THEME-MATTERS: FASHION COMPATIBILITY LEARNING VIA THEME ATTENTION

Jui-Hsin(Larry) Lai[1]*, **Bo Wu**[3]*, **Xin Wang**[4], **Dan Zeng**[5], **Tao Mei**[2], **Jingen Liu**[2]
[1]PAII-Labs, [3]Columbia University, [2]JD.com, [4]Donghua University, [5]Shanghai University
juihsin.lai@gmail.com, bo.wu@columbia.edu, wangx@mail.dhu.edu.cn,
dzeng@shu.edu.cn, tmei@live.com, jingen.liu@gmail.com

October 6, 2020

ABSTRACT

Fashion compatibility learning is important to many fashion markets such as outfit composition and online fashion recommendation. Unlike previous work, we argue that fashion compatibility is not only a visual appearance compatible problem but also a theme-matters problem. An outfit, which consists of a set of fashion items (e.g., shirt, suit, shoes, etc.), is considered to be compatible for a "dating" event, yet maybe not for a "business" occasion. In this paper, we aim at solving the fashion compatibility problem given specific themes. To this end, we built the first real-world theme-aware fashion dataset comprising 14K around outfits labeled with 32 themes. In this dataset, there are more than 40K fashion items labeled with 152 fine-grained categories. We also propose an attention model learning fashion compatibility given a specific theme. It starts with a category-specific subspace learning, which projects compatible outfit items in certain categories to be close in the subspace. Thanks to strong connections between fashion themes and categories, we then build a theme-attention model over the category-specific embedding space. This model associates themes with the pairwise compatibility with attention, and thus compute the outfit-wise compatibility. To the best of our knowledge, this is the first attempt to estimate outfit compatibility conditional on a theme. We conduct extensive qualitative and quantitative experiments on our new dataset. Our method outperforms the state-of-the-art approaches.

1 Introduction

Fashion compatibility learning, whose goal is to automatically compose compatible outfits for recommendations, is of importance to a variety of academic and industrial tasks such as outfit composition [1], wardrobe creation [2], item recommendation [3], fashion generation [4]. It has recently attracted increasing attention [5, 6, 7, 8, 9, 10, 11, 12, 3, 13, 14].

In general, we can categorize the fashion compatibility learning methods into two classes: one formulates it as a pair-wise learning task [15] [16] [6] [10], which develops measurement methods (e.g., metric learning) for pair-wise compatibility, and another one is outfit-wise compatibility learning, which models the process of forming an outfit as sequence learning (i.e., LSTM) [5] [7] [8]. Most existing works treat fashion compatibility as visual appearance compatible problem. As a result, although significant progress has been made, we are still not able to answer a question as shown in Figure 1: "is Outfit A compatible in a business occasion?".

Fashion compatibility is also a theme-matters problem. For example, as shown in Figure 1, Outfit A may be compatible based on visual appearance and can be dressed for dating. But if one wants to have it for business, she may want to adjust it to Outfit B (long shirt instead of miniskirt for business). Therefore, theme-aware fashion compatibility is very important for fashion recommendation.

*This work is completed during the time in Silicon Valley Research Center, JD.com. Jui-Hsin(Larry) Lai and Bo Wu have equal contribution.

Figure 1: An example demonstrating the importance of theme-aware fashion compatibility. In the theme-ignored case, Outfit A may be compatible based on visual appearance. But, "is it suitable for the business occasion?" In general, miniskirt may not fit an office. So in the theme-aware case, we can generate Outfit B, which may be better for business.

Most existing fashion datasets such as Polyvore dataset [5] and DeepFashion2 [17], however, do not carry the capability to estimate theme-aware fashion compatibility. Hence, we built a new real-world fashion dataset called Fashion32, which is the first one with rich annotations including outfit themes and fine-grained fashion categories. Since the annotations were labeled by fashion stylists from brand vendors, they generally are of high quality. Fashion32 contains 32 theme tags for more than 13K around outfits, and 152 fine-grained categories for more than 40K outfit items. To learn theme-aware fashion compatibility models from this dataset, we face two challenges: how to measure pair-wise compatibility of outfit items and how to associate a theme to pairwise compatibility to compute outfit-wise compatibility.

To address the above challenges, we propose a theme-attention model, which is built on the category-specific embedding space. Figure 2 illustrates the overview of our framework. Given an outfit and a specific theme, pair-wise items are projected into the category-specific subspace (Figure 2 (a)). Unlike traditional embedding, which maps all fashion items into a common space, we employ triplet network and embedding masks (Figure 2 (b)) to project items category-specific subspace embedding. This "task-orientated" embedding enables the subspace to be more discriminative for compatibility computing. We further build a theme-attention model to associate the themes with pairwise compatibility (Figure 2 (c)). As a result, a theme-specific attention matrix is learned to link the theme to pairwise compatibility of outfit items, and further to aggregate pairwise ones to estimate the outfit-wise compatibility.

To the best of our knowledge, our work is the first one to explicitly estimate fashion compatibility given a specific theme. [5] maybe able to answer a question like "what to dress for a biz meeting" thanks to their visual-semantic embedding, but their capability relies on the quality of the image captions. Also, their Bi-LSTM framework is less flexible due to specific item order and number. Yet our theme-aware model does not have such constraints. The category-specific embedding is inspired by [16][10]. But unlike [16] and [10], which simply group fashion items into coarse categories (e.g., top, bottom, shoe, etc), we employ fine-grained categories because they usually have strong connections to fashion themes due to their properties. For instance, T-shirts imply causal, shirts are more official, and Polo-shirts are in-between. All the properties essentially imply some fashion themes. The coarse category does not carry this advantage.

To summarize, our work has the following contributions:

- We introduce the first theme-aware fashion dataset, which enables to compute the fashion compatibility given a specific theme. The dataset is available to download via the webpage www.larry-lai.com/fashion.html
- We propose a theme-attention model to associate themes with pairwise compatibility to compute outfit-wise compatibility. To the best of our knowledge, this is the first attempt to study the theme-matters compatibility learning problem.
- We leverage fine-grained categories and the category-specific embedding to effectively support our theme-attention model.
- We demonstrate our proposed approach can outperform state-of-the-art approaches on Fashion32 dataset and the improved Polyvore dataset [5].

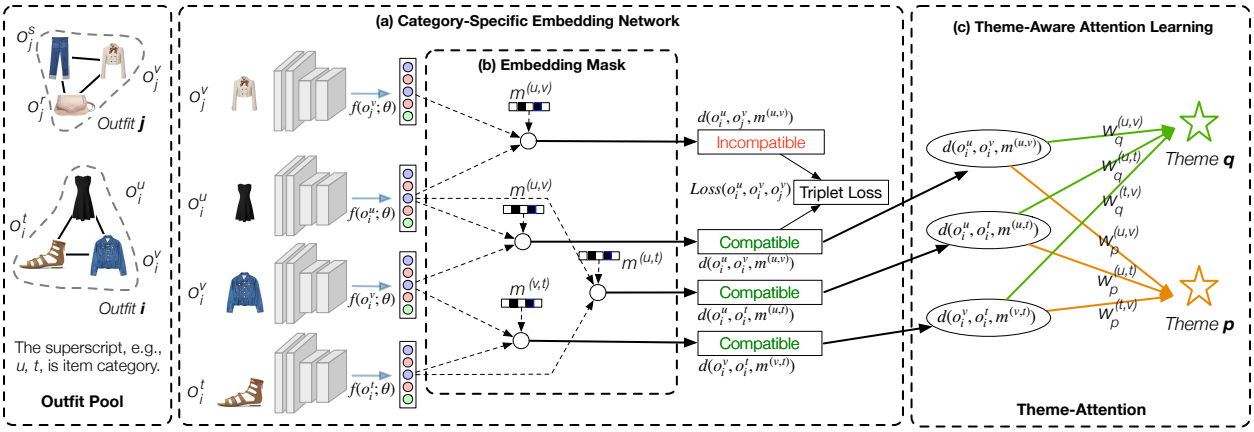

Figure 2: The framework of the proposed theme-attention model for fashion compatibility. This model is built on the fine-grained category, which serves as a bridge connecting outfit/fashion items to the theme. (a) Pairwise outfit items are projected into the category-specific embedding space, and then (c) theme-aware attention is employed to associate themes with the pairwise fashion compatibility. The outfit-wise compatibility is computed as the aggregation of pairwise compatibility using theme-aware attention.

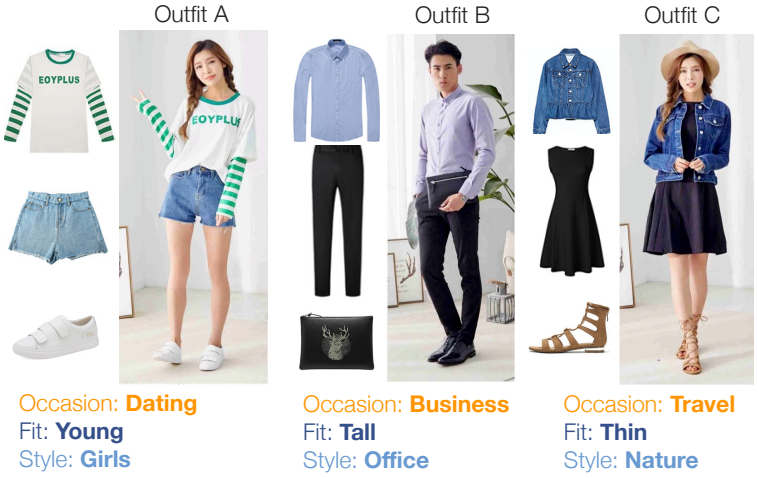

Figure 3: Sample outfits from the Fashion32 dataset. Each outfit carries theme tags, descriptions, items, and pictures of a model to illustrate its compatibility.

2 Related Work

Fashion Datasets. In general, we can group the existing fashion datasets, which are built for fashion compatibility learning, into two categories: online shopping datasets [18, 19, 2] and social media datasets [5, 10, 8]. The former datasets are mined from some online e-commerce platforms by leveraging buyer's shopping carts to form various outfits and labels. As we know, however, a shopping cart usually contains mixed items, which may not form a compatible outfit. Therefore, the labels in online shopping datasets can be very noisy. The social media datasets, such as Maryland Polyvore [5] and Polyvore Outfits [10], are collected from social media platforms. Since the outfits are created by fashion enthusiasts, these datasets have less outfit noise. But the quality of outfits is very diverse, because outfits created on social media largely depends on users' uploading and labels are also random and bias.

Unlike previous datasets, our Fashion32 not only carries fashion themes, fine-grained categories, and recommendation descriptions but also has high-quality outfit model pictures and abundant annotations. More importantly, our outfit composition and its annotations come from fashion designers of brand vendors. Consequently, our dataset is more realistic and convincing.

Another two popular fashion datasets, i.e., DeepFashion [20] and DeepFashion2 [17], are specifically built for fashion research including attribute prediction, image retrieval, and fashion synthesis, but not for fashion compatibility.

Theme Type	Theme Tag(outfit counts)
Occasion	Dating(4674), Travel(1706), Party(1206), Sports(578), School(1179), Business(2769), Home(447), Wedding(35)
Style	Sports(491), Casual(2485), Office(1116), Japanese(255), US(2733), UK(1230), Girls(1994), Ladies(373), Simple(1029), Nature(1765), Punk(224), Folk(37)
Fit	Bottom(8), Small Face(17), Long Neck(30), White Skin(215), Thin(6184), Tall(971), Breast(30), Young(1057), Strong(6)
Gender	Female(9502), Male(4412), Unisex(375)

Table 1: Number of outfits under the theme tags in Fashion32 dataset. 32 theme tags are grouped into 4 types including occasion, style, fit, and gender.

Fashion Compatibility Learning. In general, telling an outfit compatible or not is a subjective task. One needs to check all possible compatible relationships between items, which can involve very subtle difference. One current solution is to leverage metric learning [15, 10, 21] or embedding techniques [22, 13, 23] to project the fashion items into a specific space, in which the outfit compatibility is explicitly measured by pairwise items' distances [16]. This measurement is built on pairwise compatibility rather than outfit-wise (namely, computing outfit compatibility as a whole). Han et al [5] employ Bi-LSTM beyond visual-semantic embedding to estimate outfit compatibility in an end-to-end model. Meanwhile,Vasileva et al [10] proposes a category-aware embedding approach to include garment/item types or coarse-grained categories (e.g., top, bottom, etc.) during learning. Taking the garment type into consideration, the embedding space consists of a set of type-specific sub-spaces, which further improves the fashion compatibility estimation.

3 The Fashion32 Dataset

As aforementioned, most current fashion datasets lack the capability of theme-aware fashion compatibility learning because no fashion theme and labels of fine-grained item category are provided. Accordingly, we collected a new fashion dataset called Fashion32. There are about 13K outfits, and each of them has been labeled with multiple themes from a set of 32 themes. Also, each outfit fashion item is tagged with one of the 152 fine-grained categories. Figure 3 shows some outfit examples from our dataset. Every single outfit has rich meta information and various labels, as well as real model pictures. To the best of our knowledge, this is the first real-world dataset carrying both theme and fine-grained category annotations for each outfit and fashion item. The annotations were labeled by fashion stylists from the brand vendors. It can be publicly accessed through the following link: http://www.larry-lai.com/fashion.html

Dataset collection. The Fashion32 dataset is crawled from the fashion channels of the e-commerce platform JD.com, one of the largest e-commerce platforms for fashion shopping. We collected 32 fashion themes as listed in Table 1. These fashion themes were proposed by fashion designers and utilized by the platform to index its products. Each collected outfit in Fashion32 is designed by fashion designers and uploaded to the platform by the brand vendors. The brand name of designers is recorded in each outfit for reference.

We collected 13,914 outfits, as well as additional 40,667 images of fashion items and 51,415 model pictures. One outfit usually contains 2 or 3 fashion items and more than 4 pictures of a model wearing these fashion items, as shown in Figure 3. The model pictures are important in fashion compatibility because they demonstrate how to select fashion items to form an outfit. Each fashion item was assigned with one label for coarse-grained category (i.e., 6 categories including inner top, outer top, bottom, shoe, bag, and accessory) and one label for fine-grained category (i.e., 152 categories). This assignment was done when the vendors uploaded their outfits to the platform.

Besides, all fashion items carry more information including product name, Stock Keeping Unit (SKU) ID, tags of design/style, tags of texture/fabric, tags of color, and a paragraph for product descriptions. To our knowledge, this dataset has the most detailed fashion labels, which can be used not only in fashion compatibility but also in the fashion image analysis.

Theme tags and descriptions. The fashion theme of an outfit can carry rich context information on it. This high-level fashion knowledge can reflect an outfit's style, occasion, or culture. Hence, in this dataset, we mainly have the following four groups of fashion themes: occasion, style, fit, and gender. There are 32 themes in total as shown in Table 1,

which also lists the number of outfits collected for each theme. Each outfit is labeled with at least one theme. Also, for each outfit, we also collected a paragraph description, which explains the reason for fashion compatibility and recommendation. As an example, the Outfit C in Figure 3 is labeled with theme tags: *travel, thin, nature, and female*, and its description for the recommendation looks like *"Denim jackets open or bare shoulders, with straw hat sunglasses, full of holiday atmosphere"*.

Fine-grained category. As aforementioned, the Theme-Attention is built on the fine-grained categories, as they usually carry more high-level knowledge that can be used to form a theme. Each fashion items is labeled with one of 152 fine-grained categories such as T-shirt, jacket, boots, wallet, sunglasses, and so on. The fashion designers labeled the fashion attributes for each fashion item to construct the fine-grained categories. For example, the jacket of Outfit C in Figure 3 has the fine-grained category *short jacket*, and its attributes are *lapel, ruffle, simple, cowboy, long sleeve, tops, female, and jacket*. Each fashion item has 7 attributes in average.

4 Approach

In this section, we first formulate the theme-aware compatibility learning problem. Then, we propose our framework as shown in Figure 2, which consists of two main parts: (a) Category-Specific Embedding Network, and (b) Theme-Aware Attention Learning.

4.1 Problem Formulation

The proposed fashion compatibility learning framework consists of two major components: category-aware triplet embedding and attentive Theme-Attention, where category serves as a bridge connecting two components. Given a fashion outfit O, let o^u be one of the fashion items of O, where the superscript denotes this item's category is $u \in C$ (C is the fine-grained category set). Then, the theme-aware fashion compatibility of outfit O, given the fashion theme P, can be computed as,

$$y_O^P = \sum d(o^u, o^v|P) \quad \forall o^u, o^v \in O, u \neq v \tag{1}$$

where u and v are item categories, $d(o^u, o^v|P)$ is the pairwise distance between any pair of items in outfit O, and d is computed in a category-specific embedding subspace with theme-attention. The employment of theme-attention enables our approach to obtain the outfit-wise compatibility conditional on a theme P, rather than simply averaging the pairwise compatibility of an outfit.

4.2 Category-Specific Embedding Network

Given one pair of compatible items and one incompatible pair, a Triplet Network can learn a mapping function, which projects compatible items close to each other and incompatible ones separable in the embedding space. In general, the fine details of fashion items are important for the embedding to learn a fashion compatibility "metric". To better capture the details, we prefer to learn a category-specific embedding, since the compatibility measurement between one pair of categories can be different from that of another pair. In other words, a mapping function $f^{(u,v)}$ is learned to measure the compatibility between items from categories u and v. For simplicity, in this section, the superscript (u, v) will be omitted.

Given an outfit O_i and two items o_i^u, o_i^v in O_i, where u, v indicates the corresponding item's category, to compute the distance $d(o_i^u, o_i^v)$ in terms of compatibility, we adopt multiple layers CNN with deep residual block [24] (see Section 5.1 for details) to embed two items into the (u, v) category-specific space as $f(o_i^u; \theta)$ and $f(o_i^v; \theta)$, where θ represents the CNN parameters. As a result, the distance between two outfit items can be computed as,

$$d(o_i^u, o_i^v) = ||f(o_i^u; \theta) - f(o_i^v; \theta)||_2^2 \tag{2}$$

where $d(o_i^u, o_i^v)$ is the Euclidean Distance [25]. In the experiments (Section 5.1), we will introduce the details of different deep networks to construct non-linear projections.

To learn a category-specific (i.e., (u, v) category-pair) mapping function f, we form a training set $\mathcal{T}$ consisting of a group of training triplets $\{o_i^u, o_i^v, o_j^v\}$ by selecting two items o_i^u and o_i^v from outfit O_i and the third item o_j^v from another outfit O_j. The selected items are from either category u or v. One assumption in our triplet formulation is that items o_i^u and o_i^v from the same outfit are compatible, while o_j^v from another outfit is incompatible to the other two items. If o_i^u is the anchor of a specific outfit, the optimization goal is to force the distance between items in the same outfit $d(o_i^u, o_i^v)$ closer than that of items (o_i^u, o_j^v) from different outfits. Therefore, our goal during the triplet network learning is to minimize the following loss function over the training set $\mathcal{T}$:

$$\mathcal{L}oss(o_i^u, o_i^v, o_j^v) = \sum_u \sum_k max\{0, d(o_i^u, o_i^v) - d(o_i^u, o_j^v) + \mu\} \tag{3}$$

where μ is some margin.

4.3 Embedding Mask

The category-specific embedding network attempts to learn an independent mapping function $f^{(u,v)}$ for any pair of categories (u, v). It results in $|C| \times (|C| - 1)$ (C is the category set) number of CNN networks or embedding spaces. These individual embedding processes are less efficient because the CNNs could be highly redundant since the difference between two category-specific embeddings may be small. Therefore, instead of learning individual spaces, we propose to learn category-specific sub-spaces, which enables to learn a shared mapping function for all category-specific embedding. To this end, we further introduce a category-specific mask $m^{(u,v)}$ into the triplet embedding process. The mask serves as a gate function by selecting relevant bins to project an item to its category-specific subspace, which is depicted as $f(;\theta) \odot m^{(u,v)}$.

Then, the distance between two items in Equation (2) can be represented with category-specific compatibility:

$$d(o_i^u, o_i^v, m^{(u,v)}) = ||f(o_i^u;\theta) \odot m^{(u,v)} - f(o_i^v;\theta) \odot m^{(u,v)}||_2^2 \tag{4}$$

where $m^{(u,v)}$ is a $1 \times n$ vector, and n is also the output size of feature extractor $f(;\theta)$.

Therefore, the modified conditional triplet loss is represented as:

$$\begin{aligned} &\mathcal{L}oss(o_i^u, o_i^v, o_j^v, m^{(u,v)};\theta) = \\ &\sum_u \sum_k max\{0, d(o_i^u, o_i^v, m^{(u,v)}) - d(o_i^u, o_j^v, m^{(u,v)}) + \mu\} \end{aligned} \tag{5}$$

The loss can be minimized by learning the embedding mask $m^{(u,v)}$ to each category pair.

4.4 Theme-Aware Attention Learning

Given an outfit O, we can compute its outfit-wise compatibility by evaluating the pairwise compatibility of all pairs items (o^u, o^v) in outfit O. One straightforward solution is to average all pairwise compatibility as following,

$$y = \sum d(o^u, o^v, m^{(u,v)})/k \tag{6}$$

where o^u and o^v are items in outfit O, and k is total number of item pairs in O. This solution treats each pair equally without considering the outfit's theme tags. As shown in Figure 1, outfit A may be highly compatible without taking into account the fashion theme, while it may not be suitable for a business purpose. Therefore, we shall take into account fashion themes when measuring outfit-wise compatibility. To this end, an attentive Theme-Attention is proposed in this work.

In fact, the theme-aware fashion compatibility is an attention problem. The theme-attention is built to link themes to pairwise compatibility and eventually enables to add theme attention to the estimation of outfit-wise compatibility. As shown in Figure 2 (c), the pairwise compatibility (e.g., node $d(o^u, o^v)$ and $d(o^u, o^t)$) is computed based on category-specific embedding. The yellow edges imply the likelihood of associating a theme (e.g., p and q) to pairwise categories. In fact, the association likelihood is the theme-attention (e.g., $w_P^{(u,v)}$) to pairwise compatibility when aggregating all pairwise ones into the outfit-wise compatibility. Consequently, learning theme-attention is to learn the theme-attention values like $w_P^{(u,v)}$.

Therefore, given a fashion theme P, the theme-aware compatibility for an outfit O can be computed by,

$$y^P = \sum w_P^{(u,v)} \cdot d(o^u, o^v, m^{(u,v)}) \tag{7}$$

where $w_P^{(u,v)}$ indicates the attention weight for the category pair u and v. Putting all w together, we obtain an attention matrix W_P for a given theme P. As we can see, $d(o^u, o^v, m^{(u,v)})$ directly measures the compatibility based on the items' appearance. Basically, it can be treated as instance-level compatibility conditional on their category. While attention matrix W carries high-level human knowledge when measuring if an outfit compatible or not.

To learn the attention network W_P for theme P, we treat all compatible outfits associated with theme P as positive examples. To obtain the negative examples, we select outfits from other themes, as well as creating an incompatible outfit by replacing one or several items in a compatible one. We treat the compatibility prediction as a classification problem, and formulate the loss function as a Cross-Entropy Loss [26] as,

$$\mathcal{L}_i^P = y_i^P \cdot \log x_i^P + (1 - y_i^P) \cdot \log(1 - x_i^P) \tag{8}$$

where y_i^P is the output, and the x_i^P is ground-truth of theme-aware compatibility under theme P condition.

Methods	Type	Compat. AUC(%)	FITB Acc(%)
Our baseline	-	87.37	71.43
Theme-Attention	Occasion	**94.26**	76.87
	Fit	93.89	**78.85**
	Style	93.84	76.69
	Gender	87.21	74.53

Table 2: The compatibility learning performance comparison of Theme-Attention (Baseline) and Theme-Attention between different theme types.

5 Experiments

To evaluate the effectiveness of theme tags, we conduct experiments to compare proposed theme-aware approach and the state-of-the-art methods as the baselines. To show the generality of our method, we perform experiments not only on proposed dataset Fashion32 but also on theme-ignored dataset Polyvore[5], which has been used by several previous works.

5.1 Implementation Details

Training In all experiments, we use ImageNet [27] pre-trained ResNet-50 model [24] with the bottleneck network as the backbone model. The input image is resized to 224×224 and the output of embedding is a 1000-dimension feature vector. All models are trained on a single Tesla V100 GPU, and each input mini-batch has 32 outfits. It takes about 6 hours for 50 epochs of training. The learning rate is $1e^{-2}$ and exponentially decayed by a factor of 0.2 every 10 epochs. The optimizing strategy is SGD with a momentum of 0.9. Only model parameters with the best performance on the validation set will be saved. In Fashion32, We split the outfits into three parts: training (11,040), validation (853), and testing (2,021) sets. The experiment settings for the Polyvore dataset are the same as that for Fashion32. The Polyvore dataset is also available to download via the link *http://www.larry-lai.com/fashion.html*.

Negative Sample All labeled outfits in a dataset are naturally positive samples (i.e., compatible outfits). So there are no annotated negative outfits (i.e., incompatible outfits), because in real life people won't compose an incompatible outfit. To generate a negative sample, we select a positive one and substitute one item in this outfit with an item that is randomly selected from the other outfits, which carry the same category with the one in the original outfit. For training, validation and test set, the ratio of compatible and incompatible outfits are all $1:1$. During evaluating, each model is evaluated 5 times with different negative samples.

5.2 Compatibility Metrics

Area Under Curve (AUC) To evaluate the model's binary prediction of compatibility. We reported the Area Under the Receiver Operating Characteristic (ROC) curve.

Fill-in-the-blank (FITB) Accuracy FITB task aims at filling the most compatible item into the blank of an outfit. Each blank has 4 options in our experiment, the accuracy can be calculated for the option selection process. Our model chooses the answer by predicting scores for 4 possible outfits only substituting the blank item with different options.

5.3 Quantitative Experiments on Fashion32 Dataset

To verify the effectiveness of our theme-aware fashion compatibility model, we also implemented a **Baseline** version of the **Theme-Attention** method. Both of them minimize the compatibility loss via Equation (8) and Equation (5), respectively. The Theme-Attention is built on the Baseline with additional theme-aware attention.

Table 2 shows the AUC and FITB scores of both methods. In terms of AUC scores, theme-attention method achieved better performance than our Baseline in almost all groups of themes except the gender group. In general, FITB scores are proportional to AUC scores. Especially, theme-Attention achieves 6.89% of AUC increase for the occasion theme group. The results successfully demonstrate that theme-attention method is able to improve the quality of fashion compatibility. In addition, from the results, we can observe that theme-attention can perform better on some distinctive themes. For example, outfits of "sports" usually consist of T-shirt, shorts, or running shoes, which make the We theme-aware weights easy to be learned. As a comparison, the Gender theme group does not help fashion compatibility estimation. We conjecture that "Female" and "Male" are the distinctive themes because it is easy to tell if an outfit

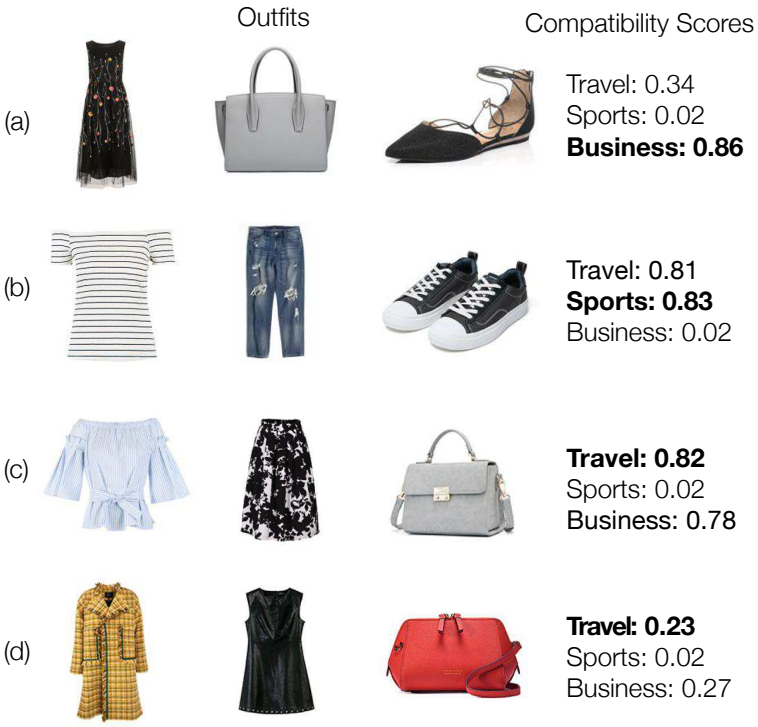

Figure 4: Some visual results of our theme-aware fashion compatibility learning. Three compatibility scores are computed given three specific themes.

	Fine-grained Polyvore		Fashion32	
Method	Compat. AUC(%)	FITB Acc(%)	Compat. AUC(%)	FITB Acc(%)
BiLSTM [5]	74.44 ± 0.95	45.41 ± 0.40	85.12 ± 0.26	60.14 ± 0.97
BiLSTM+VSE [5]	74.82 ± 0.63	46.02 ± 0.62	-	-
Concatenation [8]	83.40 ± 0.48	52.91 ± 0.59	93.82 ± 0.34	69.23 ± 1.42
Pooling [7]	**88.35** ± 0.26	57.28 ± 0.313	92.97 ± 0.43	69.28 ± 0.70
Type-Aware [10]	84.90 ± 0.52	57.06 ± 1.70	88.42 ± 0.47	71.15 ± 1.90
Our Baseline	85.85 ± 0.37	**60.66** ± 0.84	87.37 ± 0.52	71.43 ± 0.36
Theme Attention	-	-	**94.26** ± 0.62	**76.87** ± 0.87

Table 3: The comparison between different methods on Polyvore and Fashion32 dataset.

designed for male or female based on its visual appearance. Adding this incapable attention to the model actually hurt the performance. That is why its performance is worse than our Baseline.

In terms of the FITB scores, the Theme-Attention method can achieve up to 7.42% improvement compared to the Baseline. Overall, the Theme-Attention method can effectively learn the theme-aware fashion compatibility under a theme with narrow variations of category combinations.

Figure 7 illustrates some visual examples of our theme-aware fashion compatibility learning results. As we can see, our model generates different compatibility scores for an outfit given different themes. The theme with the highest score implies it is the most relevant theme to this outfit. Hence, outfit (a), (b), and (c) are correctly detected as business, sports, and travel, respectively. Example (d) is a negative example without a theme. Also, the result of (b) tells us an outfit can be suitable for multiple themes, i.e., both travel and sports are good for (b). This visualization further verifies the proposed theme-aware fashion compatibility.

5.4 Subjective Experiment on Online Fashion Shop

We further evaluate our model's performance by calculating the click rates of customers' browsing history, which is widely used by many e-commence platforms. An outfit with a higher click rate will have higher compatibility score. Unlike previous experiments, this one is more subjective.

Figure 5: Given an item and a theme, the Theme-Attention method searches on the pool of online shop to recommend the compatible outfits.

We collected 500 fashion items from an online fashion shop as the searching pool. Given a fashion item and a target theme, two experimental algorithms (i.e., our theme-attention and non-theme method) will recommend 3 to 5 items from the search pool to generate a complete outfit. Figure 5 shows some recommended outfits. As we see, the results are not only visually compatible but also suitable for the given theme. During the evaluation, 5 subjects are assigned to conduct the click rate experiments. The subjects were asked to tell which is more compatible given a fashion theme. To avoid subjects' bias on one recommendation algorithm, we randomly switched the order of two recommendations. Each subject evaluated 200 outfits on three themes: business, travel, and sports.

The final results show that Theme-Attention method is better than non-theme method with 8.6% improvement in terms of click rates. It further demonstrates that our approach can recommend theme-compatible outfits from a pool of new fashion items which have not to be seen during the training.

5.5 Performance Comparison

Since we are the first to work on the theme-aware fashion compatibility problem, it is very difficult to directly compare our approach with previous work. Since our Baseline is non-theme version of our approach, we can apply our Baseline to Polyvore dataset for comparison. However, our Baseline requires fine-grained categories which Polyvore does not have. To this end, we improved Polyvore dataset to fine-grained Polyvore (please refer to our supplemental material for this improved version). In addition, we also managed to run previous approaches on our Fashion32 dataset without leveraging the theme information.

Table 3 illustrates the detailed performance comparison on both fine-grained Polyvore and Fashion32 dataset. On fine-grained Polyvore dataset, as we can see, our Baseline is comparable to previous approaches for both compatibility prediction and FITB tasks. As our Baseline is actually a non-theme version of our approach, this comparison is somewhat meaningful. On Fashion32 dataset, our theme-attention outperforms the state-of-the-arts approaches. In terms of FITB, our approach is 7% better than state-of-the-art approaches. Although this is not a direct comparison, the results still demonstrate the advantages of theme-aware fashion compatibility. Previous approaches do not have a mechanism to leverage theme information for fashion compatibility.

6 Conclusions

To solve the theme-aware fashion compatibility problem, in this paper, we collected the first theme-matters fashion dataset, which contains 13K outfits in total over 32 themes and 152 fine-grained category classes. We further propose a novel benchmark, which leverages the attentive Theme-Attention built on category-specific embedding, to learn theme-aware fashion compatibility. we evaluate our approaches by both objective and subjective experiments. Compared with the baseline, our method Theme-Attention achieved 94.26% AUC in outfit compatibility prediction and 78.85% accuracy in FITB task, respectively. Comparing to several recent works on the Polyvore dataset, the Baseline version of the Theme-Attention method also achieves competitive on compatibility prediction and FITB tasks.

7 Supplementary Materials

7.1 A. Detailed Comparison between theme attention learning and non-theme method

Table.4 to Table. 7 provide detailed results for each theme tag group. Not only the overall performance but also the results for each kind of theme tags (e.g. sports, dating, etc.) are presented. It can be seen that theme attention method show superiority in almost every theme tags.

7.2 B. The statistics of Fine-grained Polyvore

We cleaned up Polyvore dataset with fine-grained categories. Polyvore dataset does not have theme tags, but has 149 fine-grained categories to satisfy the baseline version of the Theme-Attention method. Firstly, we filtered the 692 non-wearable fashion items in the outfits (e.g. paintings or cups). For multiple items of the same category in an outfit (e.g. multiple shoes in one outfit), only the first one will be taken. Finally, outfits with less than 3 fashion items are removed. The cleaned dataset will be publically available at www.larry-lai.com/fashion.html, its statistics are shown in Table. 8.

7.3 C. The Panel Designed for Click Rate Experiment

In the click rate experiment, the subjects were asked to click on the outfit, which is more compatible given a fashion theme, given two outfits as shown in Figure. 6. To avoid subjects to have any bias on one side, we randomly switched the order of two recommendations.

Figure 6: The panel for subjects to click on which outfit is more compatible with considering the sports theme.

7.4 D. Visual results for outfit compatibility with respect to different themes

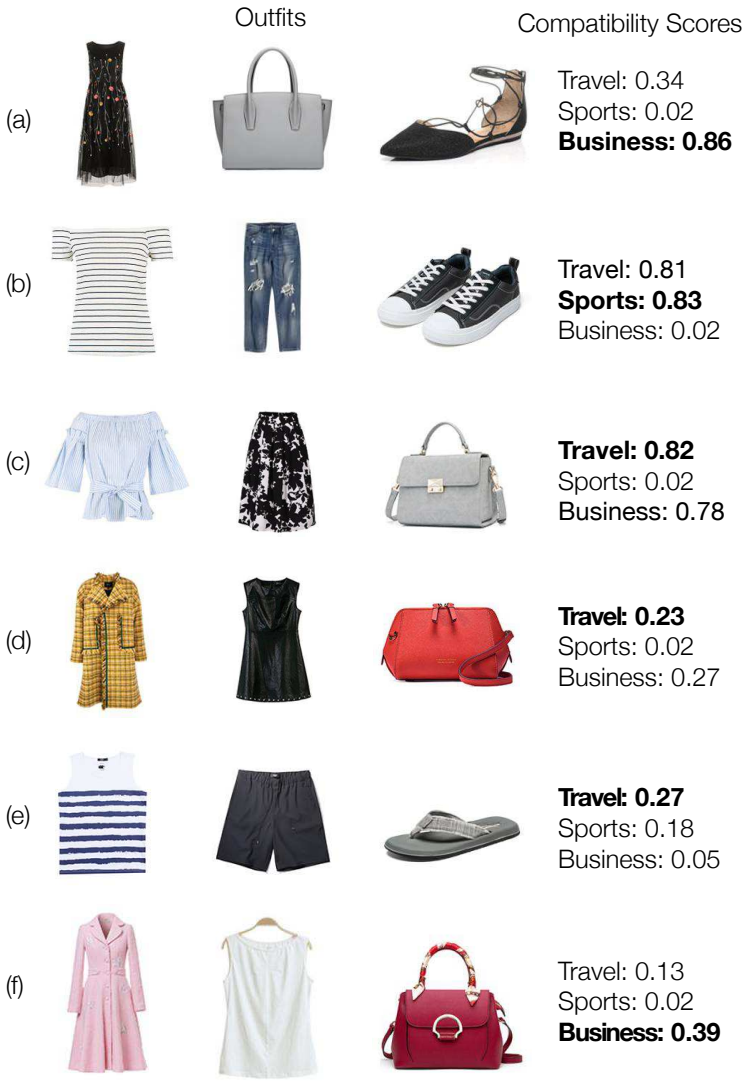

Figure 7: Some visual results of our theme-aware fashion compatibility learning. Three compatibility scores are computed given three specific themes.

Fashion	**Outfits**			**Type-Aware**		**Theme-Graph**	
Themes	Train	Validation	Test	Compat. AUC(%)	FITB Acc(%)	Compat. AUC(%)	FITB Acc(%)
Dating	3744	269	661	88.05 ± 0.79	72.62 ± 1.44	**93.29** ± 0.57	**74.36** ± 0.67
Travel	1351	120	235	90.69 ± 1.11	76.75 ± 1.74	**94.82** ± 0.75	**79.39** ± 2.16
Party	940	88	178	89.28 ± 3.34	73.64 ± 1.45	**92.54** ± 1.29	**74.78** ± 3.63
Sports	470	30	78	78.28 ± 3.01	68.60 ± 6.13	**98.61** ± 0.52	**80.04** ± 5.42
School	938	86	155	88.13 ± 2.25	74.42 ± 3.57	**96.92** ± 1.15	**82.52** ± 1.03
Business	2197	167	405	90.74 ± 1.02	72.07 ± 1.01	**94.75** ± 0.95	**78.18** ± 2.00
Home	355	37	55	85.86 ± 4.73	**78.63** ± 8.60	**91.60** ± 4.76	73.16 ± 4.36
Total	9995	797	1767	88.41 ± 0.47	71.15 ± 1.90	**94.26** ± 0.62	**76.87** ± 0.87

Table 4: The performance comparison over occasion theme group.

Fashion Themes	**Outfits**			**Type-Aware**		**Theme-Graph**	
	Train	Validation	Test	Compat. AUC(%)	FITB Acc(%)	Compat. AUC(%)	FITB Acc(%)
Bottom	7	0	1	-	-	-	-
Small Face	10	1	6	74.17 ± 17.56	49.67 ± 26.47	**82.50** ± 18.71	**56.67** ± 34.90
Long Neck	27	0	3	40.00 ± 20.00	0.00 ± 0.00	**80.00** ± 24.49	**86.67** ± 16.33
White Skin	177	9	29	44.30 ± 7.42	63.65 ± 11.99	**84.59** ± 5.37	**73.72** ± 3.60
Thin	4938	401	845	90.75 ± 1.07	76.24 ± 1.18	**94.63** ± 0.56	**79.70** ± 0.96
Tall	771	58	142	90.54 ± 2.14	70.38 ± 3.55	**93.94** ± 0.88	**75.65** ± 3.03
Breast	26	3	1	-	-	-	-
Young	846	72	139	86.62 ± 3.83	76.76 ± 2.28	**93.56** ± 1.61	**78.95** ± 1.33
Strong	4	0	2	-	-	-	-
Total	6806	544	1168	89.67 ± 0.90	74.17 ± 1.26	**93.89** ± 0.55	**78.85** ± 1.01

Table 5: The performance comparison over fit theme group.

Fashion Themes	**Outfits**			**Type-Aware**		**Theme-Graph**	
	Train	Validation	Test	Compat. AUC(%)	FITB Acc(%)	Compat. AUC(%)	FITB Acc(%)
Sports	406	7	78	89.40 ± 2.04	77.70 ± 6.53	**96.46** ± 1.24	**78.69** ± 4.34
Casual	1945	184	356	87.78 ± 1.57	76.54 ± 0.65	**93.60** ± 0.71	**76.52** ± 1.80
Office	908	64	144	94.06 ± 1.30	76.70 ± 4.53	**95.24** ± 1.85	**81.44** ± 0.69
Japanese	200	11	44	84.64 ± 0.70	73.42 ± 4.81	**90.68** ± 1.64	**74.91** ± 4.97
US	2182	151	400	90.81 ± 0.96	75.58 ± 1.00	**93.80** ± 0.97	**78.65** ± 1.73
UK	968	75	187	89.27 ± 3.14	74.24 ± 0.58	**94.58** ± 1.23	**77.01** ± 1.21
Girls	1600	116	278	88.02 ± 0.56	75.12 ± 1.79	**94.03** ± 0.46	**76.29** ± 1.72
Ladies	282	32	59	91.03 ± 2.76	71.55 ± 6.96	**95.27** ± 2.24	**70.85** ± 6.07
Simple	822	63	144	92.28 ± 0.74	78.29 ± 2.60	**93.68** ± 1.37	**76.79** ± 1.96
Nature	1398	113	254	89.43 ± 1.67	73.01 ± 4.82	**93.68** ± 1.47	**73.03** ± 2.83
Purk	180	17	27	83.79 ± 6.42	74.05 ± 5.56	**90.81** ± 3.21	**77.62** ± 5.64
Folk	29	3	5	68.33 ± 18.56	**87.00** ± 16.61	**83.33** ± 21.08	37.00 ± 28.91
Total	10920	836	1976	89.47 ± 0.61	75.43 ± 0.44	**93.84** ± 0.29	**76.69** ± 0.78

Table 6: The performance comparison over style theme group.

Fashion Themes	**Outfits**			**Type-Aware**		**Theme-Graph**	
	Train	Validation	Test	Compat. AUC(%)	FITB Acc(%)	Compat. AUC(%)	FITB Acc(%)
Female	7547	587	1368	**86.80** ± 0.57	70.35 ± 0.67	83.51 ± 0.95	**70.81** ± 0.71
Male	3493	266	653	**95.60** ± 0.38	**82.95** ± 1.79	94.38 ± 0.72	82.35 ± 0.83
Total	11040	853	2021	**89.70** ± 0.34	74.42 ± 0.42	87.21 ± 0.91	**74.54** ± 0.74

Table 7: The performance comparison over gender theme group.

		Top	Bottom	Shoe	Bag	Accessory	Item	Outfit
Before	Train	-	-	-	-	-	114806	17316
	Val	-	-	-	-	-	9070	1497
	Test	-	-	-	-	-	18604	3076
After	Train	13764	14849	15268	12640	12093	68614	16176
	Val	962	1052	1124	948	823	4904	1196
	Test	2000	2153	2314	1994	1712	10173	2463

Table 8: The statistics of items and outfits in the Polyvore dataset before and after data cleaning. The original Polyvore dataset does not have type labeling, so its type statistics are missing.

References

[1] Zunlei Feng, Zhenyun Yu, Yezhou Yang, Yongcheng Jing, Junxiao Jiang, and Mingli Song. Interpretable partitioned embedding for customized multi-item fashion outfit composition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 143–151. ACM, 2018.

[2] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7161–7170, 2018.

[3] Ruining He, Charles Packer, and Julian McAuley. Learning compatibility across categories for heterogeneous item recommendation. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 937–942. IEEE, 2016.

[4] Elaine M Bettaney, Stephen R Hardwick, Odysseas Zisimopoulos, and Benjamin Paul Chamberlain. Fashion outfit generation for e-commerce. *arXiv preprint arXiv:1904.00741*, 2019.

[5] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086. ACM, 2017.

[6] Yong-Siang Shih, Kai-Yueh Chang, Hsuan-Tien Lin, and Min Sun. Compatibility family learning for item recommendation and generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[7] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, 19(8):1946–1955, 2017.

[8] Pongsate Tangseng, Kota Yamaguchi, Takayuki Okatani, et al. Recommending outfits from personal closet. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2275–2279, 2017.

[9] Takuma Nakamura and Ryosuke Goto. Outfit generation and style extraction via bidirectional lstm and autoencoder. *arXiv preprint arXiv:1807.03133*, 2018.

[10] 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.

[11] Wei-Lin Hsiao and Kristen Grauman. Learning the latent "look": Unsupervised discovery of a style-coherent embedding from fashion images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4213–4222. IEEE, 2017.

[12] Edgar Simo-Serra and Hiroshi Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 298–307, 2016.

[13] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4642–4650, 2015.

[14] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. Neurostylist: Neural compatibility modeling for clothing matching. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 753–761. ACM, 2017.

[15] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, 2015.

[16] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 830–838, 2017.

[17] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *CVPR*, 2019.

[18] Kota Yamaguchi, Takayuki Okatani, Kyoko Sudo, Kazuhiko Murasaki, and Yukinobu Taniguchi. Mix and match: Joint model for clothing and attribute recognition. In *BMVC*, page 4, 2015.

[19] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee, 2016.

[20] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[21] Long Chen and Yuhang He. Dress fashionably: Learn fashion collocation with deep mixed-category metric learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.

[23] Michael Wilber, Iljung S Kwak, David Kriegman, and Serge Belongie. Learning concept embeddings with combined human-machine expertise. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 981–989, 2015.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Per-Erik Danielsson. Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248, 1980.

[26] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.