for the right reasons: Training differentiable models by constraining their explanations.” arXiv preprint arXiv:1703.03717 (2017).