Project Proposal

Jiaqi Song (js4979), Naiqiu Zhan (nz155), Wenxuan Tang (wt254), Zhaoyuan Qiu (zq37)

Background Introduction

E-commerce is now all around people's lives. We can very easily find the products we need through the Internet and get the products we need with the click of a finger. The role of advertising for commercial products is to generate consumer interest, win customers through publicity among similar products, and attract users who may be interested in the product. Most advertisements will emphasize the product's advantages through typography, aesthetic design, etc. However, it will take the merchants lots of time to design fancy captions to attract consumers. If there is a tool to generate vivid descriptions automatically just based on a set of brief keywords, then it can help the merchants to largely reduce manual devotion and decrease costs on the advertising part. Moreover, the main information of the product can be quickly and effectively presented to consumers, which will improve the service of the e-commerce platform and increase user viscosity, which plays a positive role in the development of e-commerce.

This project aims to form a summary by processing both textual and visual information on e-commerce product pages. This is more challenging compared to process traditional text, but ideally this summary can adequately represent more effective information on the product page, since this type of summary could capture the selling points based on the appearance of the product as it usually creates a first impression of the product to the consumers.

Data Overview

The dataset consists of two parts. One for the commodity images and the other for corresponding text descriptions.

The text descriptions are saved as 3 JSON files including 30000 records, totaling 42.8 MB. Each record is indicated uniquely by a hashed string of length 10. For each record, the

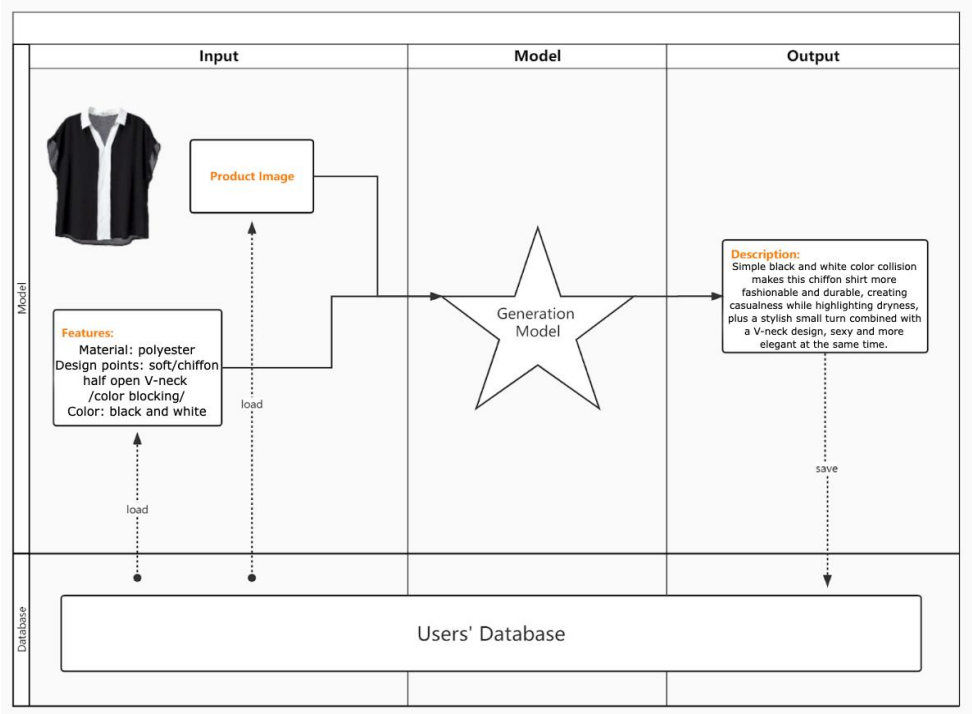
contents are saved into 3 keys, ’tgt’ for target generated text, ’cate’ for commodity category, and ’src’ for initial descriptions. The texts have gone through preprocessing including tokenization, removing stopwords, and dropping punctuations.

For image data, 30000 images of the inconsistent size and pattern are collected, whose size sum up to 1.98GB. Each image only contains one commodity. Either in a white background or with a model. Examples are shown below.

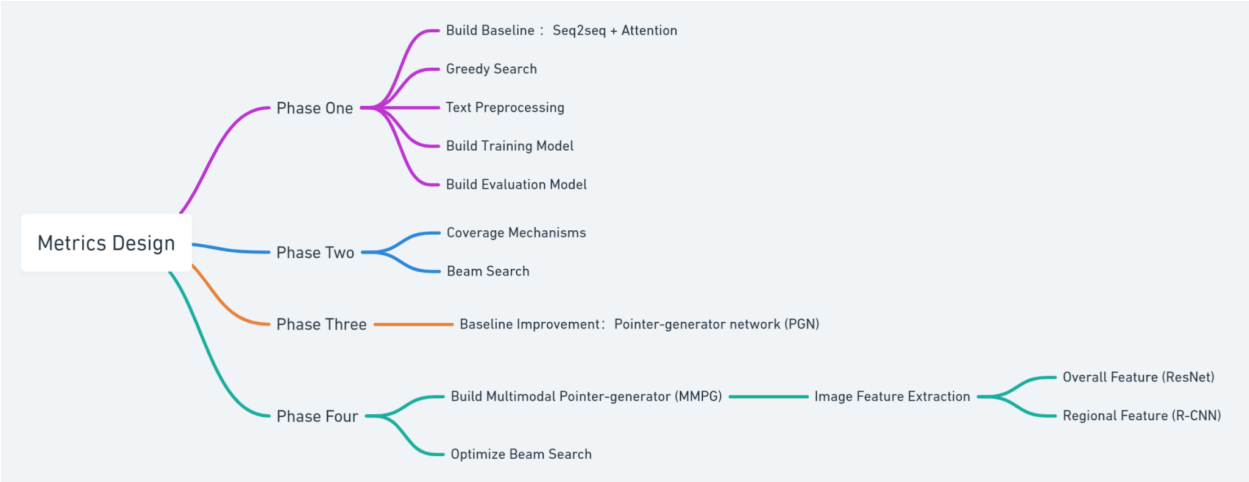


Prototyping

We are trying to design a solution to generate the clothes description automatically. The users can select and import the clothes images and their features or labels, and then the model will generate the description of the clothes for marketing purposes.



Model design



For the model's design, we first decided to use deep learning frameworks. This is still a natural language processing (NLP) task, although we will eventually incorporate information from the images into the model. We considered this a text summarization task based on the data situation we could use and what we ultimately wanted to accomplish. Based on the data available to us and what we ultimately wanted to accomplish, we considered this to be a Text Summarization task. According to the output type, there are Abstractive Summarization and Extractive Summarization. Extractive Summarization consists of key sentences and keywords extracted from the source document, and the abstracts are all derived from the original text. Abstractive Summarization allows the generation of new words and phrases to form the abstract based on the original text. Because we want to generate text that includes key information from the original text and we want to have some novel sentences, we will use a framework that combines these two approaches methods in the NLP section. To combine the information from the images, we will also add a computer vision (CV) module to build a multimodal framework

The project technique design will be conducted in four phases.

In the first phase, we will build a baseline for NLP, which is mainly based on Sequence-to-Sequence Model (seq2seq) and Attention Mechanism. For text generation, the baseline model will use a greedy search. Other work in this phase will include text pre-processing, building a training and evaluation pipeline, and other steps.

For the second phase, we will add Coverage mechanisms to reduce generation repetition and apply the Beam Search approach to our NLP model, which will increase the accuracy of language processing.

Phase three will be an improvement for our baseline by applying a pointer-generator network (PGN) to our NLP model. PGN will adjust the weight of each word to implement a word coping mechanism, and then the final output description will contain the important words in the original text.

A multimodal pointer-generator (MMPG) will be created in the last phase to complete our project technique design. The MMPG will combine the previous NLP model and a new CV model to improve our model inputs. We will extract both overall and regional features from each image to assist the NLP model in generating some significant features.

Metrics Evaluation

The metrics evaluation mainly consists of two parts. Based on our NLP model, we will use Rouge to measure the accuracy of output. In the first phase, the baseline model will convert the sequences following a one-dimensional direction, so the results of Rouge-1, which refers to the overlap of unigram (each word) between the system and reference summaries, will represent the accuracy of our baseline model. For a modified model with coverage mechanisms and beam search in phase two, we will use Rouge-2, which refers to the overlap of bigrams between the system and reference summaries, to evaluate the accuracy of the optimized model. Besides these designed metrics, we will go through the results manually to check the fluency of output sequences.

For computational/hardware considerations, we will use Google Colab to implement our model to meet the need to remove personal computers' limitations and work together.

Expected Deliverable

The most significant unknown that might dictate the success or failure of our project is the quality of images, since this project's most considerable challenge is combining text summarization with image mining. Our desired result is that the combination will perform much

better than only considering the text summarization method. Therefore, if the quality of the images from the dataset is high, then these images will clearly show the remarkable features of products, which will also be consistent with the textual information of the corresponding products. On the contrary, images with low quality cannot convey accurate information to our model so that the generator will give similar outputs for entirely different kinds of products.

We will show a live demo of our summary generator along with a brief oral presentation. In the demo, we will randomly pick up and input a set of product information into our model, including product images, titles, and descriptions, to see whether the output summary will effectively leverage the visual information in product images and capture the most salient aspects of products, as well as stay readable and non-redundant.