Improving Neural Collaborative Filtering through Ablation Studies on Model Architectures and Negative Sampling Strategies

Benard Adala Wanyande[s4581733] and Ataklti Haileslase[s4590821]

Leiden Institute of Advanced Computer Science, Leiden University, Leiden, Netherlands

Abstract. This study aims to investigate the performance of Neural Collaborative Filtering (NCF) by exploring various model architectures and negative sampling strategies. Initial tests were conducted by varying the neural network configurations. Although these tests yielded minor improvements, further enhancements were pursued. Finally, an ablation study was performed to evaluate the effects of different negative sampling ratios on NCF performance, based on the following metrics: Hit Rate@10, NDCG@10, and Recall@10.

1 Introduction

Neural Collaborative Filtering (NCF) is a recommendation model that leverages deep learning by replacing traditional inner product operations with neural networks [3]. In particular, it integrates Generalized Matrix Factorization (GMF)[3] and Multi-Layer Perceptrons (MLPs)[3] to learn intricate, higher-dimensional user-item relationships. This allows for richer and more expressive latent representations compared to traditional linear collaborative filtering methods [4].

2 Research Methodology

Initially, we implemented and evaluated a baseline NCF model (embedding dimension of 32, MLP layers [64, 32, 16, 8]), using a negative sampling ratio of 1:4. We hypothesized that varying the model's embedding dimensions (24, 32, 64, and 128) would yield significant performance improvements; however, these architectural adjustments provided incremental insights but did not notably enhance Recall@10. Given the limited gains from altering embedding sizes, we turned our focus to data-level adjustments, specifically negative sampling strategies. Inspired by Rendle et al. [5], we recognized that increasing the negative sampling ratio could improve the model's discrimination capabilities between positive and negative interactions. Thus, we performed an ablation study exploring multiple negative sampling ratios: 1:4 (baseline), 1:8, 1:10, 1:15, 1:50, and 1:100.

3 Experimental Setup

3.1 Dataset and Initial Preprocessing

For this study, we used the MovieLens dataset [1]. The initial preprocessing involved loading the dataset, which contained explicit ratings ranging from 0 to 5. We converted these explicit ratings into implicit feedback by mapping ratings greater than 4 to positive feedback ("1") and treating unrated movies as negative feedback ("0"). To generate negative feedback, we identified movies that users had not rated. For our baseline configurations, prior to the ablation tests, we sampled negative instances at a ratio of 1:4 (positive to negative).

3.2 Model Implementation

All models were implemented using PyTorch and trained using binary cross-entropy (BCE) loss, optimized with Adam, and employed early stopping to prevent overfitting.

3.3 Evaluation Metrics

We used the following standard metrics for evaluation [2]:

- **Hit Rate@10**
- **NDCG@10**
- **Recall@10**

4 Results and Analysis

4.1 Impact of Embedding Dimensions on NCF Performance

We initially explored the effects of varying embedding dimensions on the performance of Neural Collaborative Filtering (NCF) models.

Table 1. Comparison of NCF Model Architectures with Varying Embedding Dimensions. Metrics include Hit Rate@10, NDCG@10, and Recall@10.

Embedding Dim.	Hit Rate@10	NDCG@10	Recall@10
32 (Baseline)	0.9504	0.4882	0.4429
64	0.9495	0.4908	0.4414
128	0.9507	0.4933	0.4373
24	0.9478	0.4806	0.4381

The embedding dimension of 128 achieved the highest NDCG@10 (0.4933), while the baseline 32 yielded the best Recall@10 (0.4429). These marginal differences suggest that increasing embedding size offers limited benefit, highlighting the need to explore data-level strategies like negative sampling.

4.2 Negative Sampling Ablation Study

We conducted an ablation study to investigate the impact of negative sampling ratios on the performance of Neural Collaborative Filtering (NCF) models. Adjusting negative sampling ratios is an essential data-level strategy for improving the model's ability to differentiate between positive and negative user-item interactions. Table 2 summarizes our findings from varying negative sampling ratios.

Table 2. Performance Comparison of NCF Models with Various Negative Sampling Ratios. Metrics evaluated are Hit Rate@10, NDCG@10, and Recall@10.

Negative Ratio	Hit Rate@10	NDCG@10	Recall@10
1:4 (Baseline)	0.9824	0.5112	0.5785
1:8	<u>0.9872</u>	<u>0.5178</u>	<u>0.5916</u>
1:10	0.9802	0.5176	0.5775
1:15	0.9787	0.5107	0.5695
1:50	0.9724	0.4985	0.5531
1:100	0.9653	0.4872	0.5418

The negative sampling ratio of <u>1:8</u> yielded the best performance across all metrics—Hit Rate@10 (<u>0.9872</u>), NDCG@10 (<u>0.5178</u>), and Recall@10 (<u>0.5916</u>). Beyond this point, performance declined with higher ratios (e.g., 1:50, 1:100), suggesting that too many negative samples may introduce noise. Thus, a moderate ratio like 1:8 offers an effective balance between discrimination and accuracy.

5 Discussion

Adjusting model complexity alone provided limited improvement, indicating diminishing returns from increased embedding dimensionality [3]. However, increasing the negative sampling ratio notably improved performance metrics, particularly Recall@10. The embedding dimension results suggest that beyond a certain complexity (32 in this study), further increases yield marginal gains, highlighting the importance of selecting appropriate model complexity. Conversely, the negative sampling ratio ablation emphasizes that a moderate ratio, such as 1:8, achieves optimal discrimination between positive and negative user-item interactions. Excessively high negative ratios introduce noise, reducing predictive accuracy, thereby underscoring the value of balanced negative sampling strategies in collaborative filtering tasks.

6 Conclusion

Our findings highlight the effectiveness of negative sampling adjustments in enhancing recommendation metrics, particularly recall. Future research should investigate advanced techniques such as regularization and hard negative mining to further improve recommendation systems [5].

References

1. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. ACM transactions on interactive intelligent systems (TIIS) **5**(4), 1–19 (2015)
2. He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M.: Lightgcn: Simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 639–648 (2020)
3. He, X., Liao, L., Zhang, H., Nie, L., Hu, X.: Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web. pp. 173–182 (2017)
4. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. In: Computer. vol. 42, pp. 30–37. IEEE (2009)
5. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. arXiv preprint arXiv:1205.2618 (2012)

Appendix

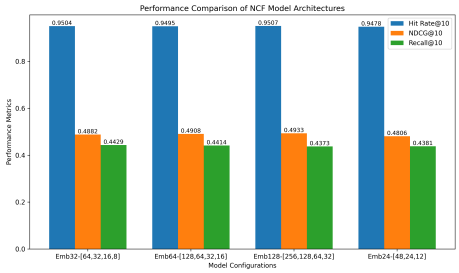

Fig. 1. Performance Comparison of NCF Model Architectures Across Embedding Dimensions

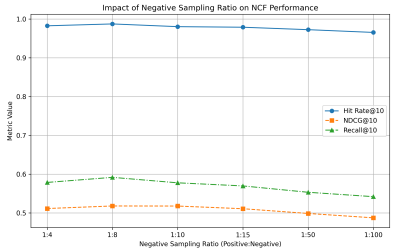

Fig. 2. Impact of Negative Sampling Ratios on NCF Performance Metrics