EECS 595 Project Proposal: Generating Valance Shifted Captions from Images

Evan Czyzycki and **Drew Davis** and **Lloyd Shatkin** and **Siddharth Venkatesan**
University of Michigan

1 Project Description

In the project checkpoint 1 report, our team decided to focus on the broad problem of image caption generation. This problem involves taking an image as input, and generating a sentence as output that describes what is occurring in the image. This is a popular topic in the fields of computer vision and natural language processing, so we explored ways we could extend beyond current work and make our project novel. After surveying the area of image caption generation, we decided to focus our project on incorporating valence shifting into the image generated captions. At a high level, valence shifting is the process of rewriting a given text to preserve the original meaning while altering the sentiment characteristics. Our system will take in both an image and slant factor as inputs, and produce a slanted caption as output. While there is previous research on both image caption generation and valence shifting, to the best of our knowledge there has not been research conducted on a valence shifted image captioning system.

We believe this is an interesting problem because it represents a process that humans subconsciously go through when they are asked to describe an image that has emotionally-charged context. As an easy example, consider politics in the US. The image shown in Figure 1 is a fairly neutral shot of Donald Trump giving a speech. Most people will understand objectively what is happening in the image; however, if many different people were asked to caption the image, their answers would likely show quite a bit of semantic variety. That variety is a result of their political biases coming through. For example, a politically right-leaning person may describe this image using words such as confident, assertive, or independent; whereas a politically left-leaning person may use words such as angry, accusational, or aggressive. Through this project we hope to capture this human process of understanding and describing an image with a perspective bias.

2 Method Description

Figure 1: Image of Donald Trump Giving Speech

We now explain in more detail the two major components of our project.

2.1 Image Caption Generation

Image caption generation takes an image as input and, through image classification and natural language generation, creates text that describes what is happening in the image. The image classification step aims to detect objects and relevant features in the image. These detected objects and features are then either sent directly to the caption generation module, or first converted to textual representations before being sent to the caption generation module. Using these features as inputs, the caption generation system aims to generate a sentence that accurately describes what is occurring in the image. Image caption generation systems generally use an RNN or an LSTM for text processing and a CNN for image processing, combining the output of both into a dense layer before final textual output. These text descriptions generally don't include many adjectives or adverbs due to subjectivity, but we intend to inject adjectives and adverbs to the generated captions for the purpose of valence shifting.

2.2 Valence Shifting

As mentioned earlier, valence shifting entails preserving the meaning of a sentence while altering its semantic perception (i.e. its ”goodness” or ”badness”). The input to a valence shifting system is a sentence or phrase, and the output is a sentence or phrase with similar meaning but with a perceivable bias. An obvious method to achieve this could be shifting adjectives and adverbs to words that are closer to the desired valence while a more complex method might include sentence rearrangement and synonym substitution. Valence shifting is a difficult task, so we expect that a primary focus of our project will be on developing an effective valence shifting module. A unique challenge of valence shifting within this project is that often shifted sentences have changed meaning compared to their original text, but our system must maintain a caption that is an accurate description of the provided image.

3 Related Work

Although the exact process of this project is unique, the procedure naturally spans two well-established sub-domains of computational linguistics: Natural Language Generation (NLG) and Valance Shifting (an extension of sentiment analysis).

3.1 Natural Language Generation

Natural Language Generation (NLG), for the purpose of caption generation, is a major part of our project and has a wide range of published material. To assist in navigating the state of the art a survey of the field was published in 2018 (Gatt and Krahmer, 2018). Within the field of NLG, conversion between images and textual descriptions has been a hot topic, which has been met with a fair amount of success over the past decade (Bai and An, 2018; Rashtchian et al., 2010; Young et al., 2014). An example of NLG in the context of caption generation is shown in a recent project, which was able to successfully generate meaningful image captions using an RNN and a CNN (Xu et al., 2015). This process is able to identify basic objects and actions occurring in an image and create a concise sentence that explains the image. We intend to use this project's model as the fundamental caption generator for our project. We assume that, much like any published code, we will be unable to use it out-of-the-box, and anticipate the need for minor additional modifications. If we are unable to get the model working, we will pursue alternative published caption generation models.

3.2 Valence Shifting

Previous published work on valance shifting discusses the methodology behind the Valentino method of shifting, which includes paraphrasing, lemmatization, insertion of downtoners and intensifiers, and modification of adjective strength (Guerini et al., 2008). For our project, we will focus on the insertion and modification methods. To know appropriate replacement words, the Valentino method uses WordNet, a language database, to gather related words (Miller, 1995). For adjectives, the ”similar-to” relation was used, and for verbs and nouns the ”hyponym” relation was used. Our version of valence shifting will take advantage of similar methods of sentence modification. The insertion method will play a prominent role in valence shifting on generated captions because they are typically generated without the use of adjectives or adverbs. Rather, captions are a combination of simple noun phrases and verb phrases using the objects found in the image processing step.

4 Data Collection

For the image caption generation portion of our system, we are planning to use the Microsoft COCO Caption dataset as our primary training and evaluation dataset (Lin et al., 2014). The COCO dataset is a popular dataset used for both training and evaluation in the area of image caption generation. This dataset contains more than 200k images containing objects belonging to 80 ’object’ categories, and

91 'stuff' categories. Each of these images is associated with 5 captions, which makes this an extremely comprehensive and well-suited dataset for image caption generation.

Collecting data for the valence shifting component of our project is a much more difficult task. There is no clear single dataset that we can use for this task, so we will need to use multiple datasets to develop a good valence shifting system. Initially, we plan on shifting valence through adjective and adverb substitution. We aim to build a language model that associates valid adjective to noun relations, and valid adverb to verb relations. For example 'comfortable chair' is a valid relation while 'spicy chair' is not. In order to designate whether a adjective-noun or adverb-verb pair is valid, we will look to see if this pair has been seen in a other texts. To observe these relationships, we will use Google n-Grams and record all possible bigrams with the proper POS pairs mentioned previously. We plan on using Amazon Web Services (AWS) to access the dataset because the AWS interfaces makes it easy to download all bigrams from a certain time range. We will then filter these bigrams using POS tags to keep only bigrams that are valid adjective to noun relations, and valid adverb to verb relations.

Given the newly inserted words, we will use WordNet as a database to find synonyms (Miller, 1995). WordNet is a lexical database that contains groupings of words that are related to each other. Then, for each synonym we found, the polarity of the given word, as well as the given sentence, will be determined using the TextBlob sentiment analyzer. While this is not necessarily data collection, it s a form of data analysis that will be highly important in our project.

5 Method Description

The method of going from an image to a valance shifted sentence is depicted in Figure 2.

5.1 Valance Model Learning

The generation of our valance model begins with the WordNet corpus. This corpus is used obtain probabilities of adjective-noun and adverb-verb pairs. Every such pair is saved, along with the sentiment analysis score received from passing the full sentence into TextBlob. Every instance of bigram and score is then used to train our valance ranking model, which takes a bigram input and outputs an expected valance score for a sentence containing that bigram. This model will be set up as a standard neural network with one hot encoded text input and a numerical output between 0 and 1.

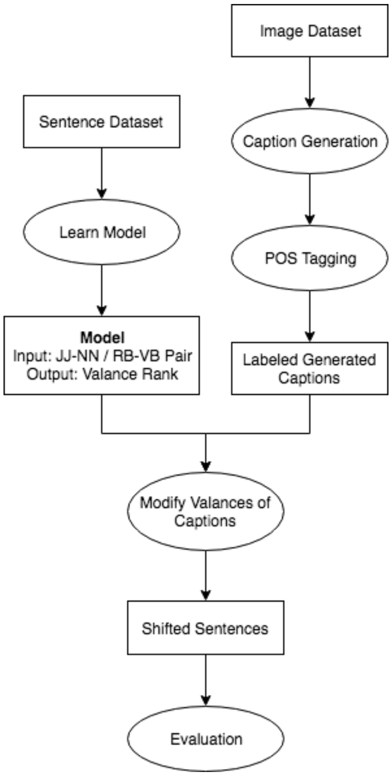

Figure 2: Process Flow Diagram of Generating a Valance Shifted Sentence from an Image

5.2 Caption Generation

Captions are generated from a large set of existing images. These images are passed through the caption generation model provided by (Xu et al., 2015). This model will produce a simple sentence for each image explaining what is depicted in the image. After the captions are generated, they will be passed through a part-of-sentence tagger to label each word in the caption. a set of data that consists of generated and labeled sentences. These two steps leave a set of captions that have been generated and labeled.

5.3 Valence Shifting

The process of valence shifting uses the learned model to alter a caption so that it has positive or negative sentiment. Given a caption - a sentence - as input, the valence shifting system will identify nouns and verbs that the text training data has record of, and add adjectives and adverbs to the sentence accordingly. The adjective-noun and adverb-verb bigram frequencies found from the original text corpus are used to make probabilistic guess of which adjectives and adverbs can be logically inserted into the sentences produced by the caption generation model. Each high probability bigram will be passed into the valance ranking model to get an prediction on how that pair affects the overall valance of a sentence. The pair that produces the valance most closely to the desired valance is selected to be inserted into the sentence.

Initially, we will simply choose a modifier, which connects to the given noun or verb, based on its frequency and valance effect. Once the model has basic functionality, we will experiment with more sophisticated modifier choices, perhaps including the word before the modifier or allowing multiple passes through the system where the model will interchange already-inserted synonyms and antonyms to further tweak the valence of the sentence.

References

Shuang Bai and Shan An. 2018. A survey on automatic image caption generation. *Neurocomputing*, 311:291 – 304.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Marco Guerini, Carlo Strapparava, and Oliviero Stock. 2008. Valentino: A tool for valence shifting of natural language texts. In *LREC*. Citeseer.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations. *Transactions of the Association for Computational Linguistics*, 2:67–78.