NR-CLIP: CLIP-Guided Multimodal News Recommendation via Multi-View Learning

Ying Guo (郭颖)[1,5], *Member*, *CCF*, *IEEE*, Shu-Ting Hu (胡淑婷)[1], Min-Jing Yu (余旻婧)[2], *Member*, *CCF*, *ACM*, *IEEE*, Ran Yi (易冉)[3], *Member*, *ACM*, *IEEE*, Qi Wang (王崎)[4], *Member*, *CCF*, *IEEE*, Jie Liu (刘杰)[1], *Senior Member*, *CCF*, and Yong-Jin Liu (刘永进)[5,*], *Senior Member*, *ACM*, *IEEE*

[1]School of Artificial Intelligence and Computer Science, North China University of Technology, Beijing, 100144, China

[2]College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China

[3]School of Computer Science, Shanghai Jiao Tong University, Shanghai, 200240, China

[4]College of Computer Science and Technology, Guizhou University, Guiyang, 550025, China

[5]Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

E-mail: guoying@ncut.edu.cn; hushuting@mail.ncut.edu.cn; minjingyu@tju.edu.cn; ranyi@sjtu.edu.cn; qiwang@gzu.edu.cn; liujie@ncut.edu.cn; liuyongjin@tsinghua.edu.cn

Abstract

In the era of social media, the evolution of news has diversified its format, incorporating texts, images, and videos. However, the majority of news recommendation methods focus solely on text data, overlooking the substantial role of news images. This paper introduces the news recommendation based on CLIP (NR-CLIP) method, which is a CLIP-guided information system with enhanced multimodal news recommendation via multi-view learning. Specifically, our method employs a CLIP encoder to embed textual and visual information into the same feature space of the neural network, where a unified news textual representation is learned by treating titles, categories, subcategories and body as different views of news. In addition, feature enhancements are applied to fully fuse textual and visual features. Finally, click history and user representations are embedded to predict the click probability of candidate news. Extensive comparisons and in-depth analyses with state-of-the-art news recommendation methods have been presented on V-MIND (Visual Microsoft News Dataset), which provides the visual information on the basis of the classic MIND (Microsoft News Dataset), demonstrating that our proposed method effectively improves the performance of news recommendation.

Context

News recommendation has gained significant attention in multi-media owing to its high timeliness, rich semantic information, and diverse user interests, compared with other recommendations such as commodities, movies, music, etc. Traditional methods primarily concentrate on textual attributes, neglecting visual information. Users are motivated to click on news articles not solely due to their interest in textual content but also because of the appeal of accompanying images. Existing multimodal approaches (e.g., MM-Rec) fail to fully capture cross-modal interactions or leverage diverse textual elements (e.g., categories, subcategories, and bodies) for comprehensive news representation. This study addresses these limitations by integrating multimodal features and multi-view learning to enhance recommendation accuracy.

Objective

The primary objective of this work is to propose NR-CLIP, a CLIP-guided multimodal news recommendation framework, to overcome the limitations of existing methods by:

1. Aligning textual and visual features in a shared semantic space using CLIP.
2. Integrating multi-view textual elements (titles, categories, subcategories, and bodies) via attentive pooling.
3. Enhancing cross-modal interactions through scaled dot-product attention mechanisms.

The goal is to achieve more personalized and accurate news recommendations by capturing both textual and visual user interests.

Method

The NR-CLIP framework consists of three core components:

1. Multimodal News Encoder:

- CLIP-based feature extractor: Maps text (titles, categories, subcategories, bodies) and images into a shared feature space.
- Multi-view attentive pooling: Assigns attention weights to different text types and aggregates them into a unified textual representation.
- Interactive feature enhancement: Uses scaled dot-product attention to fuse textual and visual features.

2. User Encoder: Learns user representations from historical click behaviors via attention mechanisms.

3. Click Predictor: Predicts click probabilities using inner products of user and candidate news representations.

Key techniques include contrastive learning with CLIP, attention-based fusion, and cross-modal feature refinement.

Results

1. Performance on V-MIND dataset: NR-CLIP achieves AUC=0.648, NDCG@10=0.351, outperforming baseline models like MM-Rec (AUC=0.644) and AT (AUC=0.645).

2. Ablation studies:

Removing CLIP reduces AUC by 3.7%.

Excluding multi-view attentive pooling decreases AUC by 0.7%.

Omitting interactive feature enhancement drops NDCG@10 by 2.8%.

3. Modality contributions:

Image-only (AUC=0.645) and text-only (AUC=0.641) underperform multimodal fusion (AUC=0.648).

Removing body text reduces AUC by 2.1%, while removing categories reduces it by 0.5%.

Conclusions

This paper introduces an innovative multimodal news recommendation approach termed NR-CLIP, which harnesses both textual and visual data to model news for recommendation purposes. Initially, news images and four distinct types of texts are fed into the CLIP encoder to facilitate the learning of news representations, to obtain the semantic alignments within the same feature space. Furthermore, multi-view attentive pooling is incorporated to amalgamate diverse types of news textual information, thereby fostering the acquisition of a unified news textual representation. Lastly, a novel feature enhancement method is deployed to bolster and enhance the capture of interactions between textual and visual elements. This constitutes a significant advancement in existing research. Exhaustive experimentation corroborates that NR-CLIP substantially enhances the efficacy of news recommendation.

Despite the advancements in news recommendation systems, the consideration of news quality, especially in multimodal contexts, remains largely underexplored. Our future research aims to bridge this gap by developing models that leverage multimodal data and assess the quality of news content, thereby creating more robust and reliable systems that better cater to users' needs and preferences. We acknowledge the limitations of our proposed method and the importance of addressing news quality and potential biases in recommendations. News quality is a critical factor influencing user trust and engagement. To mitigate these issues, future work will integrate quality assessment metrics to evaluate the credibility and reliability of news sources and incorporate diversity and fairness-aware algorithms to ensure a balanced representation of viewpoints and prevent the amplification of biased content. These enhancements will strengthen the ethical stance of our recommendation system and align it with user expectations for unbiased and high-quality news content.