Dear Editors and Reviewers:

Thanks very much for taking the time to review this manuscript. We appreciate all your comments and suggestions. We provide a detailed response to all of the previous reviews, as well as an updated manuscript highlighting in BLUE.

Thanks again.

---

1. RESPONSE TO REVIEWER #1

Comment 1: Please adjust the font size in some pictures and tables to an appropriate size.

Response: Thanks for the reviewer's comments. Due to the large number of diagrams in the manuscript, we carefully considered and adjusted them. In order to enhance the readability of the figure, we modified the original figure 3 in Section 3.2 and replaced the table in Section 4.1.3 with Figure 4. We realize that the content shown in Table 4 and Table 5 is essentially the ablation experiment of HBA mechanism in Section 4.1.1, and now combine the two into the new Table 4.

Modification:

Section 3.2 Figure3

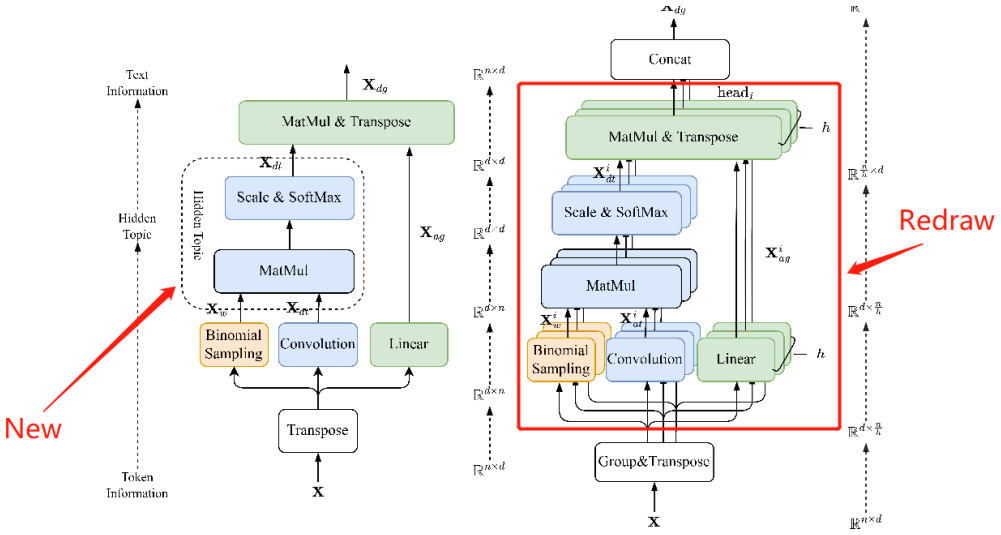

Figure 3: (left) Hidden bias attention. (right) Multi-head HBA consists of several attention layers running in parallel.

Section 4.1.3 Figure 4.

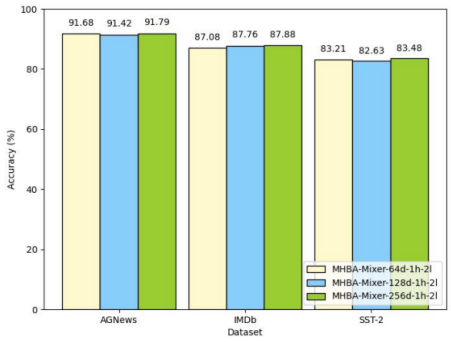

Figure 4: The Accuracy and Parameter scale comparisons of the MHBA-Mixer-1h-2l with different dimensions.

Section 4.1.1 Ablation experiment of the HBA

Firstly, we conducted an ablation experiment on the HBA The result is demonstrated in Table 4.

Table 4
Ablation experiment of the HBA in Accuracy (%)."bs": Binomial Sampling. "gc": grouped convolution

Methods	Datasets		
	AGNews	IMDb	SST-2
HBA	91.30	**87.80**	**82.10**
HBA-bs	**91.63**	86.50	81.97
HBA-gc	91.50	85.53	81.10
HBA-bs-gc	91.44	86.23	80.73

On the AGNews dataset, the HBA-gc achieves the best result with an accuracy of 91.63%. The HBA achieves an accuracy of 91.30%. However, it doesn't mean the group convolution is not work, on the contrary, it shows admirable performance on the remaining datasets.

For example, the HBA achieves the highest accuracy of 87.80% and the HBA-gc achieves the second accuracy of 86.50%. The HBA outperformed HBA-gc 1.57%. A similar phenomenon also occurs in the SST-2 dataset, which shows that the HBA is valid.

Comment 2: Please double-check the formulas to ensure their correctness.

Response: Thanks for the reviewer's suggestion. We have examined the formula carefully.

Comment 3: Please consider replacing some tables (e.g., Table 11 and Table 12) with line graphs to enhance the results' intuitiveness.

Response: The reviewer's advice was very helpful. We have drawn the tables in Appendix A.2 and A.3 as line graphs to help readers understand.

Modification:

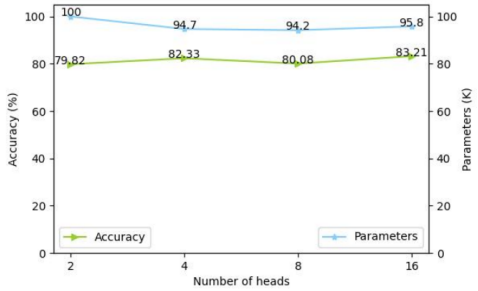

Figure 5: The comparisons of the MHBA-Mixer-64d-*n*h-2l with different number of heads *n*.

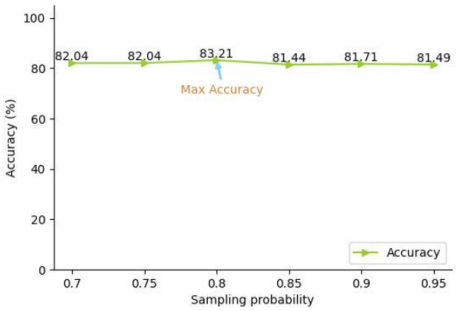

Figure 6: The comparisons of the MHBA-Mixer-64d-1h-2l with different sampling probabilities.

2. RESPONSE TO REVIEWER #4

Comment 1: The paper lacks 2023 references. Please discuss more recent relevant work, such as: Android-IoT Malware Classification and Detection Approach Using Deep URL Features Analysis Journal of Database Management (JDM) 34 (2), 2023 A deep convolutional neural network stacked ensemble for malware threat classification in internet of things Journal of Circuits, Systems and Computers, 2022

Response: We appreciate the comments of the reviewers and carefully read the recommended papers. Relevant literatures are valuable for reference, and we have added them into the Introduction.

Comment 2: Please explain the choice of the evaluation criteria, such as, explain with technical details why selected criteria are important for this paper's application.

Response: Thanks for the reviewer's advice. We add Metrics and Training explanation before section 4.1 of manuscript.

Modification:

Metrics In this study, a trade-off analysis is conducted between the model's size and its quality. The model size is measured by computing the number of parameters, which is a critical determinant of its complexity. Meanwhile, the evaluation of model quality focuses on the top-1 accuracy, which serves as a benchmark metric for both state-of-the-art (SOTA) models and their counterparts in related works.

Training The program was written using PyTorch 1.7 framework, and the effectiveness of the method was validated on a computer with an AMD R7 4800H CPU, an RTX3060 GPU, and 16GB of RAM. During the experiment, the model's

learning rate was carefully set to 5e-4 in order to strike a balance between achieving good performance and avoiding overfitting. Additionally, weight decay was set to 1e-3 to regularize the model and reduce the risk of overfitting. The batch size was chosen to be 100, which allowed for efficient processing of large amounts of data during each iteration. Finally, the number of epochs was set to 1000, ensuring that the model had sufficient time to converge and produce accurate results.

Comment 3: Explain in good technical details on validity and generalisability of the results

Response: Thanks for the reviewer's suggestions. We modify Section 4.1.1 to the ablation experiment. We believe that the effectiveness and generalization ability of the model mainly depend on the proposed HBA mechanism. First, the binomial sampling layer in the HBA mechanism samples tokens multiple times to improve the utilization rate of tokens (we added the interpretation of binomial sampling in Section 3.2). Second, the grouped convolution is used to extract text's hidden topics, which is not taken into account by previous attention mechanisms. Finally, hidden topics are used to weight the global information of the text, so that the global information of the text has a topics bias.

Modification:

Section 3.2

The Binomial Sampling method is a unique anti-masking mechanism that differs from traditional sampling and masking methods. While traditional sampling methods, such as up-sampling or down-sampling, operate on feature maps of different scales, the Binomial Sampling method operates on feature maps of the same scale. Additionally, while the Dynamic Masking mechanism masks the same tokens at the same location in each epoch, the Binomial Sampling method masks different tokens in every layer and epoch. Specifically, the masking probability of a token in the Dynamic Masking mechanism is p, whereas in the j^{th} Binomial Sampling layer, it is $(1-p)^j$.

Furthermore, we have observed that the approach of retaining 80% of tokens unchanged is still valid in our method, as supported by experimental results showing its effectiveness. We suspect that this is because increasing the perturbation through the remaining tokens enhances the model's generalization ability.

Section 4.1.1 Ablation experiment of the HBA

Firstly, we conducted an ablation experiment on the HBA The result is demonstrated in Table 4.

Table 4
Ablation experiment of the HBA in Accuracy (%)."bs": Binomial Sampling. "gc": grouped convolution

Methods	Datasets		
	AGNews	IMDb	SST-2
HBA	91.30	**87.80**	**82.10**
HBA-bs	**91.63**	86.50	81.97
HBA-gc	91.50	85.53	81.10
HBA-bs-gc	91.44	86.23	80.73

On the AGNews dataset, the HBA-gc achieves the best result with an accuracy of 91.63%. The HBA achieves an accuracy of 91.30%. However, it doesn't mean the group convolution is not work, on the contrary, it shows admirable performance on the remaining datasets.

For example, the HBA achieves the highest accuracy of 87.80% and the HBA-gc achieves the second accuracy of 86.50%. The HBA outperformed HBA-gc 1.57%. A similar phenomenon also occurs in the SST-2 dataset, which shows that the HBA is valid.

Comment 4: Please explain limitation of the current solution.

Response: As pointed out by the reviewer, the limitation of our model is the balance between model accuracy and model size. Model lightweight will lead to a slight decrease in model accuracy. However, this problem can be mitigated or even solved by some training techniques. Subsequently, we will use pre-training, whole-word masking, whole-word embedding and other technologies that can significantly improve the accuracy of the model to optimize our model, so as to obtain the performance comparable to or even exceeding that of the current advanced models (such as BERT and XLNet).

Thanks for all the valuable reviews which help us improve the quality of our manuscript and give us more insights into the research.

Best regards,

Huanling Tang, Xiaoyan Liu, Yulin Wang, Quansheng Dou, Mingyu Lu

2023.4.18