Conceptual Activation Vectors Meets Embedding-based Recommender Systems

by Langtian Ma

1 Introduction

In the previous project, we explored the application of the Integrated Gradients method [4] to attribute feature importance for specific predictions in a LightGCN recommender system. This approach effectively identified the feature contributions of both users and movies. However, Integrated Gradients is computationally expensive when used to analyze feature importance across all users and movies. In this report, we demonstrate that the Concept Activation Vector (CAV) method [3] is particularly well-suited for embedding based recommender systems, offering significantly higher computational efficiency and a more intuitive interpretation of feature influence.

2 Problem Formulation

Building on the previous project, we trained a LightGCN model [2] on the MovieLens 1M dataset [1]. In this work, we treat movie genres as interpretable "concepts" within the CAV framework and aim to explore the following key questions:

- What is the geometric structure of movie genres in the learned embedding space?
- To what extent does the model capture a user's preference for a particular genre?
- How well does the model's learned preference align with the user's actual behavior?

These investigations not only help end-users better understand the rationale behind movie recommendations, but also provide valuable insights for developers to debug and improve the underlying model.

3 Method Application

3.1 CAV for Embedding-based Recommender Systems

We find that the CAV framework is particularly well-suited for embedding-based recommender systems, and in this setting, it reduces to a simple and intuitive form. This is because both user and item embeddings are pre-trained, and the model's inference process consists of a simple inner product between these embeddings. Furthermore, unlike in deep neural networks, we do not need to select a specific "layer" for extracting CAVs, as there is only one meaningful embedding space in such models.

Formally, let C denote a movie genre, and let E_C be the set of item (movie) embeddings corresponding to genre C, with E_N denoting the embeddings of all other items. We train a logistic regression classifier to distinguish between these two sets. The learned coefficients $\mathbf{v}_C$ of this classifier then define the Concept Activation Vector (CAV) for the genre C.

To evaluate the model's sensitivity to this concept, we derive the *conceptual sensitivity* of a user u with respect to genre C. Let $\mathbf{e}_u$ denote the embedding of user u, and let the model's prediction function be defined as

$$h_u(\mathbf{e}_m) = \mathbf{e}_u^\top \mathbf{e}_m,$$

where $\mathbf{e}_m$ is the embedding of a candidate item. Then, the conceptual sensitivity of user u to concept C is given by the directional derivative of h_u along $\mathbf{v}_C$:

$$S_C(\mathbf{e}_u) = \nabla_{\mathbf{e}_m} h_u(\mathbf{e}_m)^\top \mathbf{v}_C = \mathbf{e}_u^\top \mathbf{v}_C.$$

This closed-form expression enables efficient computation of conceptual sensitivity and directly reflects a user's preference for genre C. A higher value indicates stronger alignment with the concept, while a negative value suggests disinterest. This formulation offers a simple yet effective way to interpret user preferences in the embedding space.

3.2 Experimental Results

We first investigate the geometric structure of the CAVs in the embedding space. Figure 1 shows a 2D PCA projection of the CAVs for four genres. The vectors for *Thriller* and *Horror* form a small angle, indicating strong semantic similarity. In contrast, the vectors for *Drama* and *Comedy* form obtuse angles with both *Thriller* and *Horror*, suggesting they are semantically distant from the latter two. This distribution highlights the model clusters related genres while separating dissimilar ones in the latent space, which aligns well with our intuition.

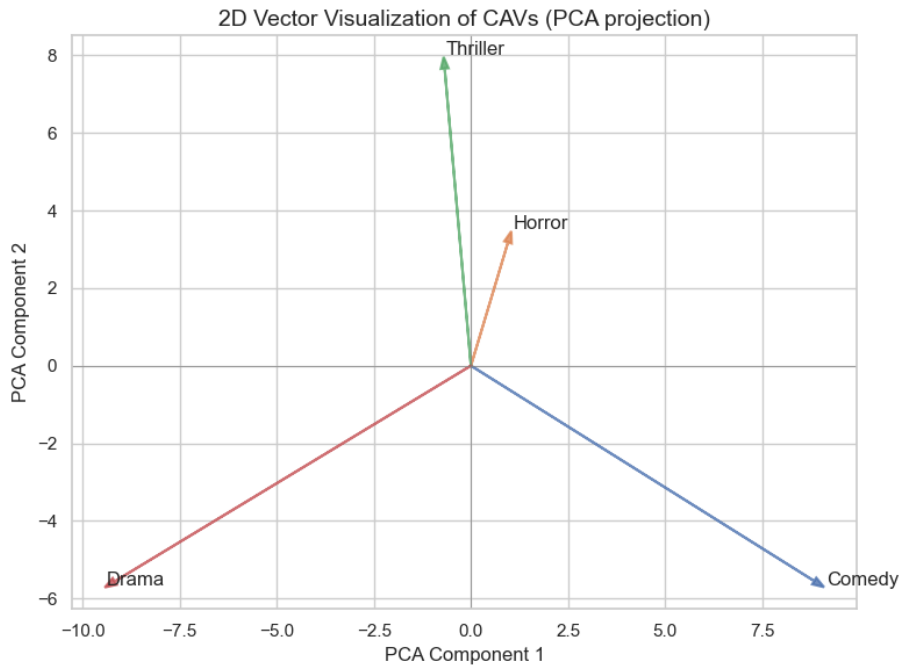

Figure 1. Visualization of CAVs

Conceptual sensitivity serves as a proxy for the user's learned preference toward a given movie genre. By comparing it with the user's actual preference, we can assess whether the model captures user interests accurately. Ideally, a user's conceptual sensitivity for a genre should be positively correlated with their average rating for movies of that genre.

For example, we compute each user's conceptual sensitivity to the "Comedy" genre and compare it against their average rating of comedy movies. As shown in Figure 2, there is no significant correlation between the two, suggesting that the model may fail to capture genre-level user preferences effectively.

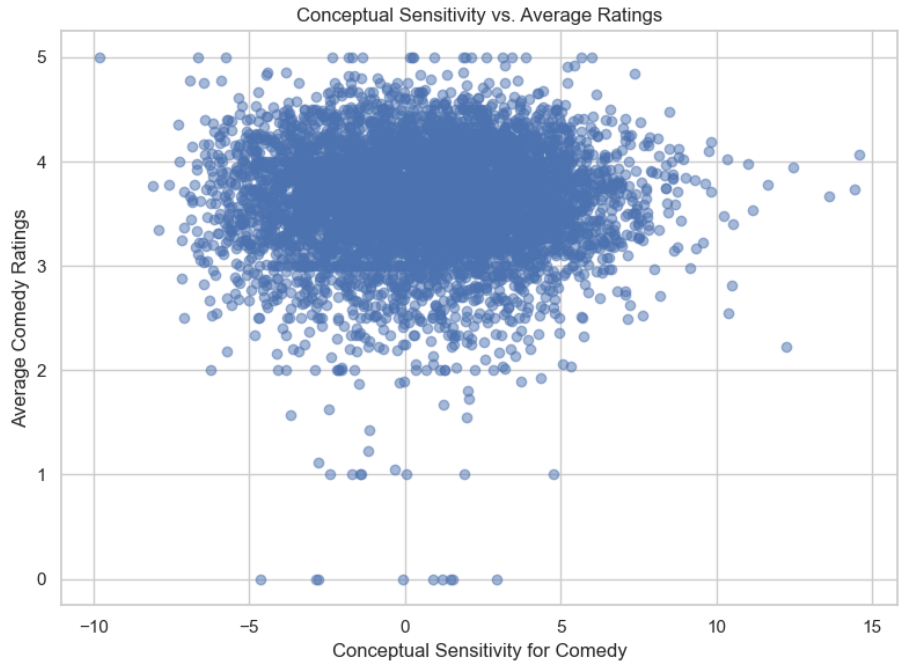

Figure 2. Conceptual Sensitivity v.s. Average Ratings for Comedy Movies

4 Discussion

In this project, we demonstrated that the Concept Activation Vector (CAV) is a powerful and efficient tool for interpreting embedding-based recommender systems. It enables concept-level explanations by quantifying how user embeddings align with interpretable directions in the latent space, offering both insight into model behavior and potential directions for improvement.

Our experiments also revealed intriguing phenomena that merit further investigation. In particular, we observed that the learned genre preferences of users—measured via conceptual sensitivity—do not always align with their actual preferences based on historical ratings. This discrepancy raises a natural and important question: can we improve recommendation quality by explicitly encouraging this alignment during training?

Addressing this question could lead to models that are not only more accurate, but also more interpretable and trustworthy. We leave this direction as a promising avenue for future work.

Bibliography

[1] F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. 5(4):1-19.

[2] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation.

[3] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668-2677. PMLR.

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319-3328. PMLR.