Image Captioning on General Data and Fashion Data

An Attribute-Image-Combined Attention-Based Network for Image Captioning on Multi-Object Images and Single-Object Images

Tu, Guoyun

Award date:
2020

Awarding institution:
Royal Institute of Technology

DEGREE PROJECT IN COMPUTER SCIENCE AND ENGINEERING,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2020

Image Captioning on General Data and Fashion Data

An Attribute-Image-Combined Attention-Based Network for Image Captioning on Mutli-Object Images and Single-Object Images

GUOYUN TU

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Image Captioning On General Data And Fashion Data

An Attribute-Image-Combined Attention-Based Network for Image Captioning on Mutli-Object Images and Single-Object Images

GUOYUN TU

Master's Programme, Data Science, 120 credits
Date: September 14, 2020
Supervisor: Ying Liu, Mårten Nilsson
Examiner: Vladimir Vlassov
School of Electrical Engineering and Computer Science
Host company: Norna AI
Swedish title: Bildtexter på allmänna data och modedata
Swedish subtitle: Ett attribut-bild-kombinerat uppmärksamhetsbaserat nätverk för bildtextning på Mutli-objekt-bilder och en-objekt-bilder

Image Captioning On General Data And Fashion Data /
Bildtexter på allmänna data och modedata

Abstract

Image captioning is a crucial field across computer vision and natural language processing. It could be widely applied to high-volume web images, such as conveying image content to visually impaired users. Many methods are adopted in this area such as attention-based methods, semantic-concept based models. These achieve excellent performance on general image datasets such as the MS COCO dataset. However, it is still left unexplored on single-object images.

In this paper, we propose a new attribute-information-combined attention-based network (AIC-AB Net). At each time step, attribute information is added as a supplementary of visual information. For sequential word generation, spatial attention determines specific regions of images to pass the decoder. The sentinel gate decides whether to attend to the image or to the visual sentinel (what the decoder already knows, including the attribute information). Text attribute information is synchronously fed in to help image recognition and reduce uncertainty.

We build a new fashion dataset consisting of fashion images to establish a benchmark for single-object images. This fashion dataset consists of 144,422 images from 24,649 fashion products, with one description sentence for each image. Our method is tested on the MS COCO dataset and the proposed Fashion dataset. The results show the superior performance of the proposed model on both multi-object images and single-object images. Our AIC-AB net outperforms the state-of-the-art network, Adaptive Attention Network by 0.017, 0.095, and 0.095 (CIDEr Score) on the COCO dataset, Fashion dataset (Bestsellers), and Fashion dataset (all vendors), respectively. The results also reveal the complement of attention architecture and attribute information.

Keywords

Image captioning, fashion data, attention based, text attributes

Sammanfattning

Bildtextning är ett avgörande fält för datorsyn och behandling av naturligt språk. Det kan tillämpas i stor utsträckning på högvolyms webbbilder, som att överföra bildinnehåll till synskadade användare. Många metoder antas inom detta område såsom uppmärksamhetsbaserade metoder, semantiska konceptbaserade modeller. Dessa uppnår utmärkt prestanda på allmänna bilddatamängder som MS COCO-dataset. Det lämnas dock fortfarande outforskat på bilder med ett objekt.

I denna uppsats föreslår vi ett nytt attribut-information-kombinerat uppmärksamhetsbaserat nätverk (AIC-AB Net). I varje tidsteg läggs attributinformation till som ett komplement till visuell information. För sekventiell ordgenerering bestämmer rumslig uppmärksamhet specifika regioner av bilder som ska passera avkodaren. Sentinelgrinden bestämmer om den ska ta hand om bilden eller den visuella vaktposten (vad avkodaren redan vet, inklusive attributinformation). Text attributinformation matas synkront för att hjälpa bildigenkänning och minska osäkerheten.

Vi bygger en ny modedataset bestående av modebilder för att skapa ett riktmärke för bilder med en objekt. Denna modedataset består av 144 422 bilder från 24 649 modeprodukter, med en beskrivningsmening för varje bild. Vår metod testas på MS COCO dataset och den föreslagna Fashion dataset. Resultaten visar den överlägsna prestandan hos den föreslagna modellen på både bilder med flera objekt och enbildsbilder. Vårt AIC-AB-nät överträffar det senaste nätverket Adaptive Attention Network med 0,017, 0,095 och 0,095 (CIDEr Score) i COCO-datasetet, modedataset (bästsäljare) respektive modedatasetet (alla leverantörer). Resultaten avslöjar också komplementet till uppmärksamhetsarkitektur och attributinformation.

Nyckelord

Bildtexter, modedata, uppmärksamhetsbaserat, textattribut

Acknowledgments

This work is supported by Norna AI, in Stockholm, Sweden. They provided with GPU machines and raw data. I would like to thank my supervisors, Ying, and Mårten for giving helpful suggestions on both literature research and model implementation during the project. I am sincerely appreciated for receiving mental support and technical support from all my colleagues, friends, and my examiner, professor Vladimir Vlassov.

Stockholm, September 2020
Guoyun Tu

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The significant growth in the number of web images has brought plenty of opportunities on computational understanding. In this thesis, the performance of deep learning models on single-object fashion image captioning is evaluated and discussed. The contribution of this paper is summarized as follows: (1) We proposed Attribute-Image-Combined Attention-Based Network (AIC-AB Net). It applies encoder-decoder architecture for image captioning, and furthermore an extension on Long Short Term Memory (LSTM). Spatial attention and "visual sentinel" are employed for generating attention distribution on images. attribute information is added for highlighting the key features of images and eliminating uncertainty. (2) We conduct our experiments on a new fashion dataset where images contain one single fashion object with one sentence of description, which is created especially for the single-object image captioning task and fashion image captioning. Although this dataset will not be published, we publish our code and its implementation on the MS COCO dataset to serve reproducibility purpose.

1.1 Background

Automatically generating captions for images has attracted attention from both academic and industrial fields. Image captioning could be roughly split into two parts, image recognizing and caption generating. Image recognizing requires to detect and recognize the objects shown in the image. In addition, the location, type of scene, amount of objects, and interactions among objects should also be under consideration. Caption generator summary the extracted information and form them into text which human could understand. Despite human can largely understand the images without any detailed captions. There

are always some scenarios that automatic image captions interpreted from machines help a lot. For example, it can help visually impaired web-users navigating the image content. Another application of image captioning is a content analysis of a large number of images.

In traditional machine learning, researchers extract hand-crafted features from input data and mostly use the combination of them to train a classifier such as Support Vector Machines (SVM). Those methods include Local Binary Patterns (LBP)[1], Scale-Invariant Feature Transform (SIFT)[2] and the Histogram of Oriented Gradients (HOG)[3]. The shortcomings of this process are obvious. Since handcrafted features only suit a specific task or dataset, it is not feasible to extract features from a large and diverse set of data. It is also not able to understand complex image dataset with a large number of semantic interpretations.

Significant successes have been achieved on the problem of image captioning based on deep learning methods. Especially in the last 5 years, a large number of articles exploring image captioning have been popularly applied on open-sourced images which contain multiple objects and rich contextual information. In previous work, lots of image captioning methods were proposed, such as Visual space-based model [4], multimodal space-based model[5], dense captioning [6], whole scene-based model, encoder-Decoder architecture-based model[7], compositional architecture-based model[8], attention-based model[9], semantic concept-based model [10], stylized captions [11].

1.2 Problem

A major blind spot of previous studies on image captioning is that their experiments were conducted on multiple-object images. It remains a hardly-been-asked and open problem to generate a caption on one single object. We identify that the main challenge faced is to ensure the generated description to include a satisfactory amount of adjectives and pay attention to details on the object. In this paper, two questions will be answered:

- Will deep learning models achieve consistent performance on multi-object images (COCO dataset) and single-object images (Fashion dataset)?
- How does an attention-based architecture and attribute information added on influence the model performance on image captioning tasks?

1.3 Purpose

Observing this problem, we create a fashion dataset which contains 144422 images from 24649 products. Each data is composed of one fashion image and one description sentence. It is noteworthy that a couple of products may map to the same label. The fashion dataset processes 10091 unique captions as labels. These images are from different vendors, Uniqlo, Toteme-Studio, Bestseller, Drykorn, Jlindeberg, Joseph-fashion, Marc-o-polo, Rodebjer, Tigerofsweden, Vince respectively. Since the form and style of captions from different vendors vary a lot, we select data from Bestseller and build a subset to further assist the evaluation. The Bestseller subset contains 89756 images from 19385 products and 8448 unique captions.

Meanwhile, we propose an attribute-image-combined attention-based architecture, the attention structure added in the output layers of LSTM model teaches our model where to look on the images, the attribute information fed into each step of LSTM model teaches our model what to say on the descriptions. Here the attributes are obtained by an auxiliary CNN classifier as a attribute extractor. The source code is available at https://github.com/guoyuntu/Image-Captioning-On-General-Data-And-Fashion-Data.

1.4 Goal

The goal of this degree project includes two fields. The first one is in the academic domain, we compared the performance of deep learning model on both multiple-object and single-object images and discuss the divergence and reason behind it. The second is about the application domain, we proposed an attribute-image-combined attention-based network, which is expected to achieve better performance on the image captioning task.

1.5 Contribution

The main contributions of this project are:

- We propose a fashion dataset that is cleaned from the raw data provided by Norna AI. It contains 144,422 images from 24,649 products, which has the following features: 1) each image contains one single fashion product with one descriptive sentence. 2) the captions include a satisfactory amount of adjectives and nouns.

- We suggest that the ability to locate the relevant region of an image when generating different words and the combination with the attribute information is crucial for accurate caption generation. Toward this end, we present an encoder-decoder network with adaptive attention architecture and attribute information, called Attribute-Image-Combined Attention-Based Network (AIC-AB Net). The adaptive attention architecture is composed of spatial attention and visual sentinel, the former determines where to "look" and the latter decides when to "look". The attribute information measures the global similarity between images and text and helps disambiguate noisy visual detection.

1.6 Benefits, Ethics and Sustainability

Image captioning could produce great commercial benefits. As an illustrative example, a fashion retailer website can generate descriptions for clothing images, saving unnecessary labor and time. Proper commercial placement would benefit from the precise location on images where the describable words point at, in addition to a high-quality overall description of products.

Ethics plays a vital role in artificial intelligence (AI) development. Whereas some research [12] suggests AI could not always perform ethically right things in the correct manner and for the proper reasons due to its unreliability. We argue that our project helps reduce bias, like other computer vision methods, thus processes ethical value. Human description varies with their cultural and educational background. Bias somehow exist in every individual's concept. However, image captioning by AI integrates captions from multiple references, thus could be considered more neutral and unbiased. It is also noteworthy that depends on the quality of training data as well.

We conduct the sustainability analysis on the economic dimension and technical dimension. On the financial aspect, this project directly builds a network that generates descriptions for fashion images and produces a spatial map highlighting image regions relevant to each generated word, saving massive labor and time costing on captioning such repetitive and ineffective work. On the technical aspect, this project extends previous image captioning works, producing new data and methods for single-object images. A new dimension is created for testing the robustness of image captioning models.

1.7 Methodology

In order to serve the two research goals, our AIC-AB network is implemented in both benchmark, MS COCO dataset and fashion dataset, compared with three ablated versions of our network, naive encoder-decoder network, attention-based network, and attribute-image-combined network.

Since the similarity and accuracy of two sentences is hard to define. Metrics play a significant role on the evaluation of image captioning tasks. In this project, BLEU[13], METEOR[14], ROUGE-L[15] and CIDEr[16] score are employed. Research claims that BLEU, METEOR, and ROUGE-L focus more on grammaticality correctness while CIDEr focuses more on semantically correctness[17]. Besides, the calculation of BLEU, METEOR, and ROUGE-L involves precision between the generated caption and the ground truth. Therefore, these scores are on different scales when the ground truth is unique (Fashion Dataset) or is composed of five captions (MS COCO Dataset). To the best of our knowledge, the CIDEr score is recommended to be the most important criterion in this project.

In this project, pre-processing and intermediate steps also affect the overall performance and evaluation. The raw data of the fashion dataset is scraped from open websites thus its descriptions are sub-standard. Most of them consist of various amount of sentences, some sentences are even incomplete, e.g. “size: M, model: 180cm”. To create a standard fashion dataset, several filters are applied for pre-processing, including sentiment filter, position-of-speech filter, and other filters. Furthermore, this project trains a text attribute extractor to extract attributes from the image. The performance of this text attribute extractor will also be discussed in the report.

Further important results will be demonstrated as well. For instance, in this project, we split the Fashion dataset into two versions, all-vendor and Bestseller, which brings an interesting question, how the model performs when transferred from one vendor to another. Besides, how the attention architecture performs is also an interesting question with respect to the approach itself. Thus, the results of transfer learning and visualization of the attention distribution map are worthy of being showed and discussed.

Further details of the research methodology and approaches, as well as the datasets used, are presented in Chapter 3.

1.8 Delimitations

Due to the time limitation and lacking up-to-standard labeling. One main delimitation we identified here is that mistakes and deviation still exist in the labels of fashion data. Therefore, perhaps not all the potentials or issues of our network are fully explored. It will be left to potential future works that build upon the results and findings of this project. According to the result of transfer learning, our network fails to be robust when transfer from one vendor to another. This may be caused by the distinct captioning style of a vendor. When the Fashion dataset is labeled in a systematical and uniform method, the robustness could be better evaluated and discussed.

Furthermore, failure cases of attention maps expose one main defect of AIC-AB network on the Fashion dataset. It sometimes fails to focus on the specific location of clothing, e.g. sleeves and waist. It still "looks at" the whole clothes. Compared with the COCO dataset, which contains multiple objects in one image, fashion images doubtlessly lack details. Raising the resolution of the attention map may bring benefits to address this problem.

1.9 Outline

This paper is organized as follows. In chapter 2, the theoretical background and related work will be introduced, including multimodal image captioning, encoder-decoder model and attention mechanism. In chapter 3, the details of the attribute-image-combined attention-based network will be explained. In chapter 4, the implementation and results of the experiments will be demonstrated. And chapter 5 is the conclusion.

Chapter 2

Theoretical Background

In this chapter, a detailed description of the background of the degree project is presented, including an outline of related work. We first describe the generic image captioning methods in Sec. 2.1.1, as it is the theoretical foundation of this paper. Then the approaches, which inspire this work and are applied in our proposed network, encoder-decoder architecture and attention mechanism are introduced in Sec. 2.1.2 & 2.1.3.

2.1 Related Work

2.1.1 Image Captioning

The main categories of image captioning approaches include retrieval-based image captioning, template-based image captioning, and language model caption generation. Template-based methods generate captions filling the blank slots based on a fixed template. For example, Farhadi et al.[18] use a triplet of scene elements to fill the template slots. Retrieval-based methods select the visually similar images with their captions which is called candidate images and the captions for the query image are generated from these captions pool. Yunchao et al.[19] use large weakly annotated photo collection to improve image sentence embeddings.

Most recent caption generation methods use deep learning based techniques. Captions can be generated by various approaches. According to the type of learning, most research applies supervised learning[4, 10, 20, 7], While Vijay et al.[21] proposed actor-critic reinforcement learning to train a network. GAN based models have been also proved successful according to Bo Dai et al [22].

From feature mapping, the visual space-based methods feed the image features. The corresponding captions are independently passing the language decoder. In contrast, multimodal methods learn multi-source information extracted from both the images and the corresponding caption-text [5].

According to the language model to generate captions, neural probabilistic language model [23], log-bilinear models [24], skip-gram models [25] have been proposed for sequence to sequence tasks. Besides, LSTMs are able to handle long-term temporal dependencies, thus is popularly used in caption generation tasks [7]. Gu et al. [26] proposed a language-CNN for statistical language modeling for image captioning.

2.1.2 Encoder-Decoder for Image Captioning

Encoder-Decoder methods have been a popular approach to tackle language tasks, such as machine translation. This architecture also provides the capability to encode visual information and decode it as a neural language for image captioning. A typical network of this category extracts global image features from the hidden activations of a convolutional neural network (CNN) and then feeds them into a long short term memory network (LSTM) in order to generate a sequence of words. Fig. 2.1 provides an illustration of these methods.

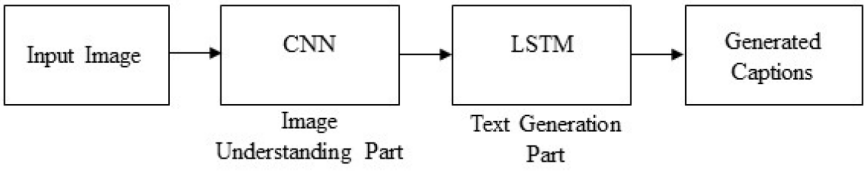

Figure 2.1: A block diagram of a compositional network-based captioning

A typical method of this category adopts the following step: 1) A vanilla CNN is used to process images, to detect objects and their relationships. 2) The output is fed into a language model to translate them into words and combined phrases that can be used to produce image captions composed of a sequence of words. Serving this purpose, LSTM is the most effective approach for such a time series problem.

Vinyals et al. [7] proposed a network called Neural Image Caption Generator (NIC), which uses a CNN for image representations and an LSTM for image captions generation. Mao et al.[27] proposed a special text generation method for images. This method, which is called Referring

Expression, is able to generate a description for each specific object or region. Recently, object detection and classification research [28, 29, 30] showed that deep, hierarchical methods perform better at learning than shallower ones. Wang et al. [31] proposed a deep bidirectional LSTM-based method for image captioning. The proposed architecture is composed of a CNN and two separate LSTM networks. It utilizes both past and future context information to generate contextually and semantically accurate image captions. In order to address the common gradient vanishing problem, Jia et al. [32] proposed a guided LSTM (gLSTM). This gLSTM adds global semantic information to each gate and cell state of LSTM to generate long sentences.

2.1.3 Attention-based Model

Following the trend of using the encoder-decoder architecture on image captioning, researchers also explored many techniques that they have found to be effective. Methods based on attention mechanisms have been increasingly popular since they provide computer vision algorithms with the capacity of knowing where to look. In fact, parts of the description are relevant to certain regions of the image. However, simple encoder-decoder networks, with CNN to extract the visual information and LSTM to generate sentences of words, are unable to link the spatial aspects of the image with the parts of the image captions. The attention mechanism has been increasingly popular since it could address this problem. Instead of regarding the image as a whole scene, attention-based network dynamically focuses on the various parts of the input image while the output captions are being generated. A block diagram of the attention-based image captioning method is shown in Fig. 2.2.

A typical attention-based method could be split into three steps: 1) Visual information is extracted from the whole scene of the input image by a CNN. 2) The language generation network, usually LSTM applied, generates words or phrases based on visual information. 3) In each time step of the language model, salient regions of the given image are focused according to the generated words or phrases.

Xu et al.[9] were the first to introduce an attention-based image captioning method. The proposed method can concentrate on the salient parts of the image and generate the corresponding words at the same time, which is the main difference between the attention-based methods and other methods. There are several variants of the attention-based techniques for image captioning. Jin et al.[32] proposed another attention-based image captioning network. This method detects the semantic relationship between visual

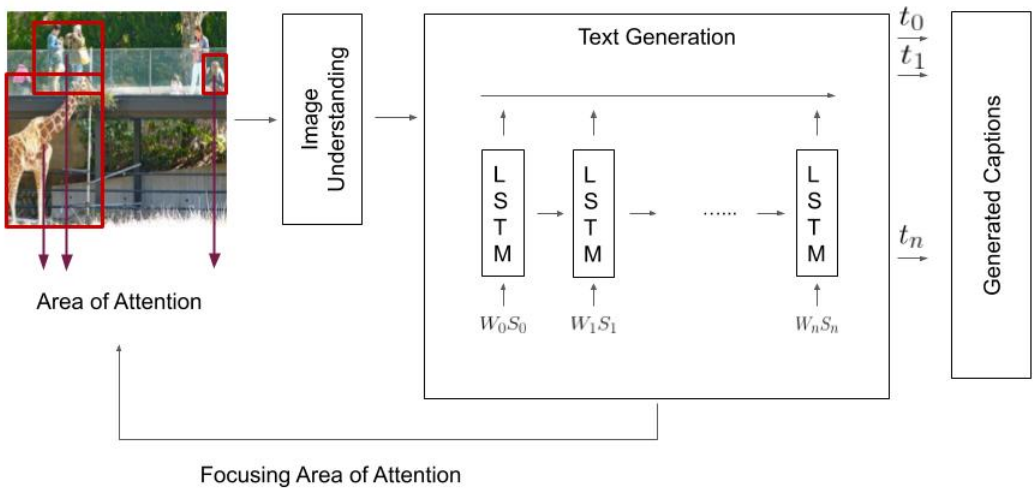

Figure 2.2: A block diagram of typical attention-based image captioning technique.

information and textual information and extracts the flow of abstract meaning. Pedersoli et al.[33] proposed an area-based attention mechanism for image captioning. The proposed method can predict the next caption word and corresponding image region as well as corresponding image regions in each time-step of RNN. Liu et al.[34] improved the method by introducing the evaluation and correction of the attention map at every time step.

2.1.4 Evaluation Metrics

In order to measure the quality of the generated captions compared with the ground truth, a number of evaluation metrics are proposed. Each metrics applies its own calculation technique and has distinct merits. Four metrics, BLEU, ROUGE, METEOR, CIDEr are popularly used.

BLEU. BLEU (Bilingual evaluation understudy)[13] is proposed in 2002 by Kishore Papineni et.al. It estimates the overall quality of the generated text with a set of references. The performance of the BLEU score varies according to the number of references and the size of the generated text. Its limitation is that the BLEU scores perform well only if the size the generated text is small.[35] Besides, it fails to consider the syntactical correctness [36].

ROUGE. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [15] is a set of metrics originally used for measuring the quality of text summary. Variants such as ROUGE-1,2, ROUGE-W, ROUGE-SU4 compare n-gram, word pairs, and word sequences in different scenarios. However, it shows poor performance in multi-document test summary tasks.

METEOR. METEOR (Metric for Evaluation of Translation with Explicit Ordering)[14] is proposed for evaluating the machine translated language by Satanjeev Banerjee and Alon Lavie. It takes word segments, stems of a sentence, and synonyms into account and consider their matching between the generated text and the reference. It outperforms at the sentence and the segment level.

CIDEr. CIDEr (Consensus-based Image Description Evaluation)[16] achieves human consensus on image captioning tasks. It is based on the term frequency-inverse document frequency. It is able to work with a various number of reference and capture the consensus between generated captions and human judgment. It achieves better performance than the metrics as stated above across sentences generated by various sources.

Chapter 3

Attribute-image-combined Attention-based Network

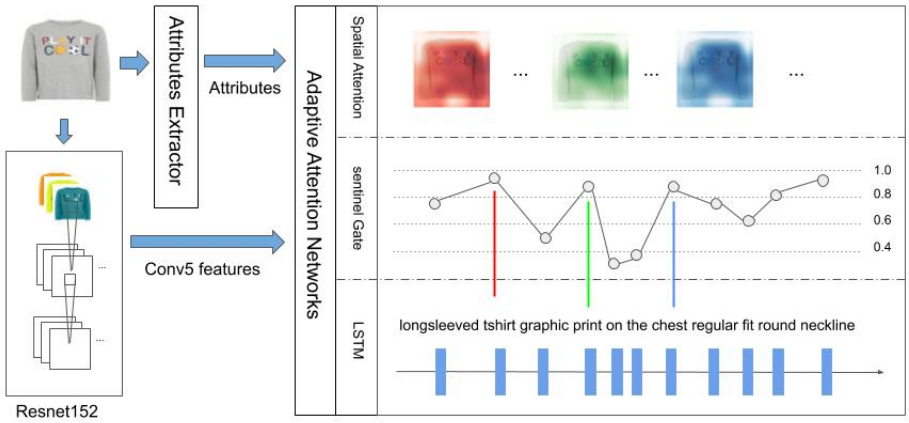

Figure 3.1: An overview of the AIC-AB Net. We extract the visual features and attribute information from the images. The visual features pass an adaptive attention architecture, which can automatically determine where to look (Spatial Attention) and when to look (visual sentinel). The attributes are fed into every time steps of the LSTM decoder for frequently updating.

Attribute-image-combined Attention-based Network (AIC-AB Net) is an end-to-end network that tackles image captioning in the meantime generate the attention map of the image. In this section, we describe its two constituents:

the adaptive attention architecture and the attribute combination method. The former is a structure proposed by [34], which endows the network with spatial attention and visual sentinel, which determine when to rely on visual information and when to rely on the text information. The latter extracts text attributes from the image and feeds it into LSTM to improve the performance. Fig. 3.1 provides an illustration of the network architecture. As the example shows, our model learns to attend to the image more when generating words "tshirt", "print", "chest" and "neckline", and depends more on the visual sentinel when generating words "graphic", "on", "the", "fit".

We extract image features using pre-trained ResNet-152[37], which implements residual learning units to alleviate the degradation of deep neural networks. ResNet-152 achieves the best accuracy among the Residual neural network (ResNet) family. The pretrained model is provided by Pytorch model zoo which achieves 21.69% top-1 error and 5.94% top-5 error on ImageNet[38]. We freeze the first 6 layers for fine-tuning and take the last convolutional layer as our visual features. We believe the features extracted retain both object and interaction information from the images.

Formally, let us denote the whole dataset as $D = \{(\mathbf{X}_i, \mathbf{y}_i\}_(i = 1, 2, ..., N)$ where X_i denotes the i^th image and $\mathbf{y}_i = \{y_1, y_2, ..., y_t\}$ denotes its caption label.

3.1 Adaptive Attention Architecture

3.1.1 Encoder-Decoder Structure

Let us first describe the generic encoder-decoder image captioning framework[7, 9]. The encoder-decoder model directly maximizes the following objective:

$$\theta^* = \underset{\theta}{argmax} \sum_{(\mathbf{X},\mathbf{y})} \log p(\mathbf{y}|\mathbf{X}, \theta) \tag{3.1}$$

where θ represents the parameters of the model. Then we decompose the log likelihood of the joint probability distribution into ordered conditionals using the chain rule:

$$\log p(\mathbf{y}) = \sum_{t=1}^{L} \log p(y_t|y_1, y_2, ...y_{t-1}, \mathbf{X}) \tag{3.2}$$

where the dependency on θ is dropped for convenience.

Usually Long-Short Term Memory (LSTM) plays the role of decoder and each conditional probability is modeled as:

$$\sum_{t=1}^{L} \log p(y_t|y_1, y_2, ...y_{t-1}, \mathbf{X}) = f(\mathbf{h}_t, \mathbf{c}_t) \tag{3.3}$$

where f is a nonlinear function that outputs the probability of y_t. c_t is the visual context vector at time step t extracted from image $\mathbf{X}$. $\mathbf{h}_t$ denotes the hidden state at t. For LSTM, $\mathbf{h}_t$ could be modeled as:

$$\mathbf{h}_t = LSTM(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{m}_{t-1}) \tag{3.4}$$

where $\mathbf{x}_t$ is the input feature map, $\mathbf{m}_{t-1}$ is the memory cell vector at time step $t-1$.

3.1.2 Spatial Attention Network

In attention-based model, $\mathbf{c}_t$ is dependent on both encoder and decoder, let us define it as:

$$\mathbf{c}_t = g(\mathbf{V}, \mathbf{h}_t) \tag{3.5}$$

where g denotes the attention function and $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_k\}, \mathbf{v}_i \in R^d$ is the visual features, each of which is a d dimensional vector corresponding to a region of the image.

Then the visual features $\mathbf{V} \in R^{d\times k}$ and the hidden state $\mathbf{h}_t \in R^k$ are fed into a single layer nerual network followed by a softmax function to generate the attention map over the k regions of the image:

$$\gamma_t = \omega_h^T tanh(\mathbf{W}_v\mathbf{V} + (\mathbf{W}_g\mathbf{h}_t)I^T) \tag{3.6}$$

$$\alpha_t = softmax(\gamma_t) \tag{3.7}$$

where $I^T) \in R^k$ is a vector with all elements equals 1. $\mathbf{W}_v, \mathbf{W}_g \in R^{d\times k}$ and $\omega_h^T \in R^k$ are weight parameters to be learnt. $\alpha \in R^k$ is the attention distribution over features on $\mathbf{V}$. Based on the attention weight, the context vector c_t can be calculated by:

$$\mathbf{c}_t = \sum_{i=1}^{k} \alpha_{ti}\upsilon_{ti} \tag{3.8}$$

where $\mathbf{c}_t$ combined with $\mathbf{h}_t$ is trained to predict the next word y_{t+1} as stated in Eq. (3.3). Here $\mathbf{c}_t$, which represents the attention map, is trained by the current hidden state instead of the previous hidden state vector[9]. The reason was explained in [34]. Similar to the superiority of residual network[37], the generated context features c_t could be considered as the residual information of the current hidden state, which reduces the uncertainty of the current state thus improve the performance of the network.

3.1.3 Adaptive Attention Architecture

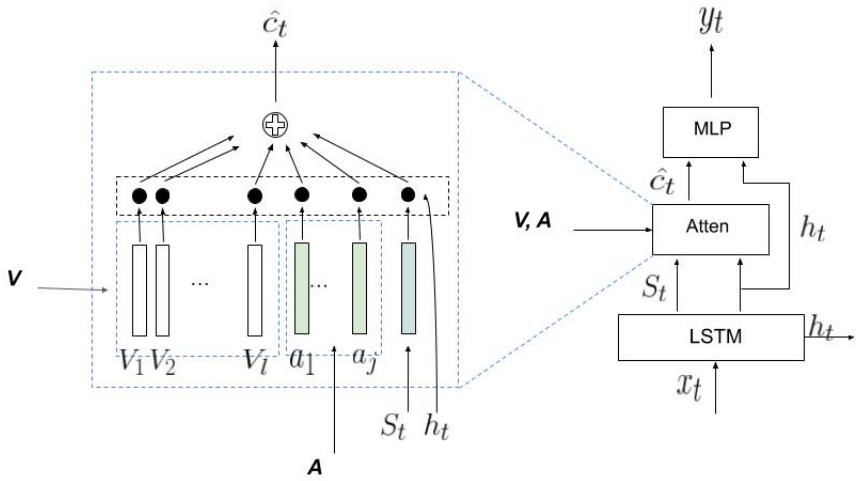

Figure 3.2: An illustration of the proposed model generating the t-th word y_t for the image. The input is the encoded image features V and attribute information A.

Adaptive attention architecture[34] (see Fig. 3.2) introduced the concept of "visual sentinel" which enables the network to know when to rely on visual signal and when to rely on the language model. It is a latent representation of existed information left in the decoder. With this, the attention architecture is able to determine whether it needs to attend the image to predict the next word.

The "visual sentinel" $\mathbf{s}_t$ can fall back on when it chooses to not attend to the image. As a extended component in LSTM, a sentinel gate accordingly decides whether to attend to the image or to the visual sentinel. This vector is

obtained by:

$$\mathbf{g}_t = \sigma(\mathbf{W}_x \mathbf{x}_t + \mathbf{W}_h \mathbf{h}_{t-1}) \tag{3.9}$$

$$\mathbf{s}_t = \mathbf{g}_t \odot \tanh(\mathbf{m}_t) \tag{3.10}$$

where $\mathbf{W}_x$ and $\mathbf{W}_h$ are weight parameters to be learnt, $\mathbf{x}_t$ and $\mathbf{h}_{t-1}$ are the input and hidden state at time $t-1$ and t respectively, $\mathbf{g}_t$ is the gate applied on the memory cell $\mathbf{m}_t$. $\odot$ denotes the element-wise product and σ denotes sigmoid activation function.

Based on the context vector $\mathbf{c}_t$ and the visual sentinel $\mathbf{s}_t$, the new adaptive context vector is modeled as:

$$\hat{\mathbf{c}}_t = \beta_t \mathbf{s}_t + (1 - \beta_t)\mathbf{c}_t \tag{3.11}$$

where β_t is the added sentinel gate at time t, which produces a scalar in the range [0; 1]. Thus, $\hat{\mathbf{c}}_t$ represents a mixture of the image features over spatial attention (i.e. context vector of spatial attention model) and previous knowledge known by the decoder (i.e. the visual sentinel). A β_t value of 0 implies only spatial image information is used and 0 means only spatial image information is processed for predicting the next word.

To compute the new sentinel gate parameter β_t, Eq. (3.6) need to be modified. An additional element is added to γ which indicates how much attention the decoder put on the sentinel rather than the visual image features. Therefore, Eq. (3.7) is converted to:

$$\hat{\alpha}_t = ([\gamma_t; \mathbf{w}_h^T \tanh \mathbf{W}_s \mathbf{s}_t + \mathbf{W}_g \mathbf{h}_t]) \tag{3.12}$$

where $[;]$ means concatenation operation, $\mathbf{W}_s$ and $\mathbf{W}_g$ are weight parameters to be learnt. It is noted that $\mathbf{W}_g$ is the identical parameter as in Eq. (3.6). To summarize, $\hat{\alpha}_t \in R^{k+1}$ represents the overall attention distribution over both the spatial image attention as well as the sentinel vector, where the first k elements is the spatial image attention and the last element is interpreted to be the sentinel gate β_t.

The probability over the vocabulary at time step t can be computed as:

$$\mathbf{p}_t = (\mathbf{W}_p(\hat{\mathbf{c}}_t + \mathbf{h}_t)) \tag{3.13}$$

where $\mathbf{W}_p$ is weight parameter to be learnt.

As stated above, this adaptive attention architecture can dynamically attend to the image or the visual sentinel when generating the next word. The sentinel

vector is updated at every step thus it could be considered as what the decoder already knows.

3.2 Attribute Information

The idea of feeding attributes to the LSTM decoder comes from the following concept. Firstly, words in the captions must contain inherently salient information. Thus it is feasible for a text attribute extractor to extract and highlight this kind of information directly from the image. These words, which are called "text attributes" could be a variety of word types, such as nouns, verbs, and adjectives. Secondly, feeding these attributes help disambiguate noisy visual detection. Captions may contain a number of abstract and ambiguous concept which increase the uncertainty when training the language model. However, by learning a multi-modal detector over both images and their captions, the global similarity between images and text is possible to measure and the majority of the suitable description for the image is easier to learn.

The first step in this part is to detect a set of words that are possible to attend in the image's description. These words may belong to the following parts of speech, including nouns, verbs, and adjectives. We build the vocabulary V using the 1000 most common words of the training captions.

3.2.1 Text Attribute Extractor

Given a vocabulary of attributes, the next step is to detect these words from images. We train the text attribute extractors using a CNN (convolutional neural network) based model. Fig. 3.3 shows an overview of the model. An image passes the pre-trained VGG-16 model and we express the Cov5 layer as the input feature map, which is fed into a 2-layer CNN following with one fully connected layer. The possibility p_i^w of image x_i containing word w is computed by a sigmoid layer:

$$p_t^w = \frac{1}{1 + \exp(-(v_t^w \phi(b_i) + u_w))} \tag{3.14}$$

where $\phi(b_i)$ is the fully connected representation for image b_i, v_t^w and u_w are the associated weights and bias with word w.

Due to the highly imbalanced ratio of the positive labels (5 words per image) vs. negative. The loss function used for training the detector is

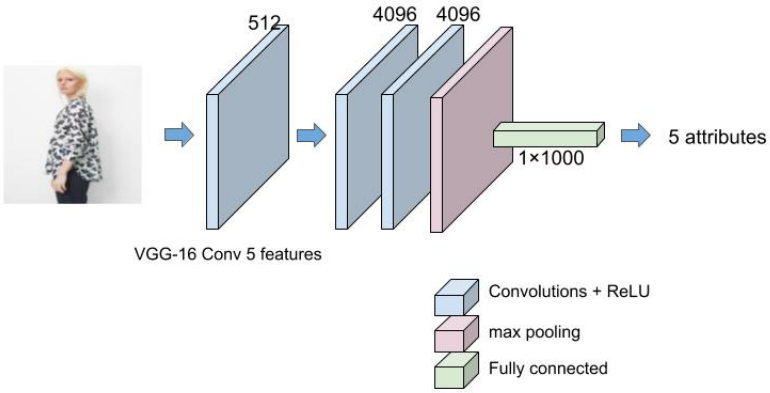

Figure 3.3: An illustration of the text attribute extractor. We extract features from every image from the "Cov5" layer of the VGG-16 network model. After passing two layers of CNN, one max pooling layer, and one fully connected layer, the output is 5 “attribute” word in a vocabulary of 1000 words.

accordingly modeled as:

$$L_i^C = -\beta_p(p(x_i)\log(q_i)) + \beta_n(1-p(x_i))(\log(1-q(x_i))) \tag{3.15}$$

where β_p and β_n are class weights assigned for giving higher penalty over false positive predictions. Here we set $\beta_p = 100\beta_n$.

3.2.2 Attribute-combined Model

With injecting the high-level attributes into the decoder, we propose our attribute-combined model (see Fig. 3.1). Let us say that in the adaptive attention architecture section whether visual sentinel or spatial image features are extracted from the image, which two we classify as visual information. In our model, the decoder is modified by additionally integrating visual information and high-level attributes. More specifically, as Fig. 3.4 shows, attribute representations are fed into LSTM as an additional input at each time step when generating words to emphasize the high-level attributes continuously. Accordingly, given the attribute representation $\mathbf{A}$ the calculation

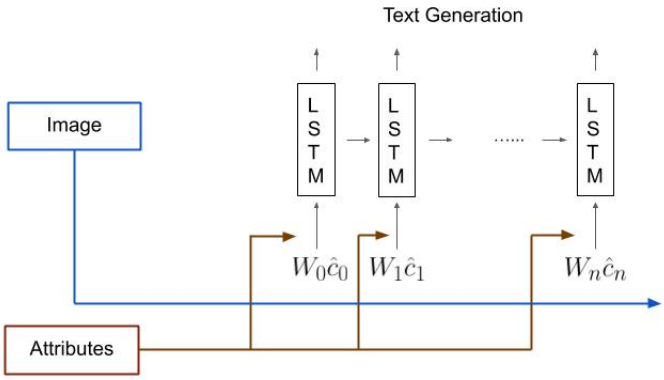

Figure 3.4: A simplified diagram of our attribute-image-combined network.

of the hidden state in each time step is converted from Eq. (3.4) to:

$$\mathbf{h}_t = LSTM(\mathbf{x}_t, \mathbf{A}, \mathbf{h}_{t-1}, \mathbf{m}_{t-1}) \tag{3.16}$$

Chapter 4

Experiments

4.1 Dataset and Preprocessing

We conduct experiments on two image captioning datasets, where one contains general images and one consists of only fashion images.

4.1.1 MS COCO Dataset

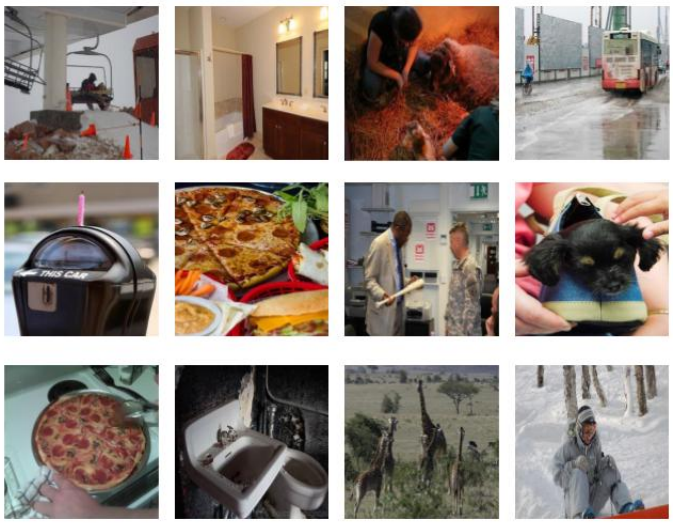

Figure 4.1: Some data examples of the MS COCO dataset.

The MS COCO dataset[39] contains 328k images with a total of 2.5 million instances. It is a very large and popular dataset for image recognition,

segmentation, and captioning. There are various features of the COCO dataset such as recognition in context, object segmentation, and multiple objects per class. The images belong to 80 object categories, such as “person”, “book”, “bowl” etc. For the image captioning task, five written caption descriptions labeled for each image are used as ground truth. Fig. 4.1 gives some examples of images in MS COCO dataset.

4.1.2 Fashion Dataset

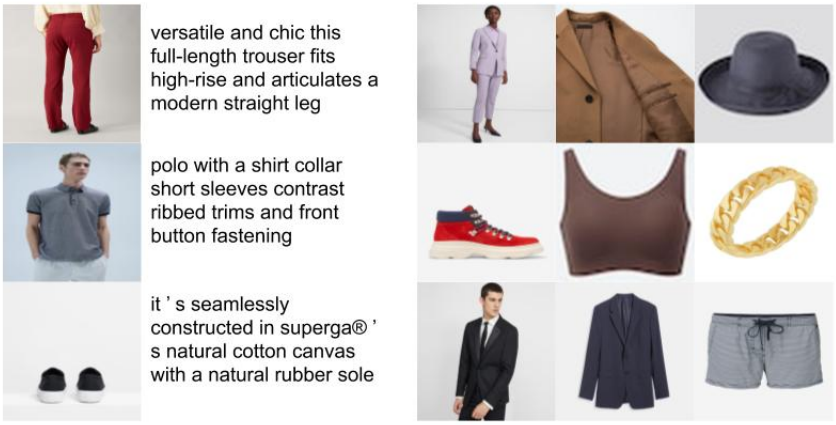

Figure 4.2: Some data examples of the Fashion dataset. The left part demonstrates three images with their corresponding caption and the right part only shows the images.

The fashion dataset is provided by Norna AI, which contains 1,511,916 images from 194,453 fashion products. Each data includes a description with various sentence amount and word length provided by its vendor. The dataset contains different categories of fashion products, including clothing, jewelry, purse, etc. These images are from different vendors, Uniqlo, Toteme-Studio, Bestseller, Drykorn, Jlindeberg, Joseph-fashion, Marc-o-polo, Rodebjer, Tigerofsweden, Vince respectively.

However, not all the data is fitting for the image captioning task. A number of data do not contain a valid caption or the caption is not related to the image content. Thus we apply some filters to select the qualified images for this project, including 144,422 images from 24,649 products. The details of

the preprocessing methods are explained in 4.1.3 Each data is labeled with only one description sentence as its caption. Couples of products are allowed to map to one same description and there are 10,091 unique captions in the fashion data. Fig. 4.2 reports some examples of fashion dataset.

4.1.3 Preprocessing

Three main filters are employed for selecting qualified caption in the fashion dataset fitting in this project, including sentiment filter, part-of-speech filter, and other filters. The details of these filters are elaborated as follows,

- **Sentiment filter**. The raw descriptions are given by vendors majorly contain advertising content. For example, *Outstanding stretch in all directions, great for kids who toss and turn in their sleep*. We clarify that those sentences are mostly exaggeration, which is normally with strong emotion. Thus, sentiment filter is employed to detect the sentiment in the description, keeping the description in neutral emotion and excluding the affective content, especially the advertising content.
  Towards this end, "SentimentIntensityAnalyzer" from "nltk" is employed to calculate the subjectivity score. A sentence is regarded as advertising content when its score exceeds the threshold th, where th is set as 0.5.

- **Part-of-speech filter**. Some sentences of the raw description are uncompleted, such as *Machine wash cold, gentle cycle*. In order to ensure the target label grammaticality correct and semantically meaningful, we apply a part-of-speech filter over the raw labels, where only the descriptions which are complete sentences are selected. Here we consider a sentence completed when it contains both nouns and verbs.

- **Other filters**. These filters exclude the descriptions about the models, item size, returning policy, and washing instructions. More specifically, sentences including "size:", "washing" or "returning" is labeled as irrelevant content and be excluded.

Passing these filters, we select the longest sentence from the remaining as the caption label for an image. If none of the sentences remains, we skip this data. The keep rate of data after preprocessing methods is 74.2%.

To train the attribute detector, we automatically generate 5 attributes for each image with the following method. Firstly we build an attribute vocabulary

comprising 1000 most common words (only nouns and adjectives) from the whole caption text. We choose 5 words in the caption which occur in the vocabulary as attributes for each data. If more than 5 words attend, the top 5 words ranking by frequency in the whole caption text is selected. On the other hand, random words from the vocabulary are added if less than 5 words attend.

We apply the same split for the two datasets, with 70% of the data used for training, 15% as the validation set, and 15% for testing. As a preprocessing step, all the images used in experiments are resized to the size of 224×224 with the method of bilinear interpolation. We also create two variations for the fashion dataset. The one-vendor condition focuses on the largest product vendor, bestseller. This subset contains 89,756 images from 19,385 products. The amount of unique captions is 8,448. The second condition employs all images in the dataset. The reason behind it is that we found the same vendor usually describe their product in similar text form and style. The fashion data is forbidden to be published by Norna AI but we publish the source code over the COCO dataset at https://github.com/guoyuntu/Image-Captioning-On-General-Data-And-Fashion-Data.

4.2 Hyperparameter Settings

4.2.1 text attribute extractor

The convolution layer of the text attribute extractor has kernel size 5×5 and stride 1×1. The max pooling layer has hernel size 8×8 and stride 0. We train the text attribute extractor for 10 epochs.

4.2.2 AIC-AB Network

In the decoding stage, words in captions are embedded into 255-dimension vectors and words of attributes are embedded into 51-dimension vectors using the default word embedding function provided by PyTorch. The hidden size is set as 512. Adam optimizer with learning rate decay is employed to train the model. The parameters is set as: $\alpha = 0.8$, $\beta = 0.999$, $learning_rate = 4e-4$. The decay of learning rate is modeled as:

$$l_r^{E+1} = l_r^E * 0.5^{\frac{E-20}{50}}, E > 20 \tag{4.1}$$

where l_r^e denotes the learning rate in epoch E. We train the text attribute extractor for 50 epochs.

4.3 Baselines and Ablation study

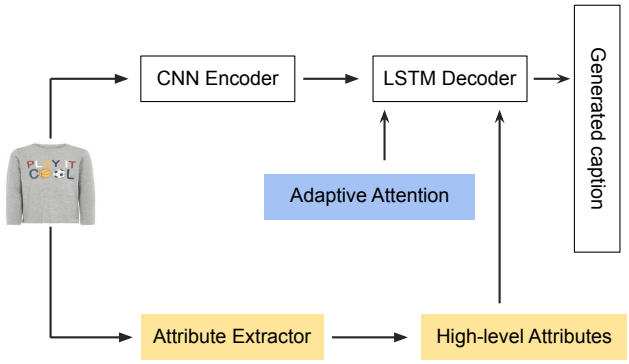

Figure 4.3: A block diagram of AIC-AB Net. Here the white blocks denote the Naive-ED network, with green blocks denote Atten-Only, and with blue blocks denote Attr-Only.

As Fig. 4.3 shows, our model is an encoder-decoder based network composed of two components, adaptive attention architecture and attribute information. Thus, it is compared to the following baselines and ablated models.

The Naive Encoder-Decoder (Naive-ED) In this version, the adaptive attention architecture and attribute information are removed. What remains is a CNN-based encoder and LSTM-based decoder network. This network is basically **NIC**[7] baseline model but with different pre-trained input and different hyper-parameter setups.

The Adaptive Attention Only (Atten-Only) As a second ablated network, we remove the attribute information from the decoding stage. The adaptive attention architecture including spatial attention and visual sentinel are kept intact. This network is the same as the state-of-the-art method among the attention-based approaches - **Adaptive**[34].

The Attribute Combined Only (Attr-Only) This is a third ablated network. We feed the attribute information into the LSTM decoder but remove the adaptive attention architecture.

Chapter 5

Results and Analysis

5.1 Attribute Detector Analysis

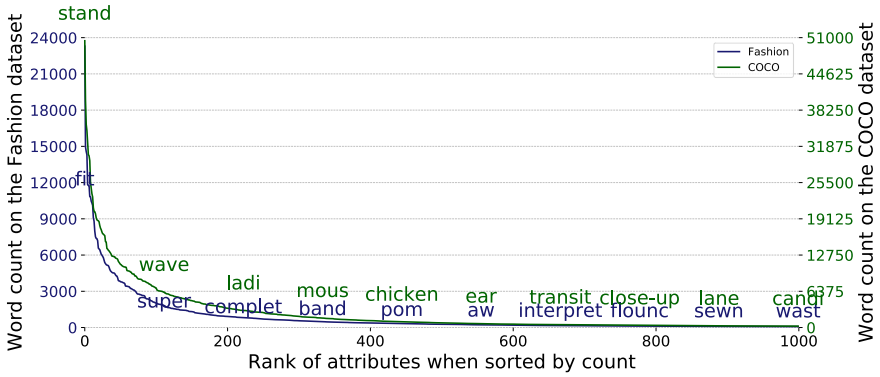

Figure 5.1: The attributes count distribution of the COCO dataset and the Fashion dataset (all vendors).

The performance of the text attribute extractor would significantly affect the quality of attribute information. We report the word count distribution of the 1000 attribute vocabulary on the COCO dataset and the Fashion dataset (Bestseller) as shown in Fig. 5.1. We observe that the distribution of the Fashion dataset is way more unbalanced than it of the COCO dataset. It indicates that the corpus of the captions in the Fashion dataset contains less semantic information than it in the COCO dataset.

Furthermore, we measure its accuracy on the word classification task. It is worth noting that to avoid the situation that conceptually similar words are classified separately, we apply a stemmer when building the vocabulary. For

example, the words *cats/catlike/catty* is identified to the stem *cat*. However, the stemmer would not work on the words which have a different part of speech. For example, it does not reduce the words *fishing/fisher* to the stem *fish*. We use the "Snowball ' Stemmer[40] provided by "nltk".

Table 5.1: Metrics for words with different parts of speech (NN: Nouns, JJ: Adjectives,). Results are shown using a random classifier and our detector.

		MS COCO			Fashion (Bestseller)			Fashion (all vendors)		
		NN	JJ	All	NN	JJ	All	NN	JJ	All
	count	952	48	1000	941	59	1000	937	63	1000
Recall	Random	0.002	0.005	0.002	0.002	0.003	0.003	0.002	0.003	0.002
	Detector	0.135	0.053	0.127	0.099	0.110	0.111	0.089	0.102	0.095
Precision	Random	0.002	0.006	0.003	0.002	0.003	0.003	0.003	0.004	0.002
	Detector	0.219	0.105	0.177	0.152	0.165	0.175	0.148	0.169	0.161
F2 score	Random	0.002	0.006	0.003	0.002	0.003	0.003	0.002	0.004	0.002
	Detector	0.143	0.059	0.138	0.119	0.124	0.120	0.098	0.113	0.101

Average recall, precision and F2 scores for different parts of speech are reported in Tab. 5.1. We also report one baseline, random, which is the result of randomly labeling each image with five attribute words. We observe that our attribute detector remarkably improves over the random classification approach for all parts of speech. Interestingly, according to the complexity of the datasets, the COCO dataset is the most complicated and the Fashion dataset (Bestseller) is the least. However, the F2 score results show the accuracy on the Fashion dataset beats it on the COCO dataset. The reason is that as claimed above, the Fashion dataset is less semantic informative than the COCO dataset. Accordingly, it contains more less-frequently attending words, which are more difficult to be detected by our model. These words pull the average accuracy down. Note that on the Fashion dataset the F2 score of adjectives is higher than that of nouns, by 0.005 and 0.015 respectively, while on the COCO dataset the result is opposite. We argue that that is caused by the visual information that adjectives could provide. On a general dataset, such as COCO, an adjective is probably less visual informative or hard to detect (e.g., *few*, which has an F2 score of 0.03). However, on the fashion dataset, an adjective is very likely to be either visually informative (e.g., *blue*, which has an F2 of 0.65) or associated

with specific objects (e.g., *printed*, which has an F2 of 0.48).

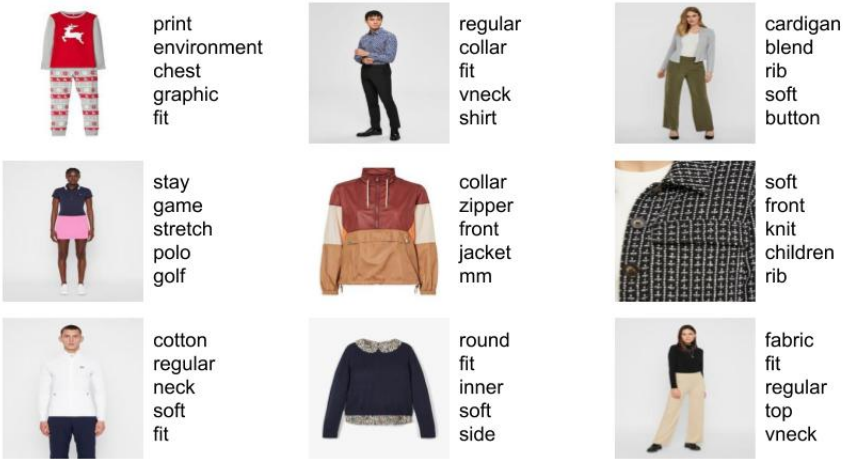

Figure 5.2: Some examples of the attribute detector results on the fashion dataset (Bestseller)

Fig. 5.2 demonstrates the attribute results and the input images for the fashion dataset (Bestseller). The three columns are all successful examples. We see that our text attribute extractor is able to extract and summarize the key attributes from the input.

5.2 Captioning Accuracy Analysis

Table 5.2: Image Captioning on MS COCO results.

Model	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
Naive-ED	0.729	0.556	0.409	0.299	0.249	0.531	0.952
Atten-Only	**0.743**	**0.572**	0.424	0.313	0.265	0.546	1.088
Attr-Only	0.671	0.495	0.366	0.276	0.255	0.535	1.059
Ours-AICAB	0.730	0.554	**0.424**	**0.339**	**0.279**	**0.550**	**1.105**

We perform image captioning on the MS COCO dataset and the fashion dataset, where the fashion dataset has two experimental conditions with different numbers of vendor sources. Tab. 5.2, Tab. 5.3, Tab. 5.4 reports the

Table 5.3: Image Captioning on Fashion (Bestsellers) results.

Model	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
Naive-ED	0.345	0.284	0.251	0.231	0.173	0.316	1.870
Atten-Only	0.365	0.306	0.275	0.255	0.185	0.334	2.094
Attr-Only	0.350	0.290	0.258	0.240	0.179	0.321	1.988
Ours-AICAB	**0.385**	**0.316**	**0.289**	**0.280**	**0.190**	**0.349**	**2.189**

Table 5.4: Image Captioning on Fashion (All vendors) results.

Model	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
Naive-ED	0.264	0.216	0.179	0.165	0.115	0.238	1.149
Atten-Only	0.276	0.218	0.193	0.178	0.121	0.256	1.202
Attr-Only	0.268	0.210	0.184	0.169	0.117	0.242	1.155
Our-AICAB	**0.290**	**0.231**	**0.202**	**0.191**	**0.125**	**0.268**	**1.297**

results on these three dataset respectively, where B-n is BLEU score that uses up to n-grams. Higher is better in all columns.

We observe that our AIC-AB model achieves the best performance among all ablated versions. The ablation study reveals the complementarity of all constituents of AIC-AB Net. The naive encoder-decoder network underperforms AIC-AB Net by 0.143, 0.319, and 0.148 (CIDEr score). The adaptive attention network underperforms by 0.017, 0.095, and 0.095 (CIDEr score). The attributes-combined model underperforms by 0.107, 0.201, and 0.142 (CIDEr score). Note that adaptive attention architecture improves the performance better than the attribute information. These results indicate that two components indeed complement each other and their co-existence crucially benefits the caption generation.

The three experimental conditions establish a comprehensive spectrum. The general image dataset, MS COCO dataset is the most complex and contains multi objects in each image, for which the CIDEr score is the lowest across the datasets. It is noteworthy that here we only compare the CIDEr score for the reason that it is the only metric that keeps the stable scale when the number of captions varies. The fashion dataset contains one single object per image. The Bestseller condition is simpler than the all-vendor condition. Although the effectiveness of our network is the same obvious, the gap widens as the task gets more complicated.

On the COCO dataset, although the attributes-combined model obtains a similar CIDEr score compared with the adaptive attention model, it observably

underperforms on other scores. As stated in 1.7, CIDEr focuses more on semantical correctness while others more reflect grammaticality correctness. These results indicate that attribute information provides significant semantic information, however, in order to demonstrate these attributes in the generated captions the model achieves this at the expense of grammaticality correctness. Interestingly, it does not happen on the Fashion dataset, we argue that this is because of the small amount of captions. The sentence pattern is easier to recognize on the Fashion dataset. However, this effect is not shown in our AIC-AB network. It reveals the attention architecture especially the sentinel gate corrects the bias brought by the attributes. The two components indeed complement each other.

On the fashion dataset, we observe that our model achieves better performance on the Bestseller condition than the all-vendor condition, with an improvement of 0.892 (CIDEr score). This is opposite to the regular pattern that the increased size of data improves the performance of the machine learning model. The detailed reason is explored in Section 5.5.

5.3 Attention Distribution Analysis

To better understand our model, we also visualize the image attention distributions α for the generated caption. We simply sample the attention map to the image size (224×224) using bilinear interpolation and pyramid expanding. Fig. 5.3 shows the generated captions and the image attention distribution for specific words in the caption. The first 5 rows are success cases, and the last row shows a failure example. We see that our model learns to paying attention to the specific region when generating different words in the caption, which correspond strongly with human intuition. Note that on the failure case although our model fails to focus on the region of the sleeves when generating “sleeves”, it still successfully recognizes the position of the printed stripe.

Since the COCO dataset provides the ground truth of objects’ bounding box, it is able to evaluate the performance of attention map generation. To the best of our knowledge, the spatial intersection over union (sIOU) score is used as the metric to measure the localization accuracy. Give the word w_t and its corresponding attention map α_t, we first segment the regions of the image with its attention value larger than a pre-class threshold th (after the map is standard-normalized to scale [0,1), where we set as 0.6). Then we take the bounding box that covers the largest connected component in this segmentation map as the predicted attention region. We compute the

Figure 5.3: Visualization of generated captions and attention distribution maps on the fashion dataset (Bestseller). Pairs of masked regions and underlined words are tagged by different colors . The first 5 rows are success cases, the last row is a failure example.

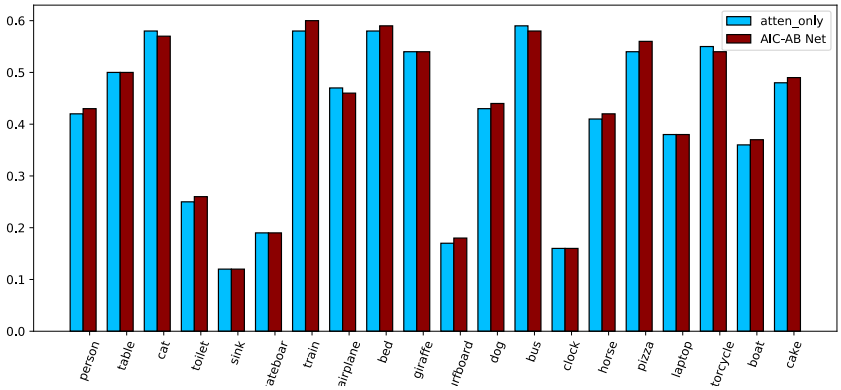

Figure 5.4: Localization accuracy over generated captions for top 20 most frequent COCO object categories. "atten_only" and "AIC-AB Net" are one of the ablated versions and our model, respectively.

sIOU between the predicted bounding box and the ground truth for the top 20 most frequent COCO object categories as Fig. 5.4 shows. The average localization accuracy for the "atten_only" is 0.415, and 0.419 for our AIC-AB Net. This implies that as a combined model, the attribute information benefits the attention map generation. We also observe that our AIC-AB network and its attention only version have a similar trend. They both perform well on visual informative objects and large objects such as "cat", "train", "bed" and "bus", and have poor performance on small objects such as "sink" and "clock". We argue that is because our attention map is extracted from 7×7 spatial map, which loses plenty of resolution and detail. This defect is remarkably exposed when detecting small objects. This reason can explain the wrong attention map on the Fashion dataset as well, where the majority of words are the description of details and refer to small regions on the image.

Since the bounding box ground truth is missing in the Fashion dataset, we apply statistic analysis for 5 typical words as quantitative analysis, *hood*, *cap*, *pants*, *dresses* and *sleeves*. On common cases, *hood* and *cap* only show on the upper part of an image, *pants* and *dresses* only on the lower part and *sleeves* only on left and right sides. We assume these regions are their ground truth respectively and apply the same approach as explained above to measure the localization accuracy. Fig. 5.5 reports the result. We observe that AIC-AB model outperforms on the first 4 words than word *sleeves* and shows a similar trend with the adaptive attention model.

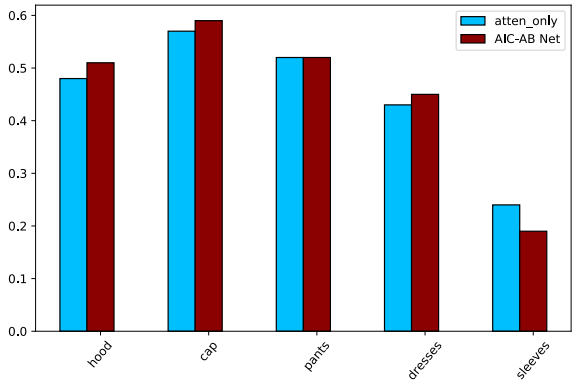

Figure 5.5: Localization accuracy over generated captions for three typical words. “atten_only” and “AIC-AB Net” are one of the ablated versions and our model, respectively.

5.4 Sentinel Gate Analysis

In order to evaluate the performance of the sentinel gate, whether it helps separate visual informative words from non-visual words such as stop words. We use $p = 1 - \beta$, which is the information from spatial attention as visual grounding probability. For each word of the vocabulary, we calculate the average visual grounding probability over all the generated captions which contain that word. Fig. 5.6 demonstrates the visualization of the top 20 words and the lowest 20 words on the COCO dataset and Fig. 5.7 shows it on the Fashion dataset (Bestseller).

We find that our model attends to the image more when generating nouns referring to specific objects such as “boys”, “ball”, “cat” and “suit”, attribute words like “short” and “plastic”. When the word is non-visual, our model learns to not attend to the image such as for “any”, “delta”, “opt” etc. Our model also learns to attend abstract words less than the visual words, such as “series” and “add”. Note that for small-object words like “mosquito” and “gold” our model leans to attend less spatial features but rely more on sentinel information.

On the fashion dataset (Bestseller), how likely a word to be visually grounded is more related to the feature of the dataset. For example, a word with rich visual information in general cases such as “gym”, “runway” would not be directly shown in the fashion images. Oppositely, they have a high correlation with other words hence chooses to not rely on the visual features.

For example, “gym” occurs more when describing sport wears. The language correlation on fashion captions differs a lot from the general captions, and it significantly affects the visual grounding probability.

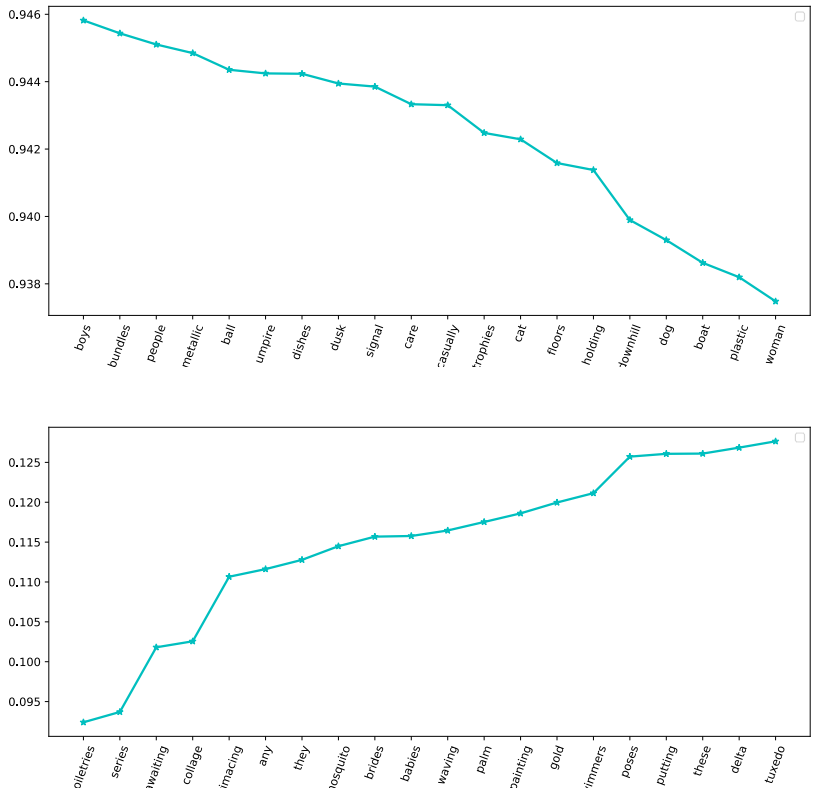

Figure 5.6: Rank-probability plots on COCO indicating how the visual grounding probability of a word when it is generated in a caption, where the top figure is the words with highest probability and the bottom figure is the lowest.

5.5 Transfer Learning Analysis

To gain a better understanding of the fashion dataset, we perform an additional experiment on transfer learning, where the data used is fashion images from the different vendors. AIC-AB Net is firstly trained on the total out-of-domain training set then fine-tuned for 10 epochs using only 20% of the in-domain training set. The comparison set is to training AIC-AB Net from scratch using the same 20% in-domain training set. Other experiment settings, including

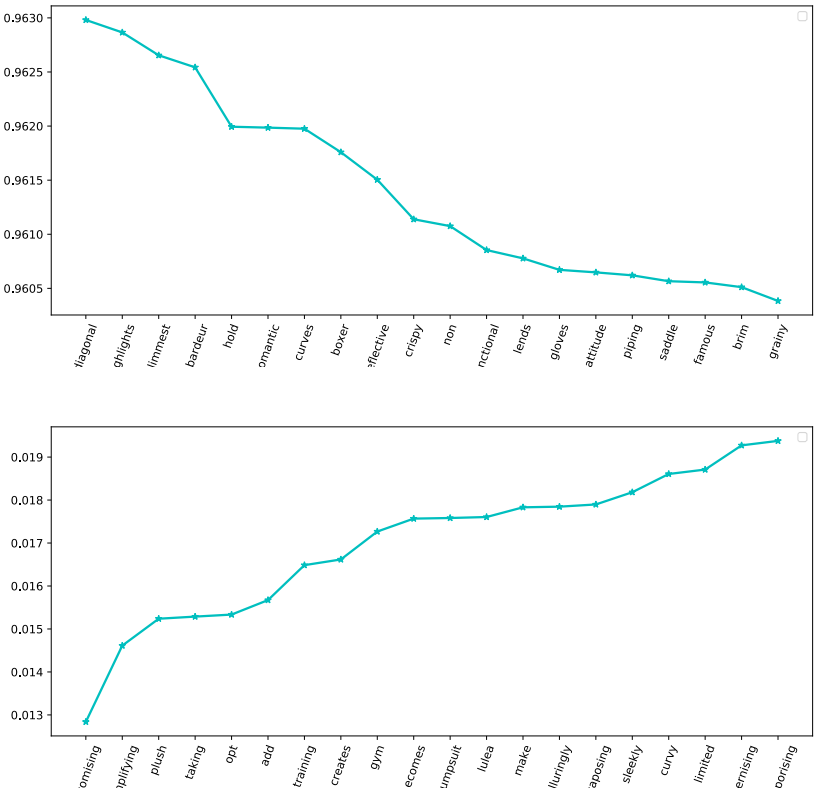

Figure 5.7: Rank-probability plots on Fashion (Bestseller) indicating how the visual grounding probability of a word when it is generated in a caption, where the top figure is the words with highest probability and the bottom figure is the lowest.

Table 5.5: Transfer Learning: Out-of-domain AIC-AB Net Fine-tuned On 20% In-domain data VS. training from scratch on 20% In-domain.

Test Set	Training Procedure	B-4	METEOR	ROUGE-L	CIDEr
Bestseller	Uniqlo → 20% Bestseller	0.013	0.018	0.066	0.032
	20% Bestseller only	**0.076**	**0.073**	**0.151**	**0.610**
Uniqlo	Bestseller → 20% Uniqlo	0.017	0.033	0.096	0.118
	20% Uniqlo only	**0.026**	**0.065**	**0.152**	**0.437**

learning rates, are kept identical to the image captioning experiments.

The results, shown in Tab. 5.5, indicates that transfer learning in the low-data setting is disadvantageous. Either Bestseller → Uniqlo or Uniqlo → Bestseller underperforms a lot compared to their corresponding control set. Vendors describe their products in different styles and formats. This would significantly influence the caption generation of our model. This suggests that the worse performance of AIC-AB Net on the full fashion dataset compared with it on the fashion dataset (Bestseller), with a drop of 0.892 (CIDEr Score) is not caused by the increase of size. The real cause is the distinct description styles.

Chapter 6

Conclusions and Future work

The conclusions from this project are presented in this chapter in Section 6.1. Furthermore, the limitations of this project and consequent potential future work are presented and discussed in Sections 6.2 and 6.3 respectively. Finally, Section 6.4 gives the benefits and ethical considerations of this project.

6.1 Conclusions

In the line of the purpose of the project introduced in Section 1.3, this work has fulfilled the goals stated in Section 1.4. Using the methodology and evaluation explained in Chapter 3, experiments conducted on datasets described in Section 4.1 and ablation study explained in Section 4.3 were carried out. Their corresponding results and analysis are showed and discussed in Chapter 5. These results corroborate the main hypothesis of this project, i.e. the viability and superiority of adding attribute information to adaptive attention assembled encoder-decoder network (AIC-AB Net).

Motivated by the task to generating captions for single-object fashion images and inspired by the adaptive attention architecture[34] and semantic concept[41], the purpose of this project is to first apply the deep learning model on general/fashion images and second test whether our attribute-image-combined attention-based network achieving remarkable performance on the image captioning task. The main contributions of this work are:

- We propose a fashion dataset that is cleaned from the raw data provided by Norna AI. It contains 144,422 images from 24,649 products, which has the following features: 1) each image contains one single fashion product with one descriptive sentence. 2) the captions include a satisfactory amount of adjectives and nouns.

- We suggest that the ability to locate the relevant region of an image when generating different words and the combination with the attribute information is crucial for accurate caption generation. Toward this end, we present an encoder-decoder network with adaptive attention architecture and attribute information, called Attribute-Image-Combined Attention-Based Network (AIC-AB Net). The adaptive attention architecture is composed of spatial attention and visual sentinel, the former determines where to "look" and the latter decides when to "look". The attribute information measures the global similarity between images and text and helps disambiguate noisy visual detection.

We give the answers to the research question introduced in Section 1.2 as following:

- *Will deep learning models achieve consistent performance on multi-object images (COCO dataset) and single-object images (Fashion dataset)?*

  Yes, it does. According to our image captioning experiments, the superiority of attention architecture and attribute information is demonstrated on both datasets. The comparison of their results indicates that the complexity of images increases the difficulty to train a model and undermines the overall performance.

- *How does an attention-based architecture and attribute information added on influence the model performance on image captioning tasks?*

  Towards this end, we propose our AIC-AB Net, which achieves the best performance on all the three datasets. The ablation study shows whether the attention-based architecture or the attribute information has a positive influence on an Encoder-Decoder based deep learning model.

6.2 Limitations

With the labels given by web scratch from product vendors, which lacks a uniform standard and format, particularly descriptions from different vendors, the explorations and experiments carried out on the Fashion dataset are chiefly to test out the performance of AIC-AB Net on single-object images. However,

all the potentials or issues are not fully explored. It shows limitations and under-performance on transfer learning. We assume that the reason behind is the huge gaps in captions from one vendor to another. The robustness of our network still remains unexplored and the question is left for further research when the fashion dataset obtains up-to-standard labels by human annotators.

From the point of attributes detector, the model employed is not art-of-the-state. This is mainly due to the time limitation of this project. Although the results already reflect the improvements brought by attribute information. We assume that a more advanced network such as noisy-OR Multiple Instance Learning [42] predicts more accurate attributes and furthermore achieves better performance on caption generation.

Furthermore, according to the failure cases of attention distribution on the Fashion dataset. It is founded that AIC-AB model fails to map to the right region when generating some words, such as “sleeves” and “waist”. The attention intends to stay on the whole body instead of moving frequently and evidently on images. The results show the satisfactory improvement brought from the attention architecture. Therefore, we argue that further optimization could be conducted on the attention architecture.

6.3 Future Work

The field of single-object image captioning being still left explored, there is still plenty of room to improve the quality of either datasets or models. This project provides some inspiration for the implementation of attention architecture and attributions information on an Encoder-Decoder based model.

However, as discussed in Section 6.2, a couple of issues and potential solutions are raised. Future research could explore the following three directions. First, create an up-to-standard labeling system for the Fashion dataset, which benefits the consistency of the data and the robustness of the models trained on it. Second, improve the attribute detector with more productive classification methods. The accuracy of attributes extracted directly affects the final performance of AIC-AB Network. Third, since in a number of cases our model is not able to pay attention to the accurate region when generating words, one suggestion is to segment the images into more regions. For instance, in this project images are segmented into 50 regions, further research could explore if the increase in the number of regions would improve the performance.

6.4 Final Words

Image captioning is a challenging and promising task for the Internet industry and computer vision. An accurate text generation from image information can enable many interesting applications such as automatically describing fashion products. We believe this work represents a significant step in improving image captioning and breeds useful applications in other domains.

References

[1] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, p. 971–987, Jul. 2002. doi: 10.1109/TPAMI.2002.1017623. [Online]. Available: https://doi.org/10.1109/TPAMI.2002.1017623

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[4] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2422–2431.

[5] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 595–603. [Online]. Available: http://proceedings.mlr.press/v32/kiros14.html

[6] J. Johnson, A. Karpathy, and F. Li, "Densecap: Fully convolutional localization networks for dense captioning," *CoRR*, vol. abs/1511.07571, 2015. [Online]. Available: http://arxiv.org/abs/1511.07571

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *CoRR*, vol. abs/1411.4555, 2014. [Online]. Available: http://arxiv.org/abs/1411.4555

[8] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1473–1482.

[9] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, p. 2048–2057.

[10] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664–676, 2017.

[11] C. Gan, Z. Gan, X. He, J. Gao, and l. Deng, "Stylenet: Generating attractive visual captions with styles," 07 2017. doi: 10.1109/CVPR.2017.108 pp. 955–964.

[12] J. Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*. USA: W. H. Freeman Co., 1976. ISBN 0716704641

[13] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," 10 2002. doi: 10.3115/1073083.1073135

[14] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://www.aclweb.org/anthology/W05-0909

[15] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[16] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[17] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, Feb. 2019. doi: 10.1145/3295748. [Online]. Available: https://doi.org/10.1145/3295748

[18] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 15–29.

[19] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 529–545.

[20] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," *CoRR*, vol. abs/1410.1090, 2014. [Online]. Available: http://arxiv.org/abs/1410.1090

[21] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K. Müller, Eds. MIT Press, 2000, pp. 1008–1014. [Online]. Available: http://papers.nips.cc/paper/1786-actor-critic-algorithms.pdf

[22] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," 10 2017. doi: 10.1109/ICCV.2017.323 pp. 2989–2998.

[23] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 1137–1155, Mar. 2003.

[24] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: Association for Computing Machinery, 2007.

doi: 10.1145/1273496.1273577. ISBN 9781595937933 p. 641–648. [Online]. Available: https://doi.org/10.1145/1273496.1273577

[25] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," 01 2013, pp. 1–12.

[26] J. Gu, G. Wang, J. Cai, and T. Chen, "An empirical study of language cnn for image captioning," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1231–1240.

[27] C. Liu, J. Mao, F. Sha, and A. L. Yuille, "Attention correctness in neural image captioning," *ArXiv*, vol. abs/1605.09553, 2017.

[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[29] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[31] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional lstms," in *Proceedings of the 24th ACM International Conference on Multimedia*, ser. MM '16. New York, NY, USA: Association for Computing Machinery, 2016. doi: 10.1145/2964284.2964299. ISBN 9781450336031 p. 988–997. [Online]. Available: https://doi.org/10.1145/2964284.2964299

[32] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding long-short term memory for image caption generation," 09 2015. doi: 10.1109/ICCV.2015.277

[33] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1251–1259.

[34] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *2017*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3242–3250.

[35] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of Bleu in machine translation research," in *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics, Apr. 2006. [Online]. Available: https://www.aclweb.org/anthology/E06-1032

[36] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, Jul. 2004. doi: 10.3115/1218955.1219032 pp. 605–612. [Online]. Available: https://www.aclweb.org/anthology/P04-1077

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[38] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft coco: Common objects in context," 05 2014.

[40] M. F. Porter, "Snowball: A language for stemming algorithms," Published online, October 2001, accessed 11.03.2008, 15.00h. [Online]. Available: http://snowball.tartarus.org/texts/introduction.html

[41] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," 11 2016.

[42] P. Viola and J. Platt, "Multiple instance boosting for object detection." vol. 18, 01 2005.

Appendix A

First Appendix

Here we report the hardware, software and libraries used in the project.

The experiments in this report were run on a machine with an Intel i7-8700K with 12 cores (2 threads per core). The GPU used is an Nvidia GeForce RTX 2080, with around 12 GB usable GPU memory.

The experiments are conducted in Python 3.7 which uses both the CPU and GPU resources for training, with emphasis on using the GPU for model learning.

We also list the libraries used in this project for training models and generating results. The libraries used are as following: *os*, *json*, *collections*, *random*, *tqdm*, *matplotlib*, *pickle*, *cv2*, *argparse*, *PyTorch*, *numpy*, *math*, *glob*, *numpy*.

Appendix B

Second Appendix

Here we report the attributes vocabulary (after stemming) of the COCO dataset and the fashion dataset (all vendors).

COCO dataset stand, peopl, next, white, woman, hold, tabl, street, top, person, larg, park, group, ride, plate, field, small, tenni, train, front, black, walk, room, sign, play, red, young, water, look, basebal, build, bathroom, skateboard, tree, food, pizza, blue, kitchen, hors, grass, side, green, bed, bus, giraff, boy, ski, player, fli, snow, car, ball, beach, clock, bear, toilet, coupl, girl, eleph, pictur, wear, lay, game, road, eat, umbrella, bench, laptop, motorcycl, board, sever, zebra, comput, phone, wooden, kite, area, cake, sink, window, glass, wall, boat, chair, truck, brown, yellow, light, frisbe, bird, cover, stop, outsid, open, track, live, desk, air, hand, surfboard, lot, mani, cow, sky, veri, bowl, banana, fill, orang, old, littl, wave, court, fire, flower, child, fenc, togeth, couch, bat, photo, bunch, airplan, sandwich, background, insid, view, color, floor, watch, plane, sheep, big, tie, mirror, display, ocean, cell, counter, cut, shirt, bike, hydrant, traffic, veget, teddi, racket, surf, sidewalk, women, differ, anim, snowboard, tower, piec, drive, jump, mountain, tall, pose, babi, fruit, head, ground, line, hit, grassi, graze, imag, slice, day, set, hill, wood, smile, full, middl, hang, use, slope, readi, pole, skier, station, skate, trick, suit, dirt, drink, box, travel, stuf, crowd, besid, show, abov, wine, guy, dress, luggag, broccoli, herd, snowi, prepar, refriger, kid, talk, wii, book, dure, camera, children, pull, pink, pair, hot, cross, wait, keyboard, passeng, decor, cup, ladi, bicycl, empti, donut, corner, paper, control, stove, surfer, doubl, televis, carri, bottl, busi, hous, soccer, video, bag, tv, suitcas, airport, rock, lie, seat, appl, various, someon, remot, contain, brick, paint, metal, throw, item, coffe, chees, river, racquet, meat, jet, someth, bedroom, cabinet, serv, tray, restaur, night, adult, place, home, catch, rest, shower, male, outdoor, row, carrot, surround,

runway, ramp, brush, toy, work, high, sleep, pile, screen, intersect, store, cook, doughnut, monitor, mouth, dish, plant, lean, number, decker, photograph, batter, offic, scissor, fri, rail, bridg, gray, vehicl, jacket, sand, meter, hair, zoo, cloth, pan, half, dark, way, pass, tub, knife, enclosur, microwav, type, meal, forest, clean, bread, move, bright, read, uniform, mous, silver, past, lake, older, fork, dock, branch, purpl, leav, furnitur, buse, gather, blanket, attach, grey, pitch, cloudi, pillow, chocol, shop, stone, post, helmet, statu, includ, scene, salad, dine, perform, femal, clear, stick, basket, shelf, rain, bath, lush, birthday, nice, lamp, arm, cart, market, turn, lit, public, case, beauti, candl, perch, sunni, shoe, edg, sun, plastic, cellphon, shot, catcher, towel, assort, reach, teeth, enjoy, put, time, fresh, tomato, platform, land, motor, toothbrush, stack, short, yard, underneath, feed, shape, leg, reflect, match, base, bush, dri, pitcher, countri, beer, wet, shore, race, engin, surfac, roll, wire, sauc, path, sandi, direct, famili, applianc, varieti, modern, stare, bar, stripe, dinner, point, flag, multipl, mount, load, chicken, nintendo, carriag, town, polic, trunk, singl, float, ice, gear, potato, kind, neck, pastri, eye, curb, distanc, bathtub, graffiti, dessert, rack, egg, sofa, tarmac, rice, pretti, hotdog, cute, arrang, rider, left, pastur, cattl, curtain, spoon, featur, sail, equip, dirti, concret, steel, electron, right, railroad, onion, center, team, style, clutter, fish, school, vintag, swim, bite, step, sheet, glove, backpack, fireplac, fridg, nearbi, alon, huge, round, trail, wheel, feet, flat, object, church, polar, size, space, stall, giant, bun, atop, desert, profession, wild, grill, asian, cream, squar, lone, bake, mother, pedestrian, commerci, breakfast, thing, friend, rug, cement, pick, subway, shade, urin, gate, stair, frost, soup, cloud, cage, design, pack, coat, blender, pool, christma, devic, ear, smoke, french, shelv, umpir, boarder, wed, duck, garden, lead, sunglass, overlook, highway, kneel, scooter, part, plain, wrap, hillsid, foot, kick, restroom, ledg, stainless, fashion, flock, signal, lift, ceil, low, construct, trash, machin, desktop, drawn, militari, mid, ship, toddler, goat, paddl, rocki, fall, hole, lawn, frame, touch, pie, befor, advertis, closeup, event, veggi, approach, pepper, airlin, tow, roof, block, attempt, outfit, parti, slide, broken, net, log, costum, picnic, blow, neat, lunch, electr, pond, trailer, tan, bottom, fanci, cap, doe, chef, indoor, partial, leash, deck, sport, overhead, narrow, wind, commut, wide, pasta, pant, beverag, new, cupcak, gold, messi, steam, tent, practic, leather, pepperoni, walkway, antiqu, structur, carpet, flip, chain, relax, foreground, toast, transit, check, bend, model, action, worker, doll, sunset, tour, paw, urban, bean, bow, palm, site, chip, doorway, stretch, posit, push, blurri, good, pipe, milk, fighter, beard, mug, held, mushroom, beneath, island, handl, mix, lid, harbor, speed, tag, fan, balanc, power, sausag, farm, lettuc, winter, wetsuit, sort, sprinkl, pier, cooki, help, write, loung,

vaniti, garag, disc, produc, bull, propel, rope, blond, sticker, hard, strawberri, cone, mound, peel, collect, soda, purs, comfort, residenti, word, stunt, utensil, platter, smart, seagul, pickl, grow, shoulder, miss, ripe, juic, upsid, vest, tooth, rural, sell, barn, tea, alongsid, pavement, bacon, noodl, patio, return, tini, sculptur, happi, teenag, raini, ornat, curl, flight, spot, smaller, sill, dresser, climb, elder, ketchup, close-up, art, parad, horn, laid, beig, enclos, sweater, lemon, shine, locat, chop, spread, glaze, mustard, jar, lamb, individu, termin, pickup, spray, groom, wash, driver, bucket, stuff, tail, onli, motorcyclist, share, suppli, american, cours, crosswalk, deep, string, fake, sailboat, laugh, holder, tank, rear, built, fun, puppi, competit, jean, gas, wagon, goe, landscap, grab, spectat, oliv, bride, stream, telephon, stadium, format, selfi, aircraft, roadway, rais, guitar, pave, stool, cycl, transport, snack, napkin, bank, strip, pattern, shirtless, hood, locomot, jetlin, dead, fold, museum, tongu, shadow, tool, shallow, seem, flown, ingredi, section, freezer, prop, motion, separ, homemad, headphon, floral, crouch, toaster, toss, fluffi, ring, marbl, patch, numer, cupboard, checker, opposit, drawer, stuck, garbag, shake, tire, lane, semi, present, fix, break, steep, asleep, microphon, fast, tourist, neon, visibl, hug, entranc, rust, habitat, dim, biker, muffin, chew, poster, skirt, taxi, downhil, whole, celebr, rang, horseback, golden, complet, stoplight, finger, newspap, motorbik, basketbal, balloon, feeder, stage, chase, figur, shini, cargo, delici, pigeon, iron, extend, pad, coach, lick, life, leap, straw, organ, curv, london, card, pancak, layer, mobil, exhibit, gravel, connect, butter, peek, observ, corn, mat, focus, liquid, furnish, beef, collar, great, thin, mark, burger, miniatur, interior, owner, mope, steepl, mask, foil, dip, figurin, rusti, fountain, natur, mini, boot, key, state, buss, start, trolley, booth, crust, condiment, plaid, hamburg, formal, hose, indic, hook, hide, cardboard, tabbi, entertain, hospit, consist, rainbow, candi

Fashion Dataset fit, cotton, soft, waist, sleev, regular, round, necklin, button, pocket, front, print, detail, fabric, look, rib, comfort, design, elast, fasten, chest, side, neck, stretchi, knit, polyest, line, style, cuff, tshirt, leather, dress, environ, adjust, top, classic, slim, low, jacket, size, high, collar, leg, qualiti, wool, closur, jean, short, trouser, dri, sole, shirt, belt, materi, inner, item, set, blend, heat, stretch, function, easi, tumbl, graphic, zip, perfect, organ, feminin, wear, skirt, longsleev, basic, heel, upper, shoulder, stripe, press, pattern, featur, viscos, colour, tie, use, breathabl, stud, lace, strap, toe, rubber, denim, allov, collect, finish, relax, textur, pair, craft, feel, loos, length, height, lightweight, tight, brush, outer, vneck, pad, onli, unlin, simpl, wide, super, loop, pullov, layer, hood, raw, contrast, combin, pack, return, pesticid, natur, floral, boot, faux, shortsleev, slit, jack, jone, warm, sneaker,

everyday, blazer, curv, crew, ankl, cardigan, ruffl, logo, frill, panel, silhouett, cool, eleg, casual, crop, product, blous, polo, day, elastan, glitteri, sweatshirt, harm, durabl, mean, insid, veri, overs, great, wrap, solid, zipper, zipup, chunki, drawstr, hoodi, touch, chemic, mix, skinni, slight, fli, conceal, flexibl, season, coat, thigh, open, year, temp, light, versatil, grown, wardrob, suit, placket, drop, onlin, synthet, formfit, taper, produkt, room, work, match, pull, midi, silver, pant, pleat, block, laceup, recycl, possibl, beauti, moda, structur, vero, grow, sidepocket, trim, mm, ensur, merino, trendi, crotch, waistband, insecticid, creat, point, complet, fur, jersey, way, string, airi, cut, lean, condit, sweat, wash, straight, embroid, mani, shoe, flounci, flat, metal, buckl, kid, extra, bottom, sterl, trend, knee, check, piec, fertil, fix, winter, mini, polybag, normal, mesh, essenti, nylon, stylish, wonder, ean, spun, code, unbrush, offer, children, age, possess, shape, mock, corduroy, good, hip, refund, sent, sandal, sourc, farm, best, worker, protect, stay, narrow, slipon, vnecklin, turn, outsol, modal, color, balloon, fibr, sweatpant, summer, origin, white, boy, wearabl, drawcord, tailor, sustain, doubl, bit, sleeveless, sweater, stretchabl, effect, pure, hook, tier, keyhol, stand, appliqu, play, movabl, reflect, disney, nurs, move, effortless, polyurethan, minimalist, occas, timeless, detach, trainer, smooth, warmth, rip, goldplat, clean, leopard, flap, movement, band, plenti, tab, sequin, temperatur, ani, amaz, comfi, babi, hemlin, technic, worn, glitter, insul, embellish, new, sock, quilt, bow, hidden, awar, abov, moistur, dot, blue, dure, differ, notch, beani, almond, romper, stapl, black, label, figur, junior, noisi, scarf, paspel, outfit, cap, construct, roll, weekend, everyth, chic, sure, toddler, trunk, gold, seam, need, breast, spaghetti, various, sharp, puffer, perfor, show, glenn, sound, tull, scratchi, insol, fine, luxuri, semish, manag, water, ski, small, youll, backsid, part, exclus, shini, motif, buttonkeyhol, edg, smock, hand, imit, gather, partial, maxi, transit, subtl, contemporari, spring, raglan, twill, silk, whi, option, longer, chelsea, night, ideal, abil, someth, acryl, heavi, everi, sleeveend, provid, welt, patch, tape, premium, team, medium, pom, lapel, build, tank, flare, women, diaper, golf, slip, support, matter, sheer, refin, cosi, favourit, becaus, excel, time, knot, type, embroideri, stitch, fill, softshel, polyamid, cooler, sewnin, roomi, cuf, turnedup, neat, modern, perform, volumin, tradit, slant, mid, guid, fashion, chino, lot, contain, cow, repel, carat, fivepocket, aim, velour, month, rise, italian, bomber, curvi, jumper, ashap, tuck, letter, anatom, hshape, cover, snap, correct, ontrend, derbi, big, cargo, alreadi, insock, express, fuzzi, teddi, initi, ring, tim, suppl, buttondown, ecovero, lenz, highlow, neon, kangaroo, thin, fineknit, choic, grip, bodysuit, christma, rollneck, technolog, extend, game, deterg, cabl, inspir, fray, squar, wrist, lil, plate, ateli, larg, garment,

sporti, piqu, drape, cold, ultim, butterfli, bag, child, formal, aw, midlay, jumpsuit, fulli, fluid, kind, fresh, weav, alfa, tieband, tonal, mould, plastic, rain, mandarin, chestpocket, tunic, recommend, fleec, period, canva, count, invalu, supplement, practic, underneath, oneck, liam, rhodiumpl, parka, half, tech, closefit, wind, outerwear, linen, poplin, windproof, retro, maintain, twist, snake, horn, pendant, mous, star, certifi, nice, tuckedin, left, statement, brand, dont, feet, smart, suspend, lurex, extrem, eyecatch, pearl, primaloft, plain, thick, highwaist, chain, highlight, alik, lyocel, copenhagen, jacquard, asymmetr, select, bore, threequart, profil, doublebreast, streetwear, fp, upgrad, studio, ear, activ, icon, loafer, alpaca, offic, signatur, volum, singlet, tee, anklestrap, accessori, velcro, onecklin, easili, romant, decor, goe, liner, deep, frontsid, import, yarn, sunglass, distress, playsuit, pressstud, scallop, littl, girl, singlebreast, interpret, silki, tube, vertic, anim, sidezip, reason, tour, width, equip, accord, fold, peplum, weather, pliss, tap, cubic, zirconia, classi, standard, visit, revers, matern, eyelet, mitten, rich, truuli, surfac, biker, boat, eva, fals, alin, smarter, autumn, puff, delic, fieldsensor, onebutton, glove, portug, lenght, tone, help, outstand, proud, full, slimfit, compart, shir, quick, member, slouchi, lantern, frozen, multicolour, level, singl, key, tread, remov, fiber, pima, paperbag, born, weight, dark, warmer, carri, superior, waistlin, realist, sinc, itali, onepiec, waterproof, increas, midris, fortnit, soldier, power, minni, asid, draw, pig, noniron, camo, velvet, secur, ameli, accentu, basebal, ventil, threebutton, cottonblend, love, ultra, emboss, clark, leav, sophist, aesthet, flip, cours, mulesingfre, youv, hybrid, text, brief, crossov, runway, flounc, brim, version, edgi, cloth, capri, boxer, trenchcoat, shawl, crepe, simplic, tencel, footb, turnup, bikini, sever, satin, subt, desert, smartcasu, desktodinn, slimleg, seamless, measur, clear, becom, parti, vest, fixtur, special, destroy, push, bootcut, shell, thermo, snow, zippocket, optim, sell, astrali, edit, freedom, sparkl, goto, true, ava, overal, lock, soap, sturdi, uniqu, heart, plung, lend, spread, interior, ultrasoft, thermoplast, pregnanc, chilli, thread, visual, diagon, stiletto, chevron, culott, r, figureskim, textil, split, flatter, sport, vcut, sneak, tough, grant, musthav, higher, hoop, poetic, scandinavian, languag, oekotex, headband, nightsuit, substanc, colourway, test, pretti, checker, bring, bestsel, cleancut, wider, bridg, oxford, pullon, certain, weve, tunit, indigo, pli, cup, enzym, cord, dinosaur, nubuck, finer, vibe, stomach, environment, sewn, cutlin, loosefit, avail, cashmer, hint, hipster, pulltab, summeri, event, easyiron, tongu, pipe, promin, twopli, box, amount, exud, right, refineri, turtleneck, blackandwhit, backpocket, hello, defin, bold, satini, favorit, vent, watson, stuff, nightwear, longlin, pick, hornlook, compress, underst, thong, softest, sartori, mill, foot, colorblock, cottonstretch, updat, bright, badg, panda, paper,

pu, selvedg, supersoft, zebra, addit, impact, footwear, crochet, profession, couldnt, standaway, lower, properti, advic, cleaner, seek, improv, snakeskin, motherofpearl, colder, rock, fast, unlinedpolyurethan, skin, insert, authent, wideleg, cableknit, traction, lustr, triangl, walk, storm, dermizaxev, attitud, bamboo, coolmax, transeason, loa, guy, serv, terri, artwork, handl, unbutton, shade, trench, stack, indispens, busi, rhodium, solidcolour, robust, slope, heelfriend, allur, gorgeous, longjohn, ecofriend, reduc, resourc, rivet, wast