# Caption Generation for Amazon Products Using Weak-Supervision

Ian Raphael Sim Ek Yan  Ma Yuan  Nicole Lee Zhi Ying  Yong Zhun Hung

*National University of Singapore*

## Abstract

Given the boom of online e-commerce in the recent years, it is important for individual merchants to stand out by conveying accurate information about each product or item that they sell online. This is particularly important in e-commerce where the customer is unable to physically assess such properties of the product. It is therefore vital for these online stores to have accurate and compelling descriptions of their products which customers can use to make an informed decision before committing to a purchase.

We see a potential in automating the process of copywriting using data inherent in the product. These include pictures, category data, metadata and short titles of the product. These features are then parsed into a machine learning model trained to generate and output a cohesive and compelling natural language paragraph describing the product with great and accurate detail.

## Dataset

Our dataset is adapted from the Amazon product/review data dataset, published by McAuley and updated by Ni (UCSD) [1]. The dataset consists of 233.1 million Amazon product reviews spanning May 1996 - Oct 2018 and includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

Since our primary focus is on using images and textual data to generate descriptions, we remove the unnecessary fields such as reviews and brand and retain only those which are useful to our model. These fields are listed below:

- Product Images (URLs of both High and Low Resolution Image of each Product)
- Categories (List of categories the product belongs to)
- Description (Short text description of the product)

Additionally, the raw data is over 20GB in size, which far exceeds what we are able to effectively train given the short time frame and limited GPU resources. Therefore, for this project, we uniformly sampled data from each of the 25 product categories, resulting in a final dataset with just over 7000 entries.

## Background

The paper *Automatic Product Copywriting for E-Commerce* (Zhang et. al. [3]) aims to achieve a very similar objective - automatically generating copywriting using product data. Our team's approach, however, includes the use of computer vision models to process the uploaded product images. In their paper, Zhang et. al. adopt a primarily text-based approach, focusing on NLP transformers and GPT. Our methodology targets a two-pronged approach, where the primary source of information is the visual features given by the product images and the secondary sources of data are the textual data and metadata. The approach is designed based on our belief that images are able to provide more accurate and descriptive features of the product that can be delivered to the customer. In order to achieve this, our project is largely guided by the methods and models proposed by the paper *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention* (Xu et. al., [2]).

The project mainly aims to apply techniques of image captioning to a highly-valued business context of product copywriting. Originally, the primary applications of image captioning are confined to datasets such as the MSCOCO or Flickr datasets which are large-scale, well-labelled datasets intended for object detection, segmentation, or captioning tasks. These datasets depict a myriad of situations and or circumstances which are not directly relevant to our objective as they mainly consist of people or animals performing activities or locations and objects.

## Methodology

The original encoder-decoder model from the show, attend, tell paper by Xu et. al. forms the foundation of our model. However, in order to combine the various forms of image, text and categorical data, we augment the model with two additional components. These are the textual description data and the one-hot categorical data. Data from the description, which consists of short sentences, was first cleaned, pre-processed and tokenized using an embedding layer. Subsequently, it is combined with the outputs of the encoder model and categorical input which is then fed into the LSTM decoder model. In order to preserve the positional data for the attention mechanism, sentence indices are also stored.

We propose and experimented with three different architectures; solution 1, 2 and 3 as can be seen in the diagram below. Solution 1 is the simplest model where we simply concatenate the output of the description embeddings with the attention block and combine it with the one-hot categorical data before feeding it into the decoder. Solution 2 increases the complexity by introducing connections from the categorical and description outputs into the attention block inputs, thereby allowing the model to attend to different parts of the text or category data. Finally, solution 3 uses zero-shot learning to try to predict previously unseen data.

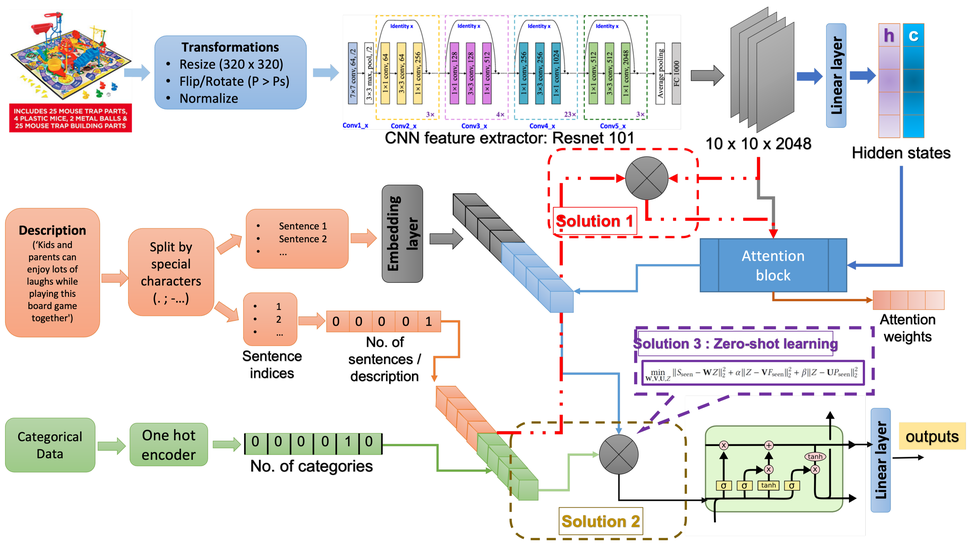


Figure 1. Architecture Diagram with Proposed Solutions

### Encoder

- Uses a pre-trained ImageNet ResNet-101 model for the encoder
- Removed linear and pooling layers of the ResNet-101 model
- Incorporated an Adaptive Average Pooling to allow for variable image sizes. This is quintessential in the business context where users often do not upload images of the same sizes.

### Attention

- The attention block uses inputs from the image encoder and LSTM text decoder to attend to regions in the image corresponding to the next predicted text token at each LSTM cell.
- In solution 2, the attention block additionally has inputs from the categorical vector which are combined with the image data with the intuition being that the attention model can attend over different product categories differently.

### Decoder

- Uses a standard LSTM network to output a series of words
- Initial Input to the decoder is
  - Solution 1: Output of Attention Model
  - Solution 2: Output of Attention model and Embedded Category/Description data
- Output of each LSTM cell is a softmax normalized probability distribution over the complete vocabulary and the word with the highest probability is chosen for that cell

## Motivation

**To generate a descriptive natural language paragraph given an input image and its meta data**

## Experimental Results Discussion

We conducted an initial experiment using only the low-resolution images and the title. However, this combination of training data was unable to generate meaningful paragraphs, in order to achieve the goal of copywriting.

| # | Training Data | Validation Loss | BLEU-1 Score |
|---|---|---|---|
| 1 | High Resolution Image + Attention for Description | 3.5003 | 0.73 |
| 2 | High Resolution Image + No Attention, Concatenated Description | | |
| 3 | High Resolution Image + Description + Zero-Shot Learning on Categorical Data | | |

Table 1. Experimental Results

While comparing the different approaches and different combinations of training data, we found that **Solution 1** provided the most realistic and natural descriptive paragraphs, which can be used for copywriting.

## Next Steps

Given the experimental results, the team will be looking into making the following improvements to the model:

- **Zero-Shot Learning - Solution 3**: Decrease the model's reliance on labelled data and further improve generated captions for unseen products
- **Beam Search**: Currently, the next word in the decoder sequence is chosen greedily, based only on the local probability of that cell. We could potentially introduce beam search to optimize the selection of the next sequence of words and maximize over the joint probability. This could potentially give better results.
- **Larger Dataset**: We are limited in scope and scale by the lack of access to high-power computation resources as well as time constraints. As we are currently using only 0.003% of the available data, it stands to reason that performance will be much improved with the larger complete dataset.

## References

[1] Jianmo Ni, Jiacheng Li, and Julian McAuley. *Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects*. 2019.

[2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2015.

[3] Xueying Zhang, Yanyan Zou, Hainan Zhang, Jing Zhou, Shiliang Diao, Jiajia Chen, Zhuoye Ding, Zhen He, Xueqi He, Yun Xiao, Bo Long, Han Yu, and Lingfei Wu. Automatic product copywriting for e-commerce, 2021.