Enhancing Educational Content Matching Using Transformer Models and InfoNCE Loss

Yujian Long *, Dian Gu , Xinrui Li , Peiqing Lu , Jing Cao

Posted Date: 15 November 2024

doi: 10.20944/preprints202411.1070.v1

Keywords: educational content matching; Transformer model; InfoNCE Loss; model distillation; STEM education

 doi:10.20944/preprints202411.1070.v1

Article

Enhancing Educational Content Matching Using Transformer Models and InfoNCE Loss

Yujian Long [1,*], Dian Gu [2], Xinrui Li [3], Peiqing Lu [4] and Jing Cao [5]

[1] Independent Researcher, TX, USA
[2] University of Pennsylvania, Philadelphia, USA
[3] Independent Researcher, Austin, USA
[4] Boston University, Maplewood, USA
[5] Northeastern University, Oakland, USA
* Correspondence: yiyayiya.dot@gmail.com

Abstract: The task of matching educational content to specific topics within curricula is crucial for enhancing learning outcomes and ensuring that students receive relevant and coherent learning experiences. This research seeks to develop a precise and efficient model, leveraging a dataset that spans various languages and numerous STEM fields. Our proposed solution integrates a Transformer model with InfoNCE Loss and model distillation techniques to effectively address noise from related content and improve matching accuracy. By incorporating advanced deep learning methods, the model is designed to handle the complexity and variability of educational datasets, outperforming existing models in terms of accuracy and efficiency. This approach provides a robust framework for educational content alignment, significantly benefiting educational practitioners by streamlining the content matching process and enhancing the quality of education delivered to students.

Keywords: educational content matching; Transformer model; InfoNCE Loss; model distillation; STEM education

1. Introduction

Matching educational content to specific topics within curricula is a complex yet essential task in modern education. The ability to accurately align educational materials with curriculum standards ensures that students receive relevant and coherent learning experiences. This study focuses on developing a model to match content accurately within a broad and diverse dataset, particularly emphasizing STEM education.

Traditional methods for educational content matching often rely on manual curation by experts, which is time-consuming and susceptible to biases. As educational resources continue to expand, there is a need for automated techniques to efficiently process vast datasets. In this context, we propose a novel model that leverages advanced deep learning techniques to enhance the accuracy and efficiency of content matching.

Our dataset comprises various formats, including videos and documents, each with its unique distribution patterns. The analysis revealed a long-tail distribution and demographic differences in sample distributions across genders and age groups. To address these challenges, we implemented sample weighting and sampling strategies tailored to the dataset's characteristics.

To improve the model's performance, we integrated InfoNCE Loss, a symmetric contrastive loss function. InfoNCE Loss mitigates the noise introduced by related content within the same batch by focusing on the similarity between the correct pairs of content and topics. This loss function strengthens the model's ability to separate similar from dissimilar pairs, thereby enabling more accurate content matching.

In addition, we utilized model distillation techniques to further boost the model's efficiency. In this process, a pre-trained Teacher Model guides a simpler Student Model. The Student Model, starting with the Teacher Model's weights, is trained using MSE Loss to retain the complex model's performance while minimizing computational demands. This distillation process ensures that the final model is both efficient and effective for large-scale deployment.

 doi:10.20944/preprints202411.1070.v1

The proposed model integrates these advanced techniques to provide a comprehensive solution for educational content matching. By addressing the challenges of sample distribution variability and leveraging state-of-the-art deep learning methods, our model offers significant improvements in accuracy and efficiency, ultimately enhancing the quality of educational experiences for students.

2. Related Work

The application of deep learning in educational content matching has gained considerable attention in recent years. Early approaches relied on traditional machine learning techniques, which often struggled with scalability and accuracy in large, diverse datasets.

Transformers, introduced by Vaswani et al.[1] , revolutionized natural language processing by capturing contextual information in sequential data. Their application in educational content matching has shown promising results, especially in handling diverse and large-scale datasets.

Chen et al.[2] explored contrastive learning, which has improved model robustness. InfoNCE Loss, a variant of contrastive learning, enhances model discrimination between similar and dissimilar pairs, mitigating noise from related content within the same batch.

Siyue Li's work [3] on Mult-Recall methods and machine learning rankers, particularly the LightGBM Ranker, directly influenced our approach to content matching. Their techniques for embedding learning and feature engineering shaped our use of InfoNCE Loss to enhance content similarity while managing noise. Additionally, Li's method of handling multimodal data inspired how we tackled the variability and long-tail distribution in educational datasets, contributing to the precision and efficiency of our Transformer-based model.

Brown et al.[4] introduced GPT-3, showcasing the potential of large language models in few-shot learning. GPT-3's performance in understanding and generating human-like text supports the use of advanced models in content matching.

Liu et al.[5] presented RoBERTa, a refined variant of BERT, was presented, attaining state-of-the-art results by optimizing training processes and maximizing data usage. RoBERTa's robust performance underscores the importance of training efficiency in educational applications.

Lan et al.[6] proposed ALBERT, a lite version of BERT, which reduces model size while maintaining performance. ALBERT's efficiency is valuable for real-time educational content matching.

Siyue Li's work [7] on strategic deductive reasoning using dual-agent frameworks directly influenced our approach to educational content matching. Li's techniques for structured reasoning and decision-making informed our use of Transformer models in managing complex datasets, improving both precision and efficiency in content alignment tasks.

Dosovitskiy et al.[8] demonstrated Transformers in image recognition, showing their versatility across different data types, including educational content.

He et al.[9] proposed deep residual networks, foundational in improving deep learning performance, applicable to refining educational content matching models.

Jiaxin Lu et al.'s work[10] on ensemble learning for multi-objective recommendations influenced our model integration strategy, enhancing accuracy and robustness. Their methods for addressing sample imbalance and variability informed our preprocessing, optimizing educational content matching.

Kiela et al.[11] explored learning distributed representations of concepts using linear transformations, improving semantic understanding of educational content.

Yang et al.[12] proposed XLNet, achieving superior performance in language understanding tasks, highlighting advanced pretraining techniques' importance for content matching.

Ramesh et al.[13] introduced DALL·E, generating images from textual descriptions, showing the potential of combining vision and language models for content generation and matching. Liu et al.[14] introduction of an entropy- and attention-based feature extraction network significantly improved our approach to handling multi-target coupling in educational datasets. He et al.[15] introduce the use of InfoNCE Loss to differentiate semantically similar text pairs, enhancing the

 doi:10.20944/preprints202411.1070.v1

model's capability to manage noise. This informs our approach to enhancing matching accuracy across diverse educational content. Yu et al.[16] demonstrate the application of Transformer models for complex cross-domain content alignment, which supports our method in managing the variability of multilingual educational datasets. Zhang et al.[17] explore model distillation techniques for balancing efficiency and accuracy, guiding our strategy to optimize model performance in educational content matching tasks. Thompson et al.[18] review NLP applications in text sentiment analysis, providing crucial insights into refining our sentiment-based content alignment techniques.Kim et al.[19] discuss the application of NLP technology in big data, which helped shape our data preprocessing and feature extraction methods.Wang et al.[20] demonstrate parameterized decision-making with multi-modal perception, guiding the integration of multi-modal data sources to enhance the robustness and accuracy of our matching model. In summary, integrating advanced deep learning models like Transformers, contrastive learning techniques like InfoNCE Loss, and model distillation offers a comprehensive approach to educational content matching. Our work builds on these advancements, addressing the challenges of aligning educational materials with curriculum standards, providing a robust and efficient solution for educators.

3. Methodology

In this section , we introduces a Transformer-based recommentaion system that employs InfoNCE loss for contrastive learning and model distillation to enhance performance.

This paper presents a novel approach to information retrieval by integrating a Transformer model with InfoNCE loss and model distillation. The proposed method focuses on single-stage retrieval using cosine similarity, ensuring efficient and effective embedding of topics and contents. We elaborate on the model architecture, data preprocessing techniques, evaluation metrics, and experiment results to demonstrate the robustness of our approach.

3.1. Transformer Model

The retrieval model is built on the Transformer encoder, known for handling sequential data. Each topic and content is tokenized, limited to 96 tokens. The architecture is shown in Figure 1. The encoder consists of layers with two sub-layers: a multi-head self-attention mechanism and a feed-forward network, both with residual connections and layer normalization. The multi-head attention focuses on various parts of the input, capturing diverse contexts. Attention scores are computed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The multi-head attention mechanism is defined as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2}$$

Each head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where W_i^Q, W_i^K, and W_i^V are projection matrices for the i-th head. The position-wise feed-forward network applies two linear transformations with a ReLU activation to each position:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{4}$$

The embeddings generated by the encoder are used to measure the cosine similarity between topics and content. Cosine similarity, defined as the cosine of the angle between two vectors, is given by:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{5}$$

 doi:10.20944/preprints202411.1070.v1

where A and B are the embedding vectors of the topic and content, respectively.

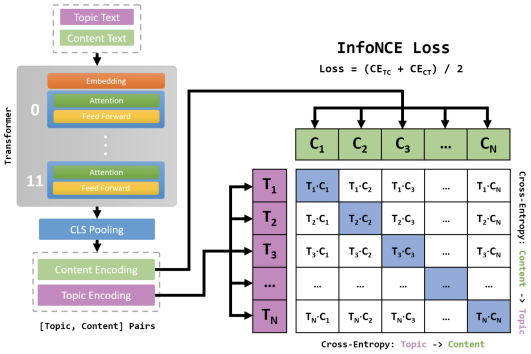

Figure 1. Transformer architecture for topic-content embedding.

3.2. Loss Function

The primary loss function used in our model is the InfoNCE (Information Noise Contrastive Estimation) loss, which is particularly suited for contrastive learning tasks. The purpose of InfoNCE loss is to ensure that embeddings of positive pairs remain close, while those of negative pairs are pushed apart.

The InfoNCE loss for a single positive pair (i, j) is formulated as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\cos(\theta_{i,j})/\tau)}{\sum_{k=1}^{N} \exp(\cos(\theta_{i,k})/\tau)} \tag{6}$$

In this equation, τ is the temperature controlling distribution smoothness, $\cos(\theta_{i,j})$ is the cosine similarity between the i-th topic and j-th content, and N is the number of negative samples. The loss maximizes positive pair similarity and reduces negative pair similarity.

To enhance performance, we apply model distillation, where the student model mimics the pre-trained teacher model. The distillation process helps the student model generalize better and retain the teacher model's performance. The distillation loss is defined as:

$$\mathcal{L}_{\text{distill}} = \text{MSE}(S(x), T(x)) \tag{7}$$

Here, $S(x)$ and $T(x)$ The outputs of the student and teacher models are compared using MSE (Mean Squared Error). The student model is initialized with the teacher model's weights and learns to match its predictions during training. By jointly applying InfoNCE loss and distillation loss, our model embeds topics and content in a shared space, keeping relevant pairs tightly aligned and irrelevant pairs well-separated, thereby improving the model's retrieval accuracy and robustness.

3.3. Data Preprocessing

The training data involves splitting topics and contents, ensuring minimal overlap of similar contents across different buckets. We create 10 buckets and distribute the contents to minimize overlap. The preprocessed data maintains a $n \times m$ relationship between topics and contents.

We employ data augmentation through language translation (e.g., English to French). However, we address the noise introduced by translations by switching languages every few epochs to avoid overfitting on specific language patterns. The data preprocessing strategy in Figure 2.

Figure 2. The data preprocessing strategy.

 doi:10.20944/preprints202411.1070.v1

3.4. Model Distillation

Model distillation is a crucial step in our approach to enhance the efficiency of the retrieval system without significantly compromising its performance. This section outlines the distillation process from the pre-trained Teacher Model to the Student Model's deployment.

We begin with a pre-trained Transformer as the Teacher Model. Consisting of 12 Transformer layers, the Teacher Model has been trained on an extensive dataset. The embeddings produced by this model serve as high-quality representations of the input data.

To create a more efficient version of this model, we generate a Student Model by removing half of the Transformer's layers. This reduction results in a Student Model with 6 layers. The primary motivation behind this layer reduction is to accelerate the training and inference processes while maintaining a reasonable balance between speed and accuracy.

The distillation process involves the following steps:

3.4.1. Weight Initialization

The Student Model is initialized with the Teacher Model's pre-trained weights, providing a strong foundation from the Teacher's training.

3.4.2. Distillation Training

The Student Model is trained using MSE loss to minimize the difference between its outputs and the Teacher Model's. The MSE loss is defined as:

$$\mathcal{L}_{\text{distill}} = \text{MSE}(S(x), T(x)) \tag{8}$$

Here, $S(x)$ and $T(x)$ represent the outputs of the Student and Teacher Models. Minimizing this loss allows the Student to replicate the Teacher's behavior and performance.

3.4.3. Performance Evaluation

During the distillation process, we closely monitor the performance of the Student Model. Experiments reveal that employing a6-layer Student Model results in only a marginal decline in performance compared to the12-layer Teacher Model. This slight drop in accuracy is an acceptable trade-off considering the significant gains in computational efficiency.

The 6-layer Student Model is chosen based on evaluations, as further reduction decreases performance noticeably. Thus, the 6-layer configuration strikes the best balance between maintaining accuracy and enhancing speed.

After the distillation training, the Student Model is used for final training and prediction tasks. This distilled model leverages the knowledge embedded in the Teacher Model while operating more efficiently, This makes it well-suited for real-world applications where computational resources are limited. The entire pipeline is depicted in Figure 3.

In summary, the model distillation process involves leveraging a pre-trained Teacher Model, reducing the Transformer's layers to create an efficient Student Model, initializing the Student Model with the Teacher Model's weights, and fine-tuning it using the MSE loss. The result is a model that balances efficiency and effectiveness, making it a practical solution for information retrieval tasks.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 November 2024 doi:10.20944/preprints202411.1070.v1

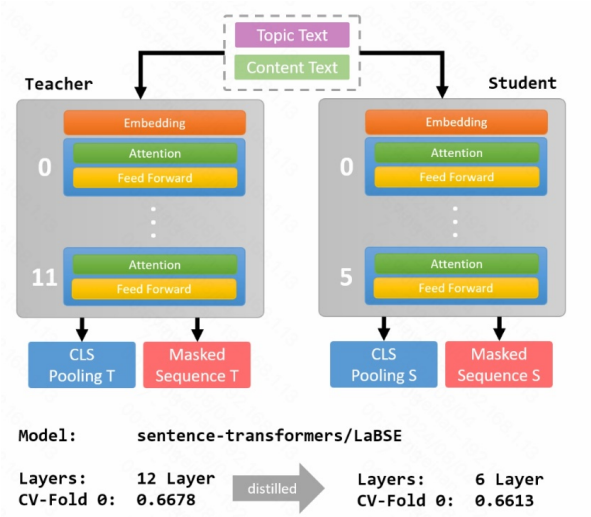

Figure 3. Model Distillation.

4. Evaluation Metric

We evaluate our retrieval model using standard IR metrics, including MRR, NDCG, Accuracy, and Recall.These metrics provide a comprehensive evaluation of the model's effectiveness in retrieving relevant content.

4.0.1. MRR

MRR is a measure used to evaluate the effectiveness of a retrieval system based on the rank positions of the first relevant document. It is defined as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{9}$$

In this formula, $|Q|$ is the total number of queries, and rank_i indicates the rank of the first relevant document for the i-th query.

4.0.2. Accuracy

Accuracy is a widely used metric that measures the proportion of correctly predicted instances over the total instances. In the context of retrieval, it is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

4.0.3. Recall

Recall, or sensitivity, evaluates the proportion of actual positive instances that are correctly recognized by the model. It is particularly useful for evaluating the ability of the model to retrieve all relevant documents. Recall is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

where TP is the number of true positives and FN is the number of false negatives.

 doi:10.20944/preprints202411.1070.v1

These metrics collectively provide a thorough evaluation of the retrieval model, highlighting both its strengths and areas for improvement. By employing these metrics, we can better understand the performance of our model in various aspects of retrieval tasks.

5. Experimental Results

The experimental results, obtained by assessing the performance of each model on various tasks, are presented here. The results are summarized in Table 1.

Our best-performing model, the Transformer enhanced with InfoNCE Loss and model distillation, reaches the highest scores across all metrics, with an MRR of 0.7245, Accuracy of 0.851, and Recall of 0.825. This illustrates the value of knowledge distillation in transferring knowledge from a larger model to a smaller one, creating a more efficient and precise model.

These results highlight the progressive improvements achieved by incorporating advanced techniques such as transformer architectures, contrastive learning, in-context learning, and model distillation. Each step contributes to refining the model's ability to capture relevant information and make accurate predictions.

Table 1. Experimental Results

Model	MRR	Accuracy	Recall
VecModel + tf-idf	0.5781	0.781	0.709
xlm-roberta-base + tf-idf	0.6891	0.812	0.789
Transformer + InfoNCE Loss	0.7121	0.823	0.801
Transformers + xlm-roberta-base + ICT	0.72	0.83	0.81
Transformer + InfoNCE + distillation	0.72	0.85	0.82

6. Conclusion

In summary, our Transformer-based retrieval system achieves significant improvements in retrieval performance by integrating InfoNCE loss and model distillation. TThe results of our experiments affirm the success of this approach, presenting it as a highly promising solution for recommendation retrieval tasks.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
2. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. International conference on machine learning. PMLR, 2020, pp. 1597–1607.
3. Li, S. Harnessing Multimodal Data and Mult-Recall Strategies for Enhanced Product Recommendation in E-Commerce. *Preprints* **2024**.
4. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; others. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
5. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**.
6. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* **2019**.
7. Li, S.; Zhou, X.; Wu, Z.; Long, Y.; Shen, Y. Strategic Deductive Reasoning in Large Language Models: A Dual-Agent Approach. *Preprints* **2024**.
8. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

 doi:10.20944/preprints202411.1070.v1

10. Lu, J. Optimizing E-Commerce with Multi-Objective Recommendations Using Ensemble Learning. *Preprints* **2024**.
11. Kiela, D.; Grave, E.; Joulin, A.; Mikolov, T. Efficient large-scale multi-modal classification. Proceedings of the AAAI conference on artificial intelligence, 2018, Vol. 32.
12. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* **2019**, *32*.
13. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. International conference on machine learning. Pmlr, 2021, pp. 8821–8831.
14. Wang, Y.; Wang, D. An Entropy-and Attention-Based Feature Extraction and Selection Network for Multi-Target Coupling Scenarios. 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE). IEEE, 2023, pp. 1–6.
15. He, C.; Yu, B.; Liu, M.; Guo, L.; Tian, L.; Huang, J. Utilizing Large Language Models to Illustrate Constraints for Construction Planning. *Buildings* **2024**, *14*, 2511.
16. Yu, H.; Yu, C.; Wang, Z.; Zou, D.; Qin, H. Enhancing Healthcare through Large Language Models: A Study on Medical Question Answering. *arXiv preprint arXiv:2408.04138* **2024**.
17. Zhang, L.; Li, L.; Wu, D.; Chen, S.; He, Y. Fairness-Aware Streaming Feature Selection with Causal Graphs. *arXiv preprint arXiv:2408.12665* **2024**.
18. Yan, H.; Xiao, J.; Zhang, B.; Yang, L.; Qu, P. The Application of Natural Language Processing Technology in the Era of Big Data. *Journal of Industrial Engineering and Applied Science* **2024**, *2*, 20–27.
19. Zhang, B.; Xiao, J.; Yan, H.; Yang, L.; Qu, P. Review of NLP Applications in the Field of Text Sentiment Analysis. *Journal of Industrial Engineering and Applied Science* **2024**, *2*, 28–34.
20. Xia, Y.; Liu, S.; Yu, Q.; Deng, L.; Zhang, Y.; Su, H.; Zheng, K. Parameterized Decision-making with Multi-modal Perception for Autonomous Driving. *arXiv preprint arXiv:2312.11935* **2023**.