# Unimodal vs. Multimodal Siamese Networks for Outfit Completion

Mariya Hendriksen
AIRLab, University of Amsterdam
m.hendriksen@uva.nl

Viggo Overes
University of Amsterdam
mail@viggooveres.xyz

## ABSTRACT

The popularity of online fashion shopping continues to grow. The ability to offer an effective recommendation to customers is becoming increasingly important. In this work, we focus on Fashion Outfits Challenge, part of SIGIR 2022 Workshop on eCommerce. The challenge is centered around *Fill in the Blank* (FITB) task that implies predicting the missing outfit, given an incomplete outfit and a list of candidates. In this paper, we focus on applying siamese networks on the task. More specifically, we explore how combining information from multiple modalities (textual and visual modality) impacts the model's performance on the task. We evaluate our model on the test split provided by the challenge organizers and the test split with gold assignments that we created during the development phase. We discover that using both visual, and visual and textual data demonstrates promising results on the task. We conclude by suggesting directions for further improvement of our method.

## KEYWORDS

Fashion Outfits Challenge, outfit completion, fill in the blank, siamese networks

## 1 INTRODUCTION

Fashion is becoming increasingly popular in modern e-commerce [2]. One of the common fashion recommendation tasks related to the problem is FITB. The task consists of predicting a missing item, given an incomplete outfit, and a list of candidates. Figure 1 illustrates the task.

**Fashion Outfits Challenge**. In this work, we focus on FITB task in the context of *Fashion Outfits Challenge* [1]. The dataset consists of approximately 400,000 products with product images and metadata. Besides, the dataset includes approximately 300,000 outfits created by stylists and fashion experts. The models are evaluated via an online leaderboard on a test set. The gold standard assignments of the test set are not public. The metric used for performance evaluation is accuracy. Additionally, we evaluate the model performance on *mean reciprocal rank* (MRR).

**Our solution**. The main contributions of this work are as follows: (1) We apply the siamese network on the FITB task and explore its effectiveness when using unimodal (Text or Image) and multimodal (Text & Image) product representations. We present a lightweight solution that uses only 697,280 trainable parameters. (2) We analyze the effectiveness and limitations of our method and discuss directions for future work. We share our code and experimental settings to facilitate reproducibility of our results [2].

## 2 RELATED WORK

**Outfit completion**. The majority of work on FITB task was done on Polyvore dataset [5]. The authors of the dataset proposed to use *bidirectional long short-term memory* (BiLSTM) network on the task. The model leverages visual data alongside one-hot encoded product descriptions and treats the task as a sequence prediction problem. Cucurull et al. [1] propose to use *graph neural network* (GNN) on the FITB task. In the work, they see each outfit as a graph and treat the outfit completion task as a missing link prediction problem. Revanur et al. [11] propose to learn fashion compatibility in a semi-supervised way by learning pseudo positive and negative outfits while training the model. Another approach implies learning type-aware use type embeddings Vasileva et al. [12] propose to jointly learn the notions of item similarity and compatibility while training the outfit completion model. Veit et al. [13] propose to learn the compatibility of items using a siamese CNN trained on dyactic co-occurrences. Unlike prior work in this domain, we investigate the performance of unimodal vs. multimodal siamese networks on the task of outfit completion.

**Multimodal fashion search**. Multimodal fashion retrieval is an important and actively developing topic [6]. Some of the related problems include fine-grained cross-modal retrieval [3], machine translation [8], and fashion recommendations [9]

Unlike prior work in this domain, we focus on leveraging multimodal fashion product data on the FITB task.

## 3 APPROACH

**Task definition**. We follow the same notation as in [7, 14]. We present the input dataset as product-product pairs $(\mathbf{x}_p^i, \mathbf{x}_p^j)$, where $\mathbf{x}_p^i$ and $\mathbf{x}_p^j$ represent information about two products. A product-product pair $(\mathbf{x}_p^i, \mathbf{x}_p^j)$ is positive if both products belong to the same outfit; the pair is negative if products in the pair do not belong to the same outfit. The product information includes images $\mathbf{x}_i$, text $\mathbf{x}_t$, and meta data, i.e., $\mathbf{x}_p = \{\mathbf{x}_i, \mathbf{x}_t, \mathbf{x}_m\}$.

For the FITB task, we take as an input a list of products in an incomplete outfit and a list of candidate products; we aim to select a product that completes the outfit from the list of candidates.

**CLIP-Siamese**. Figure 2 illustrates our approach. The model projects product information $\mathbf{x}_p$ into a $d$-dimensional space with the resulting vector $\mathbf{p}$. The model consists of an encoding and a siamese modules. It is trained with contrastive loss.

**Encoding product information**. We encode product textual and

[1] https://eval.ai/web/challenges/challenge-page/1721/overview, Last accessed: 20.07.2022.

[2] https://github.com/mariyahendriksen/OutfitComposition

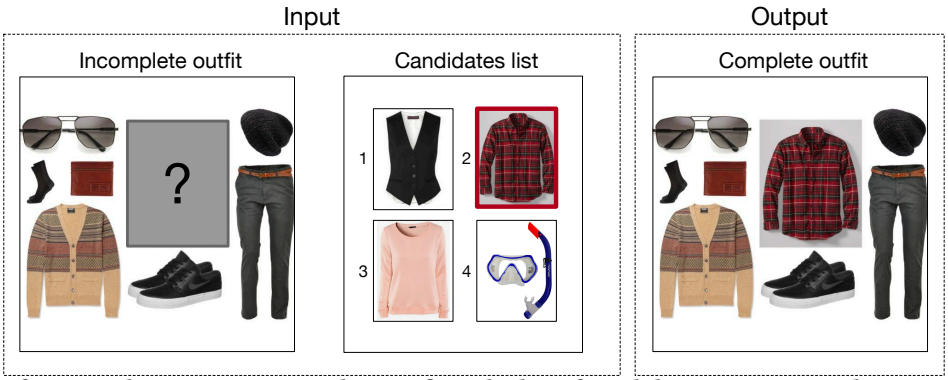


**Figure 1: Example of FITB task. Given an incomplete outfit and a list of candidates, we aim to select a candidate that would complete the outfit and return the complete outfit. In the example, the target item is the item #2 in the candidates list.**

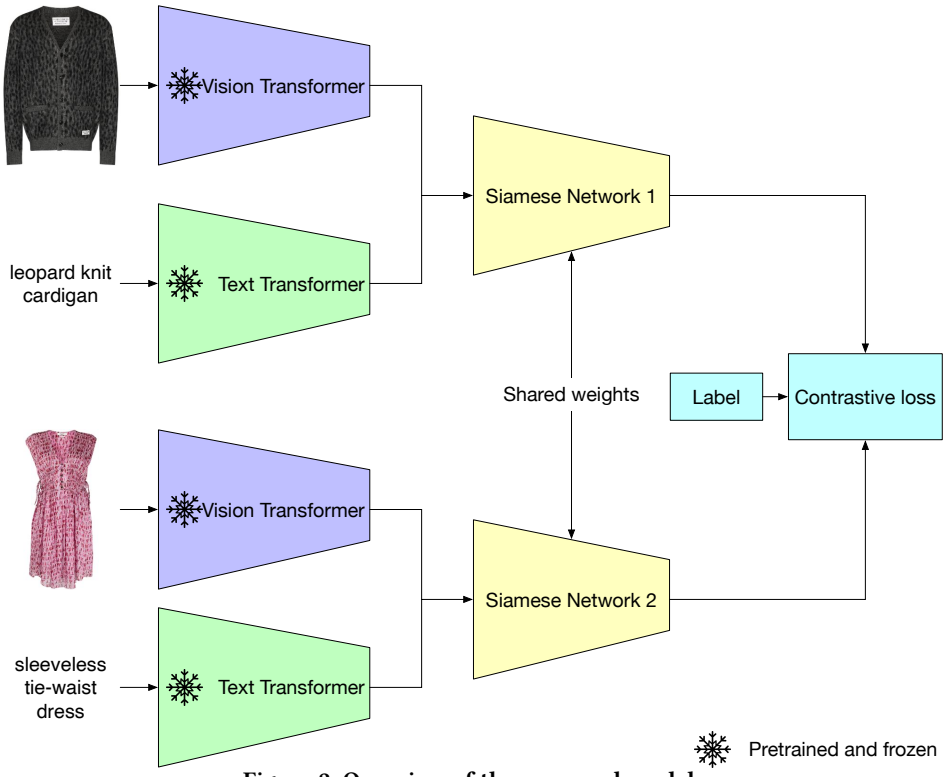


**Figure 2: Overview of the proposed model.**

visual information with text and image encoder. The *image encoder* ($f_i$) takes as input a product image $\mathbf{x}_i$. The image $\mathbf{x}_i$ is passed through the image encoder:

$$\mathbf{h}_i = f_i(\mathbf{x}_i). \tag{1}$$

The *text encoder* ($f_t$) takes a product textual information $\mathbf{x}_t$ as input and returns a text representation $\mathbf{h}_t$:

$$\mathbf{h}_t = f_t(\mathbf{x}_t). \tag{2}$$

To obtain the image and text representations, we use a pre-trained and frozen Vision Transformer and Text Transformer from CLIP model.

The image and text representations are passed to the *siamese network* ($s_p$). The network takes as an input a concatenation of the image representation $\mathbf{h}_i$ and text representation $\mathbf{h}_t$, and projects the resulting vector $\mathbf{h}_p = concat(\mathbf{h}_i, \mathbf{h}_t)$ into the into a $d$-dimensional space:

$$\mathbf{p} = g_p(\mathbf{h}_p) = g_p(concat(\mathbf{h}_i, \mathbf{h}_t)) \tag{3}$$

where $\mathbf{p} \in \mathbb{R}^d$.

**Loss function**. After obtaining product representation for pair of products $(\mathbf{p}_i, \mathbf{p}_j)$, we use contrastive loss [4] to train CLIP-Siamese. The loss goes over positive and negative product pairs. Label $Y$ indicates if the pair is positive ($Y = 1$) or negative ($Y = 0$).

$$\mathcal{L}(Y, \mathbf{p}_i, \mathbf{p}_j) = (Y)D^2 + (1 - Y)\{\max(0, m - D)\}^2 \tag{4}$$

where $D = D(\mathbf{p}_i, \mathbf{p}_j) = ||\mathbf{p}_i - \mathbf{p}_j||_2$ is the euclidean distance, $m > 0$ is a margin.

## 4 EXPERIMENTAL SETUP

**Metrics**. We evaluate the model's performance using accuracy and MRR.

**Baselines**. We use category-based baseline provided by the challenge organizers, and CLIP [10] as our baselines.

**Evaluation method**. To explore how model performance changes w.r.t. unimodal vs. multimodal product representation, we train and evaluate CLIP-Siamese on three types of product representations: (1) *Text*: we use only text data to build text-based product representations (2) *Image*: we use only product images to build image-based product representations (3) *Text & Image*: we use both textual and visual product data to build multimodal product representations

**Experiments**. We run two experiments. In *Experiment 1* we investigate how using unimodal and multimodal product representations impacts the accuracy of CLIP-Siamese when evaluated on the test split provided by the challenge organizers. We run the experiments on the test split provided by the challenge organizers, and use accuracy as the metric. We use CLIP [10] in zero-shot setting as our baseline.

In *Experiment 2* we further investigate CLIP-Siamese performance with three different types of product representations. We consider MRR scores obtained by running the model on our own test split. Similar to the previous experiment, we use CLIP [10] in a zero-shot setting as our baseline.

## 5 RESULTS

**Experiment 1: Fashion Outfits Challenge test split**. We investigate how using unimodal and multimodal product representations for training the model for the task impacts the accuracy of when we evaluate the model on the test split provided by the challenge organizers. We use CLIP in a zero-shot setting [10] as a baseline.

| Model | Accuracy | | |
|---|---|---|---|
| | Text | Image | Text & Image |
| CLIP zero-shot [10] | 0.041,60 | 0.041,46 | 0.042,46 |
| CLIP Siamese (Ours) | 0.045,93 | 0.048,64 | **0.049,20** |

**Table 1: Results of Experiment 1. Models accuracy scores when using three different types of product representations. The best performance is highlighted in bold.**

The results are shown in Table 1. In all cases, CLIP-Siamese outperforms CLIP zero-shot. The most significant relative gain is for image-based product representations where CLIP-Siamese outperforms CLIP zeros-shot by 17.32%. It is followed by 15.86% relative gain for text and image-based representations and 10.41% gain for text-based representations. Overall, CLIP-Siamese with text and image-based representations performs best.

**Experiment 2: Our own test split**. To improve our understanding of model performance, we consider the performance in terms of MRR scores. Since the the gold standard assignments for the test split is not released yet, we create our test split using scripts provided in *utils* folder available on the challenge page.

| Model | MRR | | |
|---|---|---|---|
| | Text | Image | Text & Image |
| CLIP zero-shot [10] | 0.163,48 | 0.161,40 | 0.162,77 |
| CLIP Siamese (Ours) | 0.167,90 | **0.184,87** | 0.181,55 |

**Table 2: Results of Experiment 2. Models MRR scores when using three different types of product representations. The best performance is highlighted in bold.**

Table 2, shows the experimental results for Experiment 2. Overall, CLIP-Siamese with image-based representations demonstrates the best performance, CLIP-Siamese with text and image-based representations is the second best.

## 6 CONCLUSIONS

In this paper, we present CLIP-Siamese, a model we created for Fashion Outfits Challenge. We evaluated the model on unimodal and multimodal product representations and showed that using both visual, and visual and textual data for building product representations demonstrates promising results. Future work includes further improvement of the model architecture and investigation of model performance on other datasets, e.g., Polyvore [5].

## REFERENCES


[1] Guillem Cucurull, Perouz Taslakian, and David Vazquez. 2019. Context-Aware Visual Compatibility Prediction. https://doi.org/10.48550/ARXIV.1902.03646

[2] Kinga Edwards. 2020. Key takeaways from E-commerce Region Report: Europe 2020. https://ecommercegermany.com/blog/key-takeaways-from-e-commerce-region-report-europe-2020. [Online; accessed 4-May-2022].

[3] Kenneth Goei, Mariya Hendriksen, Maarten de Rijke, et al. 2021. Tackling attribute fine-grainedness in cross-modal fashion search with multi-level features. In *SIGIR 2021 Workshop on eCommerce. ACM.*

[4] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

[5] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. 2017. Learning Fashion Compatibility with Bidirectional LSTMs. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM. https://doi.org/10.1145/3123266.3123394

[6] Mariya Hendriksen. 2022. Multimodal Retrieval in E-Commerce. In *European Conference on Information Retrieval*. Springer, 505–512.

[7] Mariya Hendriksen, Maurits Bleeker, Svitlana Vakulenko, Nanne van Noord, Ernst Kuiper, and Maarten de Rijke. 2022. Extending CLIP for Category-to-image Retrieval in E-commerce. In *European Conference on Information Retrieval*. Springer, 289–303.

[8] Katrien Laenen and Marie-Francine Moens. 2019. Multimodal neural machine translation of fashion e-commerce descriptions. In *International Conference on Fashion communication: between tradition and future digital developments*. Springer, 46–57.

[9] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2019. Improving outfit recommendation with co-supervision of fashion generation. In *The World Wide Web Conference*. 1095–1105.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[11] Ambareesh Revanur, Vijay Kumar, and Deepthi Sharma. 2021. Semi-Supervised Visual Representation Learning for Fashion Compatibility. In *Fifteenth ACM Conference on Recommender Systems*. ACM. https://doi.org/10.1145/3460231.3474233

[12] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. 2018. Learning Type-Aware Embeddings for Fashion Compatibility. https://doi.org/10.48550/ARXIV.1803.09196

[13] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. 2015. Learning Visual Clothing Style with Heterogeneous Dyadic Co-occurrences. https://doi.org/10.48550/ARXIV.1509.07473

[14] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747* (2020).