LEARNING TYPE-AWARE EMBEDDINGS FOR FASHION COMPATIBILITY WITH CLIP

Dae Lim Chung
Columbia University
dc3666@columbia.edu

ABSTRACT

The realm of fashion compatibility is an evolving landscape in the domain of machine learning, especially in leveraging deep learning techniques for nuanced tasks like creating compatible outfits. In our study, we extend the work of Vasileva et al. (1) in "Learning Type-Aware Embeddings for Fashion Compatibility" by integrating OpenAI's CLIP as a central component for both text and image embeddings. This strategic replacement is premised on CLIP's robust, pre-trained features, which encapsulate a wide range of styles and contexts, enabling more nuanced and accurate representations of fashion items. Additionally, we shift the focus towards minimizing Euclidean distances in the embedding space rather than employing the original approach of generalizing distance metrics through a fully-connected layer. This simplification aligns with the geometric intuition of embedding spaces, where distances more directly and transparently represent similarities and dissimilarities. Our modifications have culminated in a notable improvement in the model's performance, enhancing the Fill In The Blank (FITB) task accuracy and Area Under Curve (AUC) metrics by 7.6% and 5% respectively. These results underscore the efficacy of our approach, marking a stride in the quest for advanced fashion recommendation systems.

1 INTRODUCTION

In envisioning a world where artificial intelligence empowers everyone to dress in alignment with their true character and style, we confront the current limitations of AI systems in the domain of fashion outfit generation. While current systems often base recommendations on prevailing fashion trends, these can be misaligned with an individual's ambience, partly due to the lack of extensive, high-quality data encompassing the broad scope of fashion history.(2) Moreover, AI stylists have yet to demonstrate the intuitive competence of human stylists, often lacking in novelty and the natural flair that comes with human visual intuition for clothing.(2)

The paper by Vasileva et al. (1) introduced an innovative approach for understanding fashion compatibility using type-aware embeddings. This method adeptly addresses the challenges posed by the fashion domain, such as the nuanced interpretation of style and the need for contextual understanding. However, advancements in machine learning, particularly in natural language and image processing, present new opportunities for enhancing this model.

Our work is grounded in two primary improvements: the integration of OpenAI's CLIP for text and image embeddings and the focus on minimizing Euclidean distances in the embedding space. CLIP, with its extensive training on a diverse range of internet-sourced data, offers a more comprehensive and contextually rich embedding than traditional methods. This enables our model to understand and represent fashion items with greater accuracy and depth, particularly beneficial for the diverse and rapidly evolving fashion domain.

Furthermore, by focusing on minimizing Euclidean distances, our approach taps into the inherent geometric properties of embedding spaces. This strategy simplifies the model's complexity and enhances interpretability, allowing for a more straightforward assessment of compatibility based on distance metrics. This contrast with the original approach of using a fully-connected layer as a generalized distance function provides a more transparent and intuitive understanding of item relationships within the embedding space.

The efficacy of these improvements is demonstrated by a 7.6% increase in performance on the FITB task and 5% seen in AUC metrics. This enhancement not only validates our methodology but also sets a new benchmark for future research in fashion compatibility and recommendation systems. Our work represents a significant step forward in harnessing the power of advanced AI techniques for the dynamic and visually rich field.

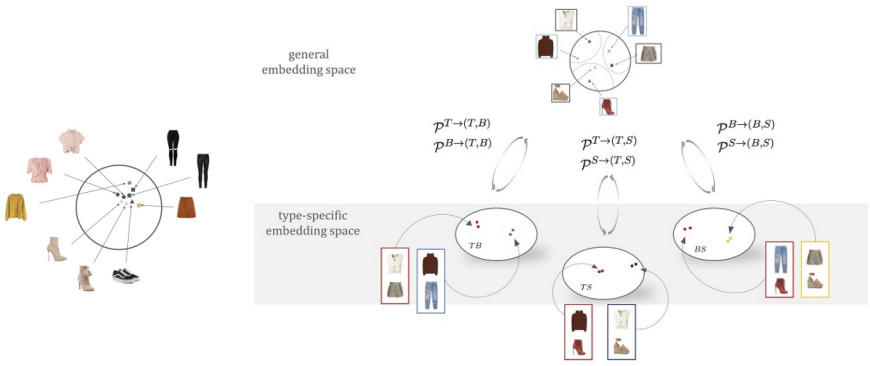

Figure 1: Left:Objects of all types sharing one embedding space Right:Type Aware Embedding where we first learn general embedding space for all pairs. Then we project from the general embedding to sub spaces according to their type

(1)

2 Related Works

The study of fashion compatibility and outfit recommendation has seen significant progress with the advent of deep learning techniques. Our work builds upon these advancements, notably drawing inspiration from Vasileva et al.'s study titled "Learning Type-Aware Embeddings for Fashion Compatibility" (1). This foundational paper presents a novel approach to understanding fashion compatibility using type-aware embeddings and a triplet loss structure. In their method, item types (e.g., shirts, pants, shoes) are considered when learning embeddings, enabling the model to capture the nuanced relationships between different fashion items. The triplet loss structure used enhances the model's ability to distinguish between compatible and incompatible items. This methodology is pivotal in our research, where we aim to refine and extend these concepts.

A key aspect of Vasileva et al.'s(1) work is the utilization of the custom Polyvore dataset, which comprises diverse fashion items and outfits. The dataset includes images and descriptions of individual fashion items, along with information on how they are paired in outfits, providing a comprehensive framework for analyzing fashion compatibility.

While our research is primarily influenced by Vasileva et al.'s work(1), we diverge from some other methodologies in the field, such as those involving Long Short-Term Memory (LSTM) networks. For instance, the approach by Han et al.(3), as detailed in their paper "Learning Fashion Compatibility with Bidirectional LSTMs", represents a different trajectory in understanding fashion compatibility. Their method focuses on the sequential analysis of fashion items in an outfit using LSTMs, offering insights into how the order and interrelationship of items contribute to the overall aesthetic of an outfit. Moreover, we point out that the polyvore outfits dataset was first constructed by authors in (3) which our dataset from Vasileva et al is variant of. ZoZo research (4) has made improvements to the sequential approach demonstrated by Han et al.(3)

In contrast, our research does not directly draw upon these LSTM-based techniques. Instead, we build upon the framework established by Vasileva et al.(1), integrating it with modern embedding and simplifying the distance metric. Our approach replaces the original image and text embedders with CLIP and emphasizes minimizing Euclidean distances, diverging from the element-wise product and fully-connected layer methodology for generalized distance function. This strategic choice has led to significant performance improvements in our models, as evidenced by a 7.6% and 5% increase in FITB (Fill in the Blank) and AUC (Area Under Curve) metrics.

In summary, our work represents a confluence of established methodologies and innovative approaches in deep learning. By harnessing the power of OpenAI's CLIP and refining the concepts of type-aware embeddings and triplet loss structures, we are pushing forward the boundaries of fashion compatibility models, contributing a novel perspective to the AI-driven fashion landscape.

3 DATA

In our research, we utilized the Polyvore dataset originally curated by Vasileva et al. (1). This dataset, sourced from the Polyvore website, comprises a rich collection of fashion items and outfits, each annotated with detailed information such as images, text descriptions, and type information. The dataset includes a total of 68,306 outfits and 365,054 individual items, offering a diverse range of fashion elements for analysis.

For our study, we opted for the "easier" split of this dataset, as defined by the original authors. This split focuses on providing a balanced yet comprehensive test bed by allowing for item overlap between the training, testing, and validation sets, while ensuring that no complete outfit is repeated across these subsets. Specifically, this split includes 53,306 outfits for training, 10,000 for testing, and 5,000 for validation. The allowance for item overlap across sets, while maintaining outfit uniqueness, presents a realistic and practical scenario for evaluating our model's capability in fashion compatibility assessment, staying true to the spirit of the dataset's original design.

4 METHODS

4.1 EMBEDDING METHODOLOGY USING OPENAI'S CLIP

In our research, we leverage OpenAI's CLIP (Contrastive Language–Image Pretraining)(5) model to transform fashion items into high-dimensional embedding vectors. CLIP, a state-of-the-art model trained on a variety of images and text pairs, offers a robust embedding mechanism well-suited for our purposes.

- **Multimodal Embedding Space:** CLIP is designed to understand and relate visual and textual information, projecting them into a shared embedding space. This feature is particularly beneficial for fashion items, which often come with visual and descriptive textual data.
- **Contrastive Learning Approach:** At its core, CLIP employs a contrastive learning framework. It is trained to minimize the distance between the embeddings of matching image-text pairs while maximizing the distance between non-matching pairs. This aligns well with the objectives of a triplet loss model in fashion compatibility.(5)

The embedding function of CLIP is denoted as $f_{\text{CLIP}}(x_i; \theta) \in \mathbb{R}^d$, where x_i is an input item (image or text), and θ represents the learned parameters of the CLIP model. The output is a d-dimensional embedding vector. For an item x_i, CLIP produces an embedding $y_i = f_{\text{CLIP}}(x_i; \theta)$, applied to both images and text descriptions.

4.2 TYPE-SPECIFIC EMBEDDING SPACE

We define $M(u, v)$ as the type-specific embedding space for objects of types u and v. The projection matrices $P_{u\to(u,v)}$ and $P_{v\to(u,v)}$ map the embeddings of objects of types u and v into $M(u, v)$, respectively. For compatible items $x_i^{(u)}$ and $x_j^{(v)}$, the distance $\left\| P_{u\to(u,v)} \left(f_{\text{CLIP}} \left(x_i^{(u)}; \theta \right) \right) - P_{v\to(u,v)} \left(f_{\text{CLIP}} \left(x_j^{(v)}; \theta \right) \right) \right\|$ is minimized.

4.3 TRIPLET LOSS

We utilize the triplet loss function from Vasileva et al(1), where a triplet comprises an anchor item $x_i^{(u)}$ of category u, a positive item $x_j^{(v)}$ of category v compatible with the anchor, and a negative item $x_k^{(v)}$ of category v but incompatible with the anchor. The loss function ensures that the distance

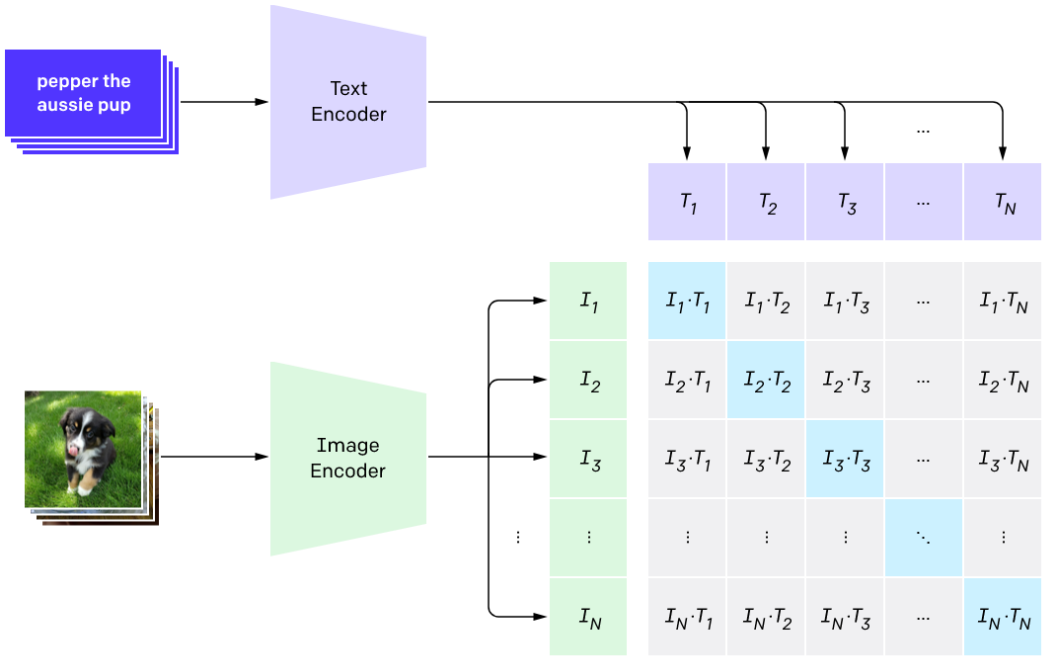

Figure 2: CLIP
(5)

between the anchor and the positive item is smaller than that between the anchor and the negative item. Compatibility is measured with

$$d_{uv}^{ij} = \left\| f_{\text{CLIP}}\left(x_i^{(u)};\theta\right) \odot w_{(u,v)} - f_{\text{CLIP}}\left(x_j^{(v)};\theta\right) \odot w_{(u,v)} \right\|^2 \tag{1}$$

where $\odot$ denotes component-wise multiplication, and $w_{(u,v)}$ are learned weights. The triplet loss is defined as

$$L_{\text{comp}} = \max\{0, d_{uv}^{ij} - d_{uv}^{ik} + \mu\}, \tag{2}$$

with μ as a margin.

4.4 Visual Semantic Embedding

We utilize CLIP's text embedding function $g_{\text{CLIP}}(t_i;\phi)$ for textual descriptions. The comprehensive loss used to learn similarity then can be defined as

$$L_{\text{sim}} = \lambda_1 l(x_j^{(v)}, x_k^{(v)}, x_i^{(u)}) + \lambda_2 l(t_j^{(v)}, t_k^{(v)}, t_i^{(u)}) \tag{3}$$

We also train a visual semantic embedding in the style of Han et al. which requires that the image $x_i^{(u)}$ is embedded closer to its description $t_i^{(u)}$ in visual-semantic space than the descriptions of the other two images in a triplet:

$$L_{\text{vse}} = l(x_i^{(u)}, t_i^{(u)}, t_j^{(v)}) + l(x_i^{(u)}, t_i^{(u)}, t_k^{(v)}) \tag{4}$$

4.5 Regularization and Final Training Loss

An L_1 penalty is imposed on the projection matrices to encourage sparsity, facilitating better disentanglement of embedding dimensions for compatibility. Additionally, L_2 regularization is applied to the learned image embedding.

$$L(X, T, P_{\cdot\to(\cdot,\cdot)}, \lambda, \theta, \phi) = L_{\text{comp}} + L_{\text{sim}} + \lambda_3 L_{\text{vse}} + \lambda_4 L_{L_2} + \lambda_5 L_{L_1} \tag{5}$$

In summary, our methodology integrates CLIP's powerful embedding capabilities with a type-aware embedding framework from Vasileva et al(1), adjusting its distance metric to be compatible with CLIP's contrastive learning framework. This allows for a nuanced and effective approach to learning fashion compatibility, leveraging both visual and textual information.

5 Experiment Setting

In our study, we adopt the evaluation methodology used by Han et al(3) to gauge the effectiveness of our model. This evaluation includes two key tasks:

Fashion Compatibility Assessment: Here, we rate how well the items in a proposed outfit complement each other. The performance metric for this task is the average area under the receiver operating characteristic curve (AUC). This measures the model's ability to correctly identify compatible and incompatible item combinations within outfits.

Fill-in-the-Blank (FITB) Fashion Recommendation: In this task, the model is presented with an outfit that is missing an item and several candidate items (four in our case). The objective is to identify the item that best completes the outfit. Performance is measured based on the model's accuracy in selecting the most compatible item for each outfit.

Dataset: All the experiments were conducted using the "easier" split of the Polyvore Outfits dataset (split containing non-disjoint outfits).

Implementation: We use Open AI's CLIP model to embed both the images and text with a general embedding size of 64 dimensions. In order to project the resulting features from the CLIP model to our chosen embedding space, we introduce a fully connected layer for our image embedder. For the text embedder, we employ adaptive average pooling to condense the embedding while preserving essential semantic information.

The model was trained with a learning rate of 5×10^{-5}, batch size of 32, and a margin of 0.3 for 5 epochs. We set $\lambda_3 = 5 \times 10^{-3}$ and 5×10^{-4} for λ_4 and λ_5 in equation (5). Adam was employed as the optimizer. For comparison, we also conducted the experiment under the same setting with the original approach from Vasileva et al. (1) and also with the image embedder replaced as ResNet50.

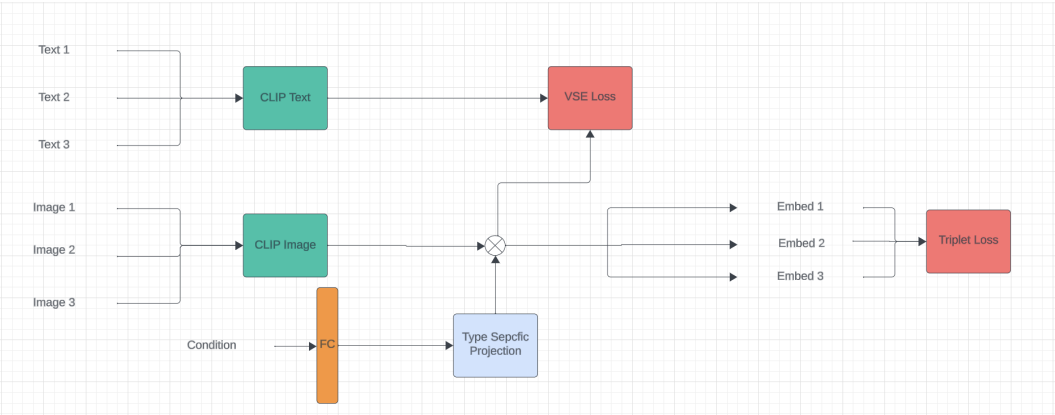

Figure 3: Model for Learning Type-Aware Embedding with CLIP

6 Result and Analysis

CSN, VSE + Sim + Metric (64-D): The original method by Vasileva et al.(1), employing ResNet18 for image embedding and a standard MLP for text embedding, demonstrates reasonable performance with a FITB accuracy of 55.3% and a Compatibility AUC of 0.86. This baseline establishes

Method	FITB Accuracy	Compatibility AUC
CSN, VSE + Sim + Metric (64-D)	55.3%	0.86
ResNet50, VSE + Sim + Metric (64-D)	59.2%	0.88
CLIP, VSE + Sim (64-D)	62.7%	0.91

Table 1: Evaluation results comparing different methods.

the effectiveness of the type-aware embedding and triplet loss structure in the context of fashion compatibility.

ResNet50, VSE + Sim + Metric (64-D): A variant of the original model, substituting ResNet18 with ResNet50 for image embedding, shows a noticeable improvement, achieving a FITB accuracy of 59.2% and a Compatibility AUC of 0.88. This increment suggests that the choice of image embedder can significantly influence the model's performance, with more advanced architectures like ResNet50 offering better visual understanding and representation.

CLIP, VSE + Sim (64-D): Our proposed method, integrating OpenAI's CLIP model for both image and text embedding, outperforms the other two approaches with a FITB accuracy of 62.7% and a Compatibility AUC of 0.91. This substantial improvement is indicative of the robustness and versatility of CLIP in handling multimodal data. By leveraging CLIP's capability to understand and relate visual and textual data in a shared high-dimensional space, our model achieves a more nuanced understanding of fashion compatibility. This is reflected in its superior performance in both outfit recommendation (FITB) and compatibility assessment (AUC).

The progression from a ResNet18 to ResNet50 framework exhibited a clear benefit, while the incorporation of OpenAI's CLIP model further extended this advantage, achieving notable gains in both FITB accuracy and Compatibility AUC. These outcomes theoretically align with the hypothesis that more sophisticated models, capable of deeper and more abstract representations, should yield superior performance in multimodal learning tasks. The success of the CLIP-based approach, even with lower-dimensional embeddings(64-D), potentially indicates its superior efficiency in capturing the essence of fashion items and the subtleties of their compatibility.

6.1 Limitations and Future Work

One of the primary challenges faced during our experimentation was the extensive size of the Polyvore Outfits dataset, which limited the number of experiments and epochs we could feasibly conduct. While our method demonstrated superior results even with 64-dimensional embeddings(the best performance by the authors was with 512-dimensional embeddings), there is evidence to suggest that larger embedding sizes could have lead to more gain in performance. Unfortunately, due to the limitations in computational resources, particularly memory capacity, we were constrained to a batch size of 32 as opposed to the original authors' batch size of 256. These constraints have inevitably impacted the breadth of our experimental analysis.

Future work could involve a more rigorous experimental framework that explores the impact of various factors, including batch size and embedding dimensions, on model performance. Investigating these aspects could provide deeper insights into the capabilities and optimization of the CLIP embedder for our triplet loss model, particularly in how different configurations affect the generalizability and robustness of the model.

Despite these limitations, our findings present a compelling case for the efficacy of CLIP in the domain of fashion compatibility, setting a foundation for further research in this area. The versatility and robustness demonstrated by our approach highlight the potential of advanced AI techniques in transforming the landscape of fashion recommendation systems and similar applications.

Author Contributions

As the sole author of this paper, I have undertaken all aspects of the work including the conception and design of the study, data collection and analysis, development of the models, and writing of the manuscript.

REFERENCES

[1] Mariya I Vasileva, Bryan A Plummer, Kedarnath Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. 2018. [Online]. Available: https://arxiv.org/pdf/1803.09196.pdf.

[2] McKinsey & Company. Generative ai: Unlocking the future of fashion. https://www.mckinsey.com/industries/retail/our-insights/generative-ai-unlocking-the-future-of-fashion, 2021. Accessed: [Insert Date Here].

[3] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. 2017.

[4] Takuma Nakamura and Ryosuke Goto. Outfit generation and style extraction via bidirectional lstm and autoencoder. October 2018. [Online]. Available: https://arxiv.org/abs/1810.xxxxx.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020.