Summary

**Title**: Improving Personalized Explanation Generation through Visualization

**Authors**: Shijie Geng, Zuohui Fu, Yingqiang Ge, Lei Li, Gerard de Melo, Yongfeng Zhang from Rutgers University, Hong Kong Baptist University, HPI/University of Potsdam

**Problem Addressed in this work**: In modern recommender systems, there are usually comments or reviews from users that justify their ratings for different items. Trained on such textual corpus, explainable recommendation models learn to discover user interests and generate personalized explanations. Though able to provide plausible explanations, existing models tend to generate repeated sentences for different items or empty sentences with insufficient details. This paper proposes an approach called Multimodally-Enhanced Transformer for Explainable Recommendations (METER), which generates visually enhanced explanations based on conditional image generation and text-image matching.

**Summarize prior work:** There are two previous applications of using METER. One for visually Guided Language Learning and another for Generating Explanations for Recommendations. Using visual information to improve language tasks through visual grounding, involves aligning visual and textual information to improve performance. DALL-E is a recent approach that merges text and visual tokens as a single stream of data and employs a universal Transformers to autoregressively model the multimodal stream. For generating explanations for recommendations, Early approaches to this task involved making latent factor models interpretable. Recently, neural models have been proposed to explain recommendations based on user reviews. There have also been attempts to generate purely visual explanations. However, sentence-based methods have been at the center of attention in recent times. METER is a new approach that integrates textual and visual features to provide multimodal explanations. METER is the first approach to use vision to improve textual explanation generation.

**Unique Contributions of this paper**: To the best of their knowledge, METER is the first exploration of a multimodal explainable recommender system that jointly generators rating scores, textual explanations, and images. The system will also be promising in creative advertising applications. By immersing the model into a multimodal environment, we help it explore the real-world concepts mentioned in the text explanations and in turn enable it to generate more diverse and faithful natural language rationales that are consistent with visual grounding.

**Describe how the authors evaluated their work**: They conducted their evaluations from three perspectives - Explanation generation performance, test-image matching performance and rating description performance. For explanation performance, they measured the text quality, diversity and explainability of the generated explanations. Automatic and manual forms of evaluation are adopted. For explanation performance, text quality, diversity, and explainability of the generated explanations are measured using BLEU-1, BLEU-4, ROUGE-1, ROUGE-2, Unique Sentence Ratio (USR), Feature Diversity (DIV), Feature Matching Ratio (FMR), and Feature Coverage Ratio (FCR). CLIPScore (CS) is used to measure text-image matching, while Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used for rating prediction performance. The aim of the recommendation experiment is to prove that the predicted rating scores are sufficiently accurate to merit explanation generation, as inaccurate predictions will result in less meaningful explanations.

**Author Citations:** Shijie Geng (888), Zuohui Fu (1192), Yingqiang Ge (908), Gerard de Melo (6918), Yongfeng Zhang (8448). Yongfeng Zhang has the highest number of citations.

**Importance**: This work is significant as it introduces a new approach to generate personalized and visually enhanced explanations in recommender systems, which can have practical applications in various fields such as advertising. The evaluation metrics used in this paper can also serve as a standardized method for evaluating machine translation outputs, which can lead to more accurate comparisons between different machine translation systems.