BAG OF TRICKS AND A STRONG BASELINE FOR NEURAL NEWS RECOMMENDATION

Yichao Lu
Layer 6 AI
yichao@layer6.ai

ABSTRACT

Personalized news recommender systems play an important role in improving online news reading experience. However, news recommendation still remains a greatly unexplored research field, partly due to the lack of a high-quality benchmark dataset in the news recommendation community. The MIND News Recommendation Competition, which is based on the Microsoft News Dataset (MIND), is a first-of-its-kind competition dedicated to benchmarking existing news recommendation algorithms and to advancing the state-of-the-art. This technical report presents a strong baseline for neural news recommendation, built on top of the NRMS model with a bag of task-specific tricks. The proposed baseline approach achieved the second place out of over 200 teams in the MIND News Recommendation Competition, with less than 0.4% relative difference in terms of AUC from the first place team.

1 INTRODUCTION

With the explosive growth of digital news, personalized news recommender systems have become an ever-present part of online news browsing platforms. Through learning user preferences from user behavior logs, personalized news recommender systems are able to present the right news to the right user in the right time.

News recommendation is a challenging task in the sense that (i) it is important for news recommendation models to understand news content in a fine-grained manner so as to capture user interests, (ii) there is often no explicit feedback from users, and (iii) news articles are produced daily and their ephemeral timelife requires accurate modelling of short-term user preferences. Therefore off-the-shelf recommendation algorithms typically cannot achieve desirable performance in the task of news recommendation. Despite the necessity of recommendation algorithms specifically designed for news recommendation, not least because the news recommendation community lacks an authoritative benchmark dataset.

The Microsoft News Dataset (Wu et al., 2020) is dedicated to facilitating the research in news recommendation. To date the Microsoft News Dataset (MIND) is the largest news recommendation dataset, consisting of about 160k English news articles and more than 15 million impression logs generated by 1 million users. The accompanying MIND News Recommendation [1] serves as a testbed for evaluating the performance of various news recommendation models. Given the anonymized behavior logs collected from the Microsoft News website [2], the task is to rank the candidate news in each impression log so that clicked news appear at the top of the list, while non-clicked news appear at the bottom of the list.

This technical report demonstrates that a properly tuned NRMS model (Wu et al., 2019b), when combined with a bag of task-specific tricks, serves as a very strong baseline for neural news recommendation, showing highly competitive performance in the MIND News Recommendation Competition. The final solution, which is based on an ensemble of NRMS models only, achieved the second place out of over 200 teams, with less than 0.4% relative difference in terms of AUC from the first place team. In addition, this technical report describes a number of alternative approaches

[1] https://msnews.github.io
[2] https://microsoftnews.msn.com

that have been explored during the competition, together with a discussion of the potential cause of their inferior performance compared to the optimized NRMS model.

2 APPROACH

The NRMS model (Wu et al., 2019b) is a state-of-the-art neural news recommendation approach with multi-head self-attention. NRMS learns news representations from news titles, using a multi-head self-attention network. Briefly speaking, the news encoder first maps each word in the news title to the the corresponding vector, and then uses the self-attention network to learn word-level representations. Finally, a query vector is used to locate the important words in the news title, and an attention-based pooling method is used to aggregate the word-level representations into the learned title representations. In order to learn user representations from their browsed news, NRMS again uses the multi-head self-attention network on top of the learned news representations. The probability of a user clicking a candidate news is given by the dot product between the user representations and the news representations.

2.1 BI-LSTM WITH ATTENTION MECHANISM

While self-attention based methods have shown superior performance in a variety of natural language processing tasks, especially in pre-trained language models with Transformers, the bidirectional LSTM remains a popular alternative when modelling short texts, and when pre-training is not used. The promising performance of bidirectional LSTM in terms of news modelling has also been reported in Wu et al. (2019b).

According to the experiments during the competition, several tricks can further improve the performance of the bidirectional LSTM based model in the task of neural news recommendation. These inculde: using fastText instead of Glove for initializing the word vectors, orthogonal initialization of the LSTM weights, data augmentation on the text input (e.g., shuffle, replace, and using synonyms), and using the lazy Adam optimizer for sparse word embeddings update.

2.2 POSITION EMBEDDINGS

While the self-attention mechanism can effectively capture the long-term dependency within sequences, it does not have the inherent ability of RNN based approaches to reason about the absolution and relative positions. In other words, the standalone self-attention network model is inherently position-agnostic. However, the order in which the user viewed the news is important for modelling short-term user preference. One would thus expect that enabling position-aware self-attention for the NRMS model would lead to a boost in terms of performance.

The common approach to enabling position-aware self-attention is to introduce the position embedding layer, which essentially encodes each position into a vector representation. It should be noted that, in order to deal with the variable lengths of user history, it is important to ensure that the user browsing history is reversed before feeding into the position embedding layer. In other words, the last viewed news for all users share the same embedding vector, and the second-to-last viewed news for all users share the same embedding vector, etc. Additionally, experimental results suggest that the sinusoidal initialization of the position encoding can result in improved performance in comparison with random initiliazation.

2.3 MULTI-VIEW REPRESENTATION LEARNING

The original NRMS model only learns news representations from titles. However, news articles often contain rich information such abstract and body, which is helpful for learning better news representations. In Wu et al. (2019b), the authors show that the attentive multi-view learning (AMV) approach (Wu et al., 2019a), which models different sources of news information independently and fuses the learned representations via the attention mechanism, outperforms the simple baseline that directly concatenates all sources of information into a long document. However, experimental results during the competition suggest that, simply averaging the representations learned from the multi-views is able to get better performance than the AMV approach. Averaging can be regarded as a special case of the attention-based approach where the model assigns uniform view-level weights.

This suggests that, given sufficient training data, a neural network based model can learn to perform multi-view learning without the need of designing a specific attentive multi-view learning module.

2.4 Scaled Dot-Product Attention with Parameter Sharing

For the original NRMS model, the representation learned by the k-th attention head is formulated as

$$a_{i,j}^k = \frac{\exp(e_i^T Q_k e_j)}{\sum_{m=1}^M \exp(e_i^T Q_k e_m)}, \tag{1}$$

$$h_{i,k} = V_k(\sum_{j=1}^M a_{i,j}^k e_j), \tag{2}$$

where Q_k and V_k refer to projection matrices, and $a_{i,j}^k$ denotes the relevance score between the i-th news and the j-th news.

Recently studies have found that in particular tasks, weight sharing in the self-attention layer can lead to better performance, presumably due to the fact that weight sharing inherently provides regularization to the model. In the MIND dataset, weight sharing again turns out to be an effective approach. In addition, the scaled dot-product attention proposed by Vaswani et al. (2017) is an effective approach to stabilize the training of the self-attention based models by enforcing the output to have unit variance. Thus the modified scaled dot-product attention with parameter sharing can be formulated as follows:

For the original NRMS model, the representation learned by the k-th attention head is formulated as

$$a_{i,j}^k = \frac{\exp(\frac{1}{\sqrt{d_k}} e_i^T Q_k^T K_k e_j)}{\sum_{m=1}^M \exp(\frac{1}{\sqrt{d_k}} e_i^T Q_k^T K_k e_m)}, \tag{3}$$

$$h_{i,k} = K_k(\sum_{j=1}^M a_{i,j}^k e_j), \tag{4}$$

where $\frac{1}{\sqrt{d_k}}$ is a scaling factor described in Vaswani et al. (2017).

2.5 Ensemble

The ensemble technique is a widely used strategy to improve the performance of machine learning models. Ensemble methods work by aggregating the predictions of a number of single models. There are two common strategies to ensemble models: bagging and boosting. Since NRMS is a neural network based model, it is naturally a low bias and high variance model. Therefore bagging based ensemble works better than boosting for NRMS since it can help reduce the bias.

Three different ensemble approaches have been experiment during the competition:

- Rank ensemble: using the averaged rank predicted by single models as the prediction by the ensemble model.
- Raw logits ensemble: using the average of the pre-softmax logits as the prediction by the ensemble model.
- Scaled probabilities ensemble: using the average of the unnoramlized softmax activations as the prediction by the ensemble model.

Scaled probabilities ensemble works the best among all the three explored approaches, due to the fact that the scaled probabilities more accurately reflect each single model's confidence about the predicted click.

3 ALTERNATIVE APPROACHES

To facilitate the research in news recommendation, this section describes some other explored yet unsuccessful approaches during the competition, and discusses the potential causes of failure of these approaches. It should be noted that, while all these approaches fail to outperform the NRMS model under the settings of the MIND competition, i.e., to predict the click behaviors for one week in the future, some approaches are superior when considering a shorter horizon of time; see Section 4 for more details.

3.1 END-TO-END FINE-TUNING WITH BERT

Wu et al. (2020) reported an improved performance in terms of news recommendation quality when replacing the word embedding layer in the news encoder with a pretrained-then-finetuned BERT model. However, such model is cumbersome and is infeasible to run in a timely manner. Therefore, in the interest of time, an alternative approach, which directly uses pooled representations from BERT as the news representations, was explored during the competition. However, this approach failed to achieve comparable performance against the bidirectional LSTM based approach. The possible reason that such way of using end-to-end fine-tuned BERT is not showing improvement is that the pooled representation of BERT is more suitable for the classification task.

3.2 FEATURE ENGINEERING

In traditional recommender systems problems, hand-engineered features can significantly boost the performance of machine learning models. Experiments on the MIND dataset show that, when the model is equipped with engineered features such as the length of the user history and the number of news in the user history with the same category as the candidate news, the model achieves better performance on the validation set, while at the same time the test set performance drops. The result suggests that engineered features probably have less generalization ability compared to neural network based models like NRMS, hence the need for research on neural news recommendation.

3.3 TWO-STAGE RE-RANKING

The two-stage re-ranking model is a common technique used in a lot of data science competitions Volkovs et al. (2018). On the MIND dataset, a two-stage re-ranking model which uses the NRMS as the first stage and XGBoost as the second stage is able to improve the NRMS baseline model by over 10% in terms of AUC on the validation set, while getting significantly worse performance on the test set. This suggests that two-stage re-ranking models are more suitable for making short-term predictions, while neural news recommendation models such as NRMS has more consistent performance over a longer period of time.

3.4 TRAINING EXPERT MODELS

In the MIND dataset, the model needs to handle different recommendation scenarios, e.g., recommending news when there is different number of available history news. Instead of applying one model to solve all these tasks, one common practice is to train individual models each for solving one particular task (Lu et al., 2019). The rationale behind this is that, training expert models that specialize in one particular task allows the model to learn task-specific patterns and at the same time enables easier optimization. While such approach is shown to be effective on a number of other tasks, on the MIND dataset, it fails to bring about gains. This observations suggests that the NRMS model, especially the self-attention model it uses, is well-suited for news recommendation in all kinds of scenarios.

3.5 ADDRESSING COLD-START WITH POPULARITY

Cold start is a common phenomenon in news recommendation. For example, about 1.2% of the impression data in the MIND test data is cold-start. A common approach to addressing the cold-start problem is to recommend the most popular item to the user, i.e., recommending the most clicked news in the task of news recommendation. However, the success of such method relies heavily on

Table 1: Ablation study on the MIND News Recommendation Competition.

Method	Validation AUC	Test AUC
Default NRMS	0.6856	0.6826
+ Bi-LSTM	0.6872	0.6855
+ Position encoding	0.6951	0.6923
+ Multi-view learning	0.7074	0.7033
+ Attention with parameter sharing	0.7122	0.7042
+ BERT fine-tuned	0.6832	—
+ Feature engineering	0.7031	0.6545
+ Two-stage re-ranking	0.7733	0.6486
+ Expert models	0.6845	0.6825
+ Popularity	0.6877	0.6814

Table 2: Final leaderboard results for the top-5 teams. The result of the solution described in this technical report is in **bold**.

Rank	Team	AUC	MRR	nDCG@5	nDCG@10
1	chenghuige	0.7131	0.3608	0.3960	0.4521
2	**oahciy**	**0.7096**	**0.3540**	**0.3883**	**0.4454**
3	Ravox	0.7048	0.3505	0.3845	0.4416
3	Qinne	0.7032	0.3496	0.3830	0.43976
3	gcc_microsoft	0.6979	0.3479	0.3806	0.4373

the how recent the supervised data is, so that the model can accurately estimate the popularity of the news. Therefore, while the popularity based method show improved performance on the validation set (predicting click behaviors in the next day), it fails to achieve desirable performance on the test set (predicting click behaviors in the next week).

4 Experiments

4.1 Implementation Details

The model is trained with the Adam optimizer with a batch size of 32 and a learning rate of $1e-4$. Following (Wu et al., 2019b), the max length of the user history is set to be 50. For training the two-stage re-ranking model, the first six days of the training data is split to be the first stage training data, and the last day of the training data is split to be the second stage training data.

4.2 Results

An extensive ablation study is shown in Table 1. The test performance for some of the approaches is unavailable because the MIND News Recommendation Competition adopts a very rigorous evaluation protocol, and each team is only allowed to submit the test results for evaluation for a limited number of times. It should also be noted that the test AUC reported here refers to the public leaderboard performance, which is evaluated on 10% of the test data.

The final leaderboard results for the top 5 teams is presented in Table 2. The final solution, which based on an ensemble of the optimized NRMS model, achieves competitive performance against the solutions provided other teams.

5 CONCLUSION

This technical report demonstrates that a properly tuned NRMS model (Wu et al., 2019b), when combined with a bag of task-specific tricks, serves as a very strong baseline for neural news recommendation. An extensive ablation study proves the effectiveness of each proposed trick.

REFERENCES

Yichao Lu, Cheng Chang, Himanshu Rai, Guangwei Yu, and Maksims Volkovs. Learning effective visual relationship detector on 1 gpu. *arXiv preprint arXiv:1912.06185*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Maksims Volkovs, Himanshu Rai, Zhaoyue Cheng, Ga Wu, Yichao Lu, and Scott Sanner. Two-stage model for automatic playlist continuation at scale. In *Proceedings of the ACM Recommender Systems Challenge 2018*, pp. 1–6. 2018.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576*, 2019a.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6390–6395, 2019b.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3597–3606, 2020.