Context-Aware Compatibility Prediction

Praguna Manvi
Team 11 : Pushpa
Topics in Deep Learning

Vivek Talwar
Team 11 : Pushpa
Topics in Deep Learning

Ankita Maity
Team 11 : Pushpa
Topics in Deep Learning

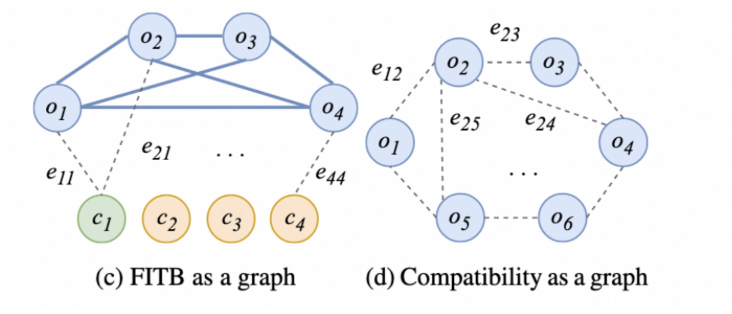
Abstract

Compatability prediction is applied in many fashion and e-commerce recommendation systems. Modeling contextual information is central in determining compatibility. Modern GNN-based approaches model contextual information effectively and provide superior results. In this project, we review the state-of-the-art paper for compatibility prediction on the Polyvore dataset [3] and improve on it by introducing multi-modality.

1 Introduction

1.1 Problem Definition

Compatibility decides whether an item can fit in a set of other entities. Modeling Compatability Prediction involves predicting the membership of an object in a collection (context). This calls for associating an item/ object in its context and modeling the context itself. Graphs are a natural choice for modeling associations between connected entities. From this modeling, membership is determined using various distance measures. Previous approaches have used siamese networks that model similarity rather than compatibility, and GNNs have demonstrated superior performance on this problem. The state-of-the-art method for compatibility using visual features is reviewed here.



(c) FITB as a graph (d) Compatibility as a graph

Figure 1: Modeling Compatibility using Nodes and Edges

The nodes are represented using features describing the item. The association between them is established via context or when they occur together in fashion choices that are obtained from the dataset used in this project. The membership of an item is decided using a decoder module that obtains encoder representation from GNN layers and calculates:

$$p = \sigma(||h_i - h_j||) \tag{1}$$

where h_i and h_j are last GNN layer feature representations where hidden weights are optimized using binary-cross-entropy loss. This is an edge prediction problem that can either be tested as FITB (Fill

Preprint. Under review.

in the blank) where the correct choice has to be made for a context or complete graph compatibility prediction, as shown in the figure 1.

1.2 Example

A simple real-world illustration of the problem is shown in figure 2. A full-arm T-shirt has to be matched for compatibility with Jeans or a long skirt. The second part of the figure shows that both sides have learned their features in their contexts after being trained using deep GNN layers. The model would choose Jeans as it has appeared in a similar context as that of a full-arm black T-shirt.

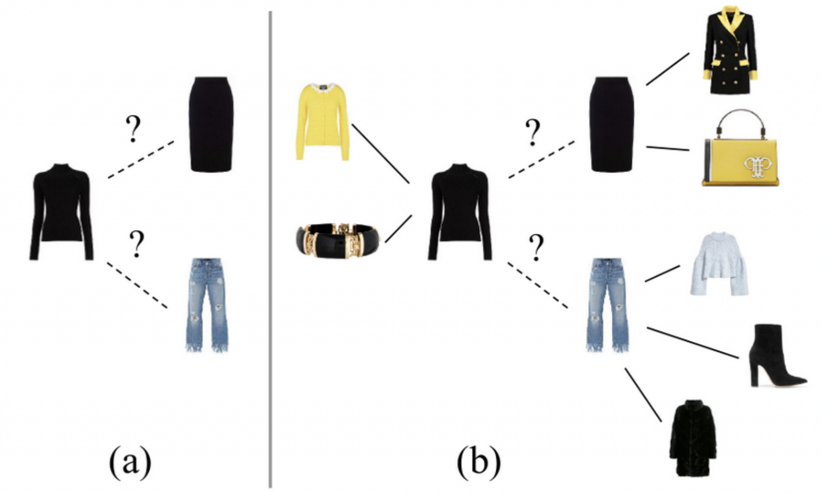


Figure 2: Illustration of Context-Based Compatibility

1.3 Motivation and Challenges

The problem has many applications in modern recommendation systems. Using graphs for modeling the context provides significant improvements in results. The complexity in modeling is high because of the variety of data in which compatibility problems can be defined, lack of clarity, and susceptibility to bias. Multi-modal methods can be explored to add more contextual information.

2 Main Contributions

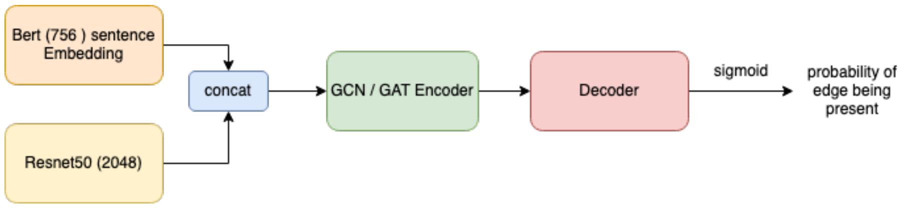

Bert (756) sentence Embedding
Resnet50 (2048)
concat
GCN / GAT Encoder
Decoder
sigmoid
probability of edge being present

Figure 3: Proposed Model

The proposed model 3 follows a multi-model approach that extracts textual as well as image features from the input. These features are combined using concatenation (arrived after testing out other approaches) which is encoded using a standard GAT / GCN (three-layered). This passes to the decoder, which outputs the logits for positive and negative examples, which are improved upon by optimizing cross-entropy-loss.
The main contribution of this project includes :

- Reproducing the results of the Context-Aware Visual Compatibility Prediction paper [1], which has state-of-the-art results in context-based compatibility tasks.
- Ablation studies on a few important optimization parameters, namely degree and hidden weight sizes, provide a deep insight into the inner working of the referred approach.
- Experimentation of different GNN architectures on this problem. We also compare their performance against the state-of-the-art approach.
- Introducing a novel multi-modal recommendation system for context compatibility that beats state-of-the-art results by more than 2%.

3 Experiments

The reference work which used GCN [2] is reproduced as shown in figure 4. The model is trained for 4000 epochs, and the nature of training is self-supervised. We use negative sampling to generate labels for unlabeled edges for the sparse dataset graph. The accuracy obtained is comparable to the three neighbor approach discussed in the referred state-of-the-art paper.

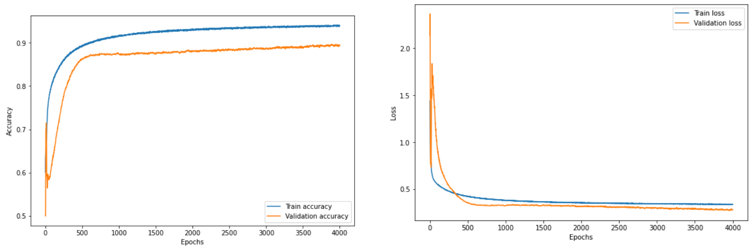


Figure 4: Reproducing Results of [1] on Polyvore [3] Dataset

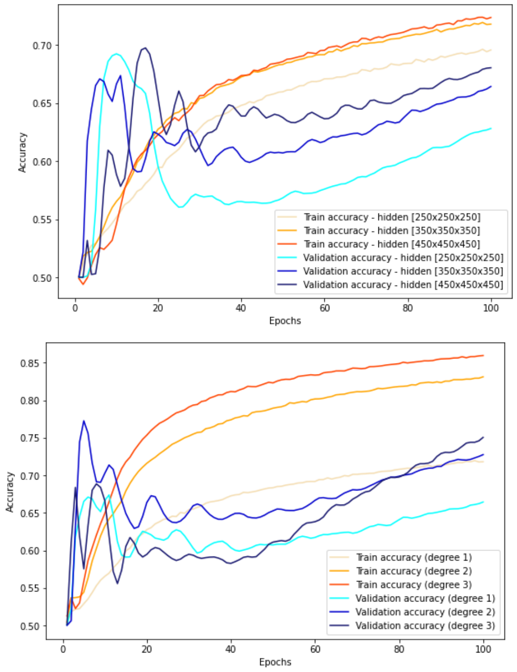


Figure 5: Ablation Studies on [1]

Few ablation studies were conducted on the reproduced model by varying some of the parameters and comparing its performance. We observe from figure 5 that the model performs best at degree 1, where degree refers to the depth of the neighborhood. The hidden layer dimensions of [350, 350, 350] for each GCN layer have the best performance. The following parameters are set in the reproduced model to attain the best results. These experiments are addons to deepen the understanding of the state-of-the-art approach.

Table 1: Comparison of Accuracy performance

Model	Accuracy
Visual Compatability[1]	90.01
Visual Compatability [1] GAT	90.007
Visual Compatability [1] GAT + Textual Features	93.0561
Visual Compatability [1] + Textual Features	92.0621

As a part of this project, different GNN architectures were explored. However, GAT [4] is considered with multi-heads (two) and compared to the reproduced model as it generalizes aggregation by using self-attention from its neighbors. The performance of GAT is comparable to the results obtained from the state-of-the-art model. The accuracy increases when text features extracted from Bert-based sentence embeddings are introduced. The model combines these features with the existing image features extracted using a Resnet50. The overall feature size in each node becomes 2804. After experimenting with various combinations of GAT and on the reproduced model, we observe that GAT + text feature-based approach has the highest performance, taking additional contextual information better than regular GCN as shown in table 1.

4 Novel Ideas

This project introduced a multi-model context compatibility approach that beats state-of-the-art models for the FITB task. We further performed ablation studies on the effect of increasing the degree of GCN models. The proposed method is less explored, and through this project, we have shown that the multi-model approach can add additional contextual information, which improves the performance even with a simple composition layer. The speed of GCNs during inference was observed to be high, which can be used to design practical solutions that do 1:n contextual compatibility prediction.

References

[1] Guillem Cucurull, Perouz Taslakian, and David Vázquez. Context-aware visual compatibility prediction. *CoRR*, abs/1902.03646, 2019.

[2] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2016.

[3] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European conference on computer vision (ECCV)*, pages 390–405, 2018.

[4] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2017.