FashionVQA: 
A Domain-Specific Visual Question Answering System

Min Wang, Ata Mahjoubfar, Anupama Joshi
Target Corporation
{Min.Wang, Ata.Mahjoubfar, Anupama.Joshi}@target.com

Abstract

Humans apprehend the world through various sensory modalities, yet language is their predominant communication channel. Machine learning systems need to draw on the same multimodal richness to have informed discourses with humans in natural language; this is particularly true for systems specialized in visually-dense information, such as dialogue, recommendations, and search engines for clothing. To this end, we train a visual question-answering (VQA) system to answer complex natural language questions about apparel in fashion photoshoot images. The key to the successful training of our VQA model is the automatic creation of a visual question-answering dataset with 168 million samples from item attributes of 207 thousand images using diverse templates. The sample generation employs a strategy that considers the difficulty of the question-answer pairs to emphasize challenging concepts. We see that using the same transformer for encoding the question and decoding the answer, as in language models, achieves maximum accuracy, showing that visual language models (VLMs) make the optimal visual question-answering systems for our dataset. The accuracy of the best model surpasses the human expert level. Our approach for generating a large-scale multimodal machine learning model to automatically, consistently and precisely infer the visual attributes of the fashion items.

1. Introduction

Fashion is about 2% of the world’s GDP and a significant sector of the retail industry. Whenever a new fashion item like apparel or footwear is launched, the retailer needs to prepare and show rich information about the product, including pictures, text descriptions, and detailed attribute tags. The attributes of the fashion products, including color, pattern, texture, material, occasion-to-use, etc., require domain experts to label them piece by piece. This labeling process is time-consuming, costly, subjective, error-prone, and fundamentally imprecise due to the interdependency of the attributes. To address these issues, we introduce a multitask multimodal machine learning model to automatically, consistently and precisely infer the visual attributes of the fashion items.

Each item is typically labeled with multiple tags that describe different attributes of the item. For example, an item can be labeled with “shirt”, “red”, “solid pattern”, “blue collar” and “short sleeve”. An intuitive way of learning such information is to train a multi-label classifier, which outputs the probability of multiple labels of each input sample. However, such a model cannot encode the relationship between different attributes. For example, “short sleeve” is a suitable attribute for “shirt”, but not for “jeans”, and “red” only describes the body part of the shirt, but not the collar. The model needs to learn attribute and object relationships and adjusts its output accordingly.

We propose designing a Visual Questioning Answer (VQA) framework for fashion items, in which the model is trained to answer complex natural-language questions, such as “is the person wearing a red shirt with a solid pattern and blue collar?”, given the input image. The VQA task is more challenging than the simple attribute classifier since it requires a thorough understanding of both the question and the structure and relationship between various visual attributes in the image. By training such a model, we convert the manual process of tagging new products with visual attributes into automated answering of a series of questions with visual intents (auto-labeling). The model also generates multimodal embeddings of the product images attended to the questions for downstream dialogue, search, and recommendation systems.

Prior to our work, there exists a large-scale VQA v2 dataset [11], which includes 0.6 million question-answer-image triplets. It has been widely used as the benchmark in recent research on VQA tasks. However, this general dataset only contains a small number of samples related to fashion. In this work, we build a fashion VQA dataset from a diverse
apparel product database. The questions, including both binary and non-binary, are automatically composed by filling question templates with the given attribute information. The dataset contains 207 thousand images and 168 million question-answer-image triplets. The automatic generation of the VQA dataset from a limited number of images and attributes allows us to achieve the scale required for training a multimodal domain expert model.

We leverage a cross-modality fusion model mapping representations from visual and text space to the same latent feature space and performing answer prediction with classifier modules. Given an image that contains a fashion item and the corresponding questions regarding its different attributes, the model predicts the answers to the given questions. We can then use the model to generate the missing or alternative attribute information based on its answers.

Additionally, given different but similar text descriptions on the same item, we can generate consistent feature embeddings that enable us to build better online search services. The existing search engines cannot attend to the relevant visual parts of a fashion item given the query and do not adapt the attention mask according to the chained adjectives. With this work, we can map the input query to the learned embedding space and perform a robust and fuzzy search in that multimodal space. We can also provide a visual dialogue service, in which the customers can ask consecutive questions to narrow down the item list according to their apparel preferences. We can also build a fashion recommendation system in the multimodal embedding space. The customer-item interaction history is mapped to this space, and the neighboring items are recommended.

2. Related work

Cross-modality fusion models: Cross-modality fusion model is a core component of the VQA framework. It aligns the features from the visual and language modalities. Initially-proposed VQA models identify the high-level cross-modal interactions by Bilinear Fusion [9]. MCB [7], MLB [17] and MUTAN [2] are later introduced to achieve better fusion performance at much lower computational cost and parameters. Motivated by the remarkable performance of the attention mechanism in language and vision models [6] [31], the attention module becomes the fundamental block in designing the cross-modality fusion models [4, 8, 21, 22, 24, 30, 37, 38].

Fashion datasets: In recent years, many valuable fashion datasets [23] [10] [39] [40] [33] [12] [15] [3] [36] [32] [34] have greatly contributed to clothing item recognition and apparel attribute understanding. However, most of them suffer from some limitations when considered for training versatile VQA models.

3. Methods

In this section, we describe how we designed and generated a novel VQA dataset for fashion. We named the new dataset FashionVQA dataset. Each sample in the dataset is a question-answer-image triplet.

3.1. Question templates

We adopt a templating mechanism to automatically create question-answer pairs from fashion items’ meta-information. The question templates are designed based on a set of fixed rules that meet the English grammar and result in human-readable sentences. By filling the question templates with specific item attribute (e.g., color, pattern...etc), attribute value (e.g. red, green, stripe...), category (e.g. shirt, pants...etc), and location (e.g. “on the top”, “on the bottom”...), we can generate a variety of questions for each image. The answer to each question can be “Yes/No” for binary questions and multiple choices from the relevant attribute values for non-binary questions.

Since the images from the FashionVQA dataset are all photoshoot images with a solid background, the question templates ask only attribute-related questions about the fashion items in the image. For example, “what is the sleeve length of this shirt on the top?” or “is this a white v-neck sweater?”. The basic template is structured as “{question type} {this/these} {a/an/} {pair of/pairs of/} {object} {location}?”. When filling the template to expand into a full sentence, the choices between “is/are”, “this/these”, “a/an”, “a pair of/pairs of”, and singular or plural format of category are required to follow the English grammar and be aligned with the number of targeted fashion item in the image. The question templates fall into two primary categories based on the answer types: binary and non-binary templates.

Binary question templates: Binary question templates typically start with “is this/are these”, “can you see”, or “is there any {part} on this/these”, followed by the description of the targeted item in the format of “{location} {a}/{a pair of}/( ) {attribute value 1} {attribute value 2} {category}”, where attribute value 1 and attribute value 2 are two attribute values from different attributes. Permuting attribute value 1, attribute value 2, category in different orders yields different question templates. Conjunction words like “with”, “and”, or “in” can be used in templates when attribute value 1 or attribute value 2, or both are located after category. The most common question types used in binary questions are “is/are” and “can”.

Non-binary question templates: Non-binary question templates typically start with question words like “what” / “why” / “when” / “how” followed by terms of attribute. The formats of the question type vary from attribute to attribute.
3.1.1 Balance positive and negative samples for each binary question

Given binary and non-binary question templates and attribute values for a specific image, we can easily generate non-binary question-(multiple answers)-image triplets and binary question-(positive answer)-image triplets.

For a balanced VQA dataset, we expect each binary question to come with the same number of positive and negative samples, i.e., balanced (question, “Yes”, image ID) triplets and (question, “No”, image ID) triplets. Here is the strategy to achieve it.

First, we build an attribute-value-to-images dictionary to map each distinct attribute value to a set of eligible image IDs. Given a specific attribute value, we collect a set of positive answer image IDs directly from this attribute-value-to-images dictionary using given attribute value and its synonyms. The negative answer image IDs are collected from all image IDs of the same attribute excluding the positive image IDs. More concretely, to maximally reduce the noise in the positive/negative answer image IDs, we need to verify the relationship among attribute values as alternative, hierarchical, or exclusive terms. Examples of alternative terminologies are “sweatpants”, “jogger pants”, and “lounge pants”; examples of hierarchical terminologies are “blue”, “light blue”, and “sky blue”; and, examples of exclusive terminologies are “light blue” and “dark blue”. We expect attribute values with similar terminologies (alternatives and parents of hierarchical terms) to contain the same set of positive samples, so they are considered synonyms. In this manner, we can build an attribute-value-to-(positive/negative answer)-images dictionary.

Then, we consider all the combinations of assorted attributes with category. For example, (color, pattern, category), (color, category), (material, neckline type, category), etc. For each combination, we further expand the attribute-value-to-(positive/negative answer)-images dictionary by mapping the combination of one specific attribute value and one specific category (e.g., (red, shirt)) to its positive/negative answer image ID set.

With the attribute-value-to-(positive/negative answer)-images dictionary, we can easily generate different binary questions via filling the question templates with each combination of attribute value and category in the dictionary. We can pick a fixed number of positive and negative answer image IDs to guarantee the sample balance for each question. Following the same formula, we can easily expand the combinations to multiple attribute values and one category.

3.2. Dataset description

FashionVQA: FashionVQA dataset includes 207,654 unique photoshoot images. We use 169,406 images in the train split for training and 38,248 images in the validation split for evaluation. The train split is composed of 163M question-answer-image triplets and the validation split includes 5.2M question-answer-image triplets.

4. Benchmarks

We benchmark the FashionVQA dataset by training several VQA models to learn the interaction between images and questions. Given the visual embedding of the input image and text embedding of the input question sentence, we train the model to output the given answer to the question. The dataset is used to train two variants of the MCAN [37] model and a MUTAN [2] model. One MCAN variant, named MCAN*-v1, is a modification of the MCAN-small, which includes only two encoder-decoder modules. The other variant is named MCAN*-VLM, which has a similar structure to MCAN*-v1, but instead of an answer classifier, it has a token classifier covering all of the question and answer tokens. For MCAN*-VLM, the answer to each question is tokenized as one token and concatenated with the question tokens as the language input. The special token ‘SEP’ is inserted between the question and the answer. Also, ‘EOS’ token is used at the end of the answer. During the training of MCAN*-VLM, we randomly mask one token and predict the masked token as in the masked language modeling, similar to BERT [6]. Table 1 lists the benchmark results of the three aforementioned models on the validation split of our FashionVQA dataset. The results show that MCAN*-VLM works better than MCAN*-v1 and MUTAN, indicating that a decoder-only visual language model (VLM) performs better than the dedicated VQA architectures. Figure 1 visualizes the attention map from two validation samples for a series of binary and non-binary questions.

Table 1. Benchmarks of different VQA models

<table>
<thead>
<tr>
<th>Model</th>
<th>All Acc</th>
<th>Non-binary Acc</th>
<th>Binary Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUTAN</td>
<td>81.38%</td>
<td>61.62%</td>
<td>87.43%</td>
</tr>
<tr>
<td>MCAN*-v1</td>
<td>84.42%</td>
<td>64.32%</td>
<td>90.58%</td>
</tr>
<tr>
<td>MCAN*-VLM</td>
<td>84.69%</td>
<td>64.65%</td>
<td>90.84%</td>
</tr>
</tbody>
</table>

Table 2. Compare the MCAN*-VLM model to a human expert

<table>
<thead>
<tr>
<th></th>
<th>All Accuracy</th>
<th>Non-binary Accuracy</th>
<th>Binary Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Expert 1</td>
<td>62.3%</td>
<td>30.5%</td>
<td>74.3%</td>
</tr>
<tr>
<td>MCAN*-VLM</td>
<td>77.7%</td>
<td>47.6%</td>
<td>89.0%</td>
</tr>
<tr>
<td>p-value</td>
<td>1.9e-05</td>
<td>0.0125</td>
<td>3.4e-05</td>
</tr>
</tbody>
</table>

Table 2. Compare the MCAN*-VLM model to a human expert

5. Comparison to human performance

Human accuracy for FashionVQA dataset: We asked 9 human annotators including 2 experts (trained each expert with at least ten examples per fashion term) to answer each
we generate a set of binary questions from the search query. A baseline search engine returns the top 24 items for an apparel search query. Another variant of the search results generates the cross-modality embeddings of the vision and language model trained on the FashionVQA dataset. The model outperforms all of the human annotators at a 95% confidence level, and the differences in the accuracies between the model and human accuracies are statistically significant.

**Accuracies for human-generated questions:** We also stress-tested the model by measuring its performance on human-generated questions. We asked another expert annotator to paraphrase the questions of 300 random samples (218 binary and 82 non-binary) from the validation set. We used these questions instead of the original questions in the validation set to measure the accuracies of the MCAN*-VLM model and a human annotator, Expert 1 (Table 2).

We performed a one-sided t-test to analyze the statistical significance of the difference between the human and the model accuracies. At a significance level of 0.05 ($\alpha = 0.05$), the p-values reject the null hypothesis of the human accuracy being greater than or equal to the model.

**Impact on downstream tasks:** We performed a side-by-side comparison of the apparel search with/without FashionVQA. A baseline search engine returns the top 24 items for an apparel search query. Another variant of the search results is formed by reranking these 24 items with FashionVQA: we generate a set of binary questions from the search query and use MCAN*-VLM model trained with FashionVQA to answer these questions for each of the 24 items. The average confidence scores of the yes and no answers are used as additional features to rerank the top 24 items.

For a number of randomly-selected search queries with two attribute values and one category, e.g., “green crew neck dress”, a human annotator is presented with the original and reranked search result pages (randomly located on the left and right sides of the screen) and gets to choose her/his preferred result page. Out of 150 search queries, the human annotator preferred 117 search pages reranked based on the FashionVQA. Binomial statistical test results in a $p$-value of 3.2e-12, showing that the human annotator significantly prefers the search result page reranked using FashionVQA.

**Conclusion**

In this work, we design a fashion VQA dataset and generate non-binary and binary questions via diverse templates. The templates allow us to flexibly scale the dataset to the size and complexity required for training a domain-specific multimodal model. We benchmark this large-scale dataset on different VQA models. The best model is a visual language model trained on the FashionVQA dataset. The model generates the cross-modality embeddings of the vision and language domains applicable to downstream tasks of fashion dialogue, search, and recommendation.
References


