Final Project: Attention Analysis and Visualization on Multimodal Models

Abstract

Visualization is a powerful tool for humans to understand and interpret large amounts of data. Attention mechanisms in natural language models appear like a prime target for visualization, but these can be spurious or uninterpretable in many models, such as recurrent neural networks. This project explores the ability for qualitative and quantitative analysis of attention in multimodal models, specifically Cai et al.'s sarcasm classifier, to provide insight into decisions and highlight weak spots in the model's performance.

1 Introduction

Interpretability is a constant struggle in natural language processing (Jacovi and Goldberg, 2020), where models are increasingly being trusted to perform real-world tasks while simultaneously becoming more complicated and opaque. Using attention weights (Bahdanau et al., 2016) as a stand-in for feature importance is a tempting solution, but faces a myriad of problems. For once, attention over latent features is meaningless to humans. Additionally, in some models, attention over observed inputs is misleading. For example, attention over timesteps in a recurrent neural networks has shown to be a poor indicator of feature importance due to the manner in which RNNs opaquely encode information from other timesteps (Jain and Wallace, 2019). As an experiment, I visualized attention from the trained seq2seq Geoquery (Zelle and Mooney, 1996) model from Project 2. Figure 1 shows an example of a correctly parsed query with expected attention in some places ("texas" to `texas`) but not others ("largest" to `_largest` or "bordering" to `_next_to`). Nearly all of the examples I observed from this dataset showed similar behavior. In models where the task is less simple, relying on attention weights to interpret model decisions would lead to spurious conclusions. However, Jain and Wallace also concluded that attention weights in feedforward neural networks were much more similar to feature importance metrics.

The rest of this paper is focused on analyzing attention in the model from Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model (Cai et al., 2019). The model predicts sarcasm in Tweets that include image and text. Attention analysis in this model and those similar is not only feasible (most of the layers are feedforward, rather than recurrent) but valuable, due to the ubiquity of multimodal data across social media and other internet mediums and the nuance of a task such as sarcasm classification.

The version of the model I used was a Pytorch implementation from GitHub (D-Blue, 2020) which I cloned and trained on my laptop using a discrete GPU. I ensured the performance metrics were on par with the original paper's before moving forward.

A diagram of Cai et al.'s model is included in figure 2. Three modalities are used: image, text, and attribute. To extract the image features, the image is split up using a 14-by-14 grid and each of the 196 squares are fed into a pretrained ResNet-50 V2 model (He et al., 2016). Five single-word attributes are extracted from the image using another ResNet model trained on an image-captioning dataset. The Tweet text is fed through a Bi-LSTM (Huang et al., 2015). In addition to the feature vectors, each modality computes a guidance vector by arithmetic or attention-weighted average over each element in the feature vector. The guidance vector can be thought of as a dimension-reduced version of the feature vector, and all three are the same size.

This paper is centered around analyzing the attention weights of the next steps: representational

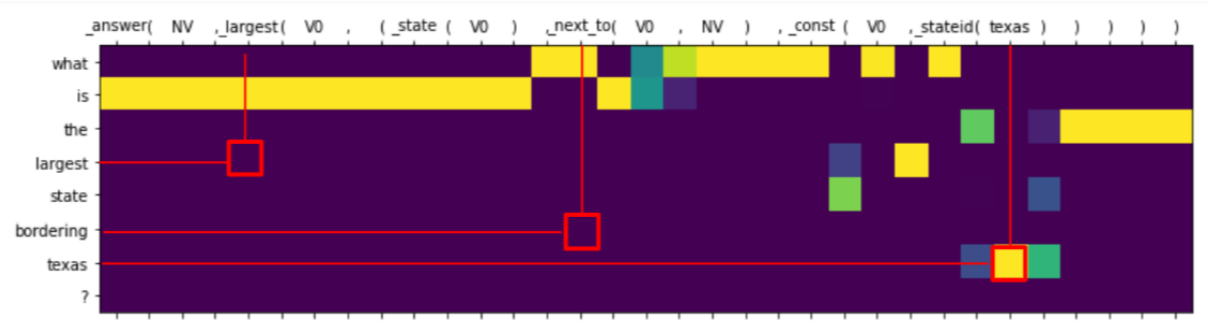

Figure 1: Attention visualization of an example from Project 2.

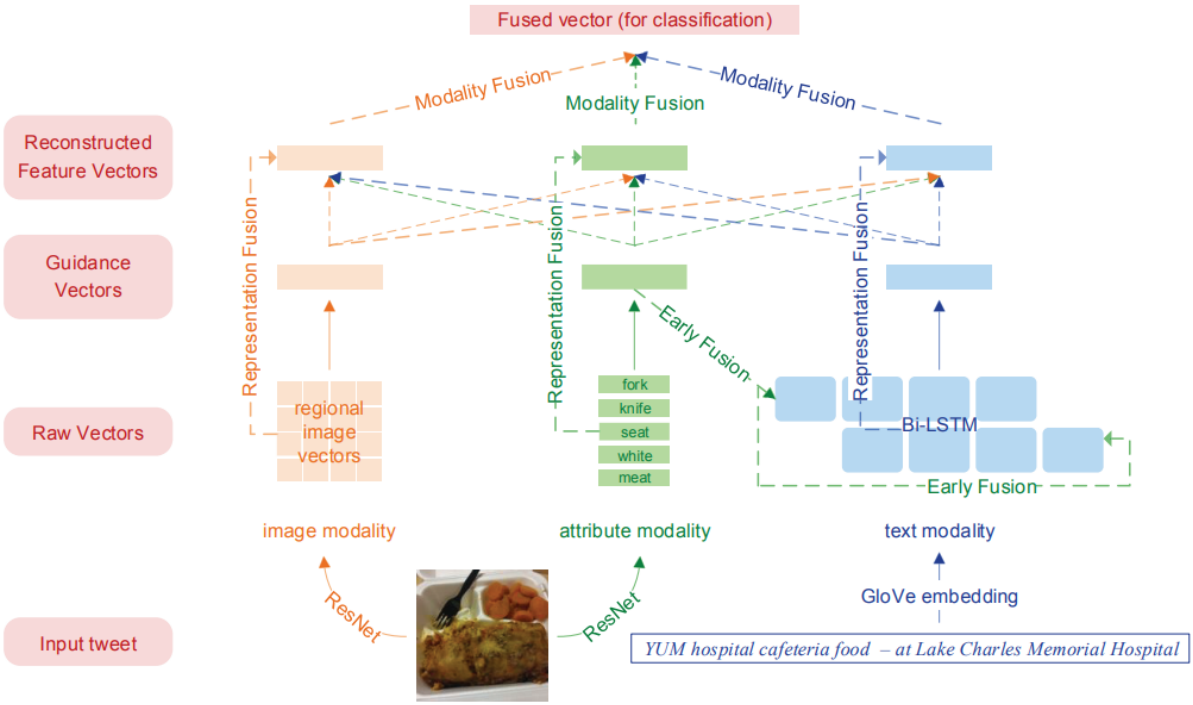

Figure 2: Architecture of Cai et al.'s model.

fusion and modality fusion. **Representational fusion** re-weights the feature vector of each modality m by computing three attention weights over each element in m, one computed from each modality n's guidance vector. We now have nine $\alpha_{m,n}$ attention vectors that represent attention over m computed using information encoded in n. The three attentions over one m are arithmetically averaged and then used to recompute the feature vector.

$$\alpha_{m,n}^{(i)} = W_2 \cdot \tanh(W_1 \cdot [X_m^{(i)}; g_n] + b_1) + b_2$$

$$\alpha_{m,n} = \text{softmax}(\alpha_{m,n})$$

$$\alpha_m^{(i)} = \frac{1}{3} \sum_{n \in \{\text{txt, img, attr}\}} \alpha_{m,n}^{(i)}$$

$$X'_m = \sum_i \alpha_m^{(i)} X_m^{(i)}$$

where X_m is the feature vector being attended and g_n is the guidance vector for the source modality

Next, the three re-weighted vectors are combined into one fused vector using **modality fusion**. Attention across each modality (three scalars) is computed. The three feature vectors are transformed to equal lengths and the final feature vector is computed by taking the attention-weighted average. This vector is then fed into a fully connected layer that outputs a classification decision.

$$\alpha_m = W_2 \cdot \tanh(W_1 \cdot X'_m + b_1) + b_2$$

$$\alpha = \text{softmax}(\alpha)$$

$$v_m = \tanh(W_3 \cdot X'_m + b_3)$$

$$v_{\text{fused}} = \sum_{m \in \{\text{txt, img, attr}\}} \alpha_m v_m$$

These twelve attention vectors together encode a large amount of information and when used together may provide deep insights into the model's decisions. The next two sections will cover the qualitative and quantitative analyses I conducted using these attention weights.

2 Visualization and Quantitative Analysis

We can visualize attention over each of the three modalities, as seen in Figure 3. Again, we have to remember to exercise a level of caution regarding attention over the text tokens, as RNN attention is not always directly representative of true importance. Because of the unique way that representational fusion works in this model, we can also split attention into each $a_{m,n}$ and view by each source modality, as in Figure 4. We can compare and contrast attention from different modalities to draw inference on the model's decision-making process. For example, in attention over the image, image- and text-sourced attention tends to highlight words within the image while attribute-sourced attention highlights people and faces. Attribute-sourced attention focuses on hashtags in the text. Text tends to self-attend to semantically significant words such as "awesome", "great", and "excited". One common example of a sarcastic Tweet is a screenshot or wall of text accompanied by enthusiastic punctuation and words. These tend to have attributes such as "screen" or "picture". Attention over these attributes by the text guidance vector suggests that perhaps the presence of enthusiasm of text generates interest in the fact that the image is a screenshot, as the combination of both is usually a sarcastic Tweet.

We can also examine the attention over the feature vectors of each modality, as in Figure 5. We can perhaps think of these values as a measure of importance of each modality, and this claim is tested in a later section. We can also generate confusion matrices for examples with the highest attention in each modality as in Figure 6. Combined with empirical observations from visualization, these suggest that the model relies on key words or phrases in the text to identify sarcasm ("really", "monday", and "memes" are such keywords) and looks for clues otherwise in images and attributes. Images and attributes often identify the presence of people in images which empirically indicates a non-sarcastic example (examples with "man" and "posing" attributes are strong indicators of this).

In the original paper, Cai et al. showcase several hand-picked attention visualizations and suggest that the model picks up on semantic "inconsistencies" between two models, and includes a picture of a dark day with "amazing weather" in the Tweet. The presence of the sarcastic screenshot archetype challenges this somewhat, as the model is unable to determine the semantics of the screenshot (it is able to determine the image is text, but the ResNet that runs on the image is incapable of determining what the words mean) and empir-

Figure 3: Attention visualization of an example from the Twitter dataset.

Figure 4: The same example with attention split by source modality.

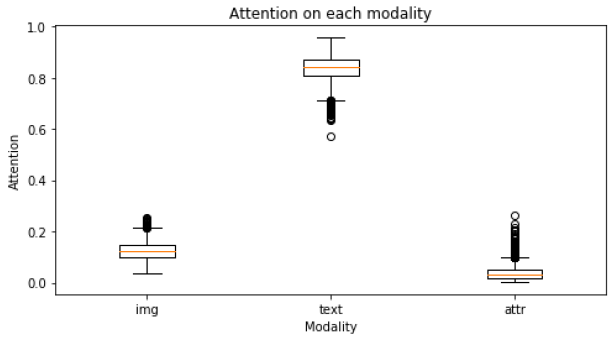

Figure 5: Distribution of attention over final modality vectors.

ically, even non-screenshot images included with sarcastic Tweets are nuanced in ways that humans understand in conjunction with the text (political figures or items with cultural significance), but are not picked up by the model through images or attributes. These examples are still correctly identified as sarcastic through the evaluation of text features, however.

2.1 Attention Permutation

Another experiment, inspired by Jain and Wallace, was to randomly permute one type of representational or modality fusion attention and analyze how it changed the model's predictions. For image, text, attribute, and modality attention, I define *impact* as $|p - p_a|$, where p is the probability output of the model and p_a is the output after randomly permuting attention type $a \in \{\text{img, text, attr, modality}\}$. Impact should be an approximate measure of the importance of that feature (the meaning of the modality attention is a little harder to interpret). For every example in the test dataset, I calculate the impact of 10 random permutations (5 for modality attention), and the average is the impact for that example. Figure 7 shows the distribution of impact for each attention type. Unsurprisingly, text has a significantly higher mean and is more variable than attribute and image, and image has a slightly higher mean than attribute. This looks similar to the distribution of modality attention (Fig. 5). A natural question arises: is modality attention (the weight of each modality in the classification) a good indicator of modality importance at an example level? Table 1 shows r values of a linear regression between modality fusion attention and permutation impact. There is a moderate correlation for text, but a weak correlation for image and attribute. An additional consideration is that 10 random permutations is a small number, particularly for images ($196! \approx 5e{+}365$ possible permutations), and increasing this number may reduce variability and increase correlations (I limited myself to 10 due to time and computational constraints).

Another interesting similarity to modality fusion attention is that the confusion matrix of the top 100 impacted examples for each modality (Figure 8) looks very similar to the most attended examples (Figure 6), where image and attribute were largely true negatives while text was true positives. Impact also allows us to view the confusion matrix for modality fusion attention itself, which is mostly true positives. One possible explanation for this is that the model has a somewhat strict region for classifying positives, which means that noise tends to bring down the classification probability for more positive classifications.

Modality	r
text	0.588
img	0.094
attr	0.286

Table 1: Correlation between modality attention and permutation impact

3 Future Work

Attention visualization is particularly powerful in Cai et al.'s model due to the presence of different modalities that interact through fusion. A multimodal fusion model including video or audio (Kumar and Vepa, 2020) could also provide interesting or insightful visualizations.

Additionally, such visualization could be used as a transparency tool for user-facing models, or as an exploration tool for academics and researchers. Creating an interactive, web-based visualization tool for a model could be an intermediate step towards interpretability, until we can coerce models to provide succinct explanations for their decisions.

4 Conclusion

Through a combination of example-level observation and quantitative aggregated analysis, attention over a multimodal model can provide a wide range of insights, although it is difficult to prove empirical hypotheses without deeper probing of the

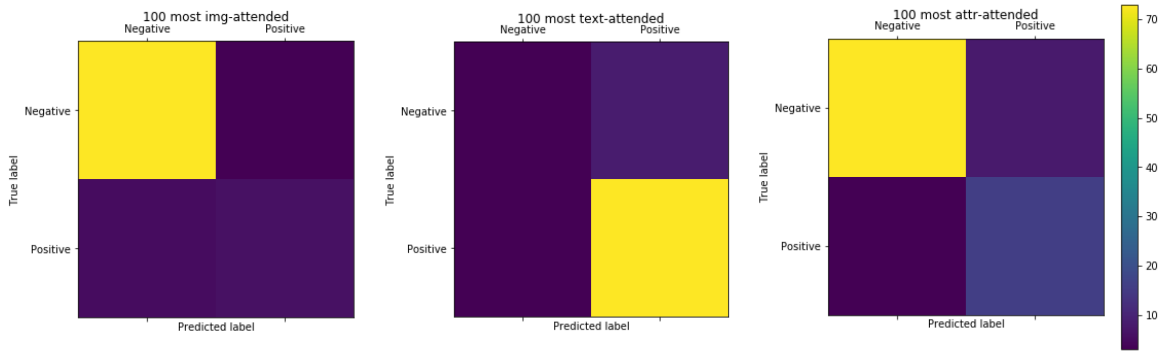

Figure 6: Confusion matrices for highest attended examples for each modality.

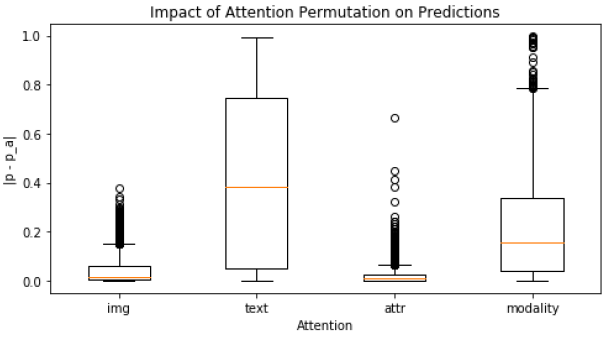

Figure 7: Distribution over examples of impact of attention permutation for each type of attention.

model. Additionally, attention visualization is an effective tool for non-technical personnel or end-users to understand how models make decisions, and their ability to do so is critical to create fair and transparent NLP-driven applications.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy, July. Association for Computational Linguistics.

D-Blue. 2020. pytorch-multimodal-sarcasm-detection. https://github.com/wrk226/pytorch-multimodal-sarcasm-detection.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *CoRR*, abs/2004.03685.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multimodal sentiment analysis. *CoRR*, abs/2003.01043.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, page 1050–1055. AAAI Press.

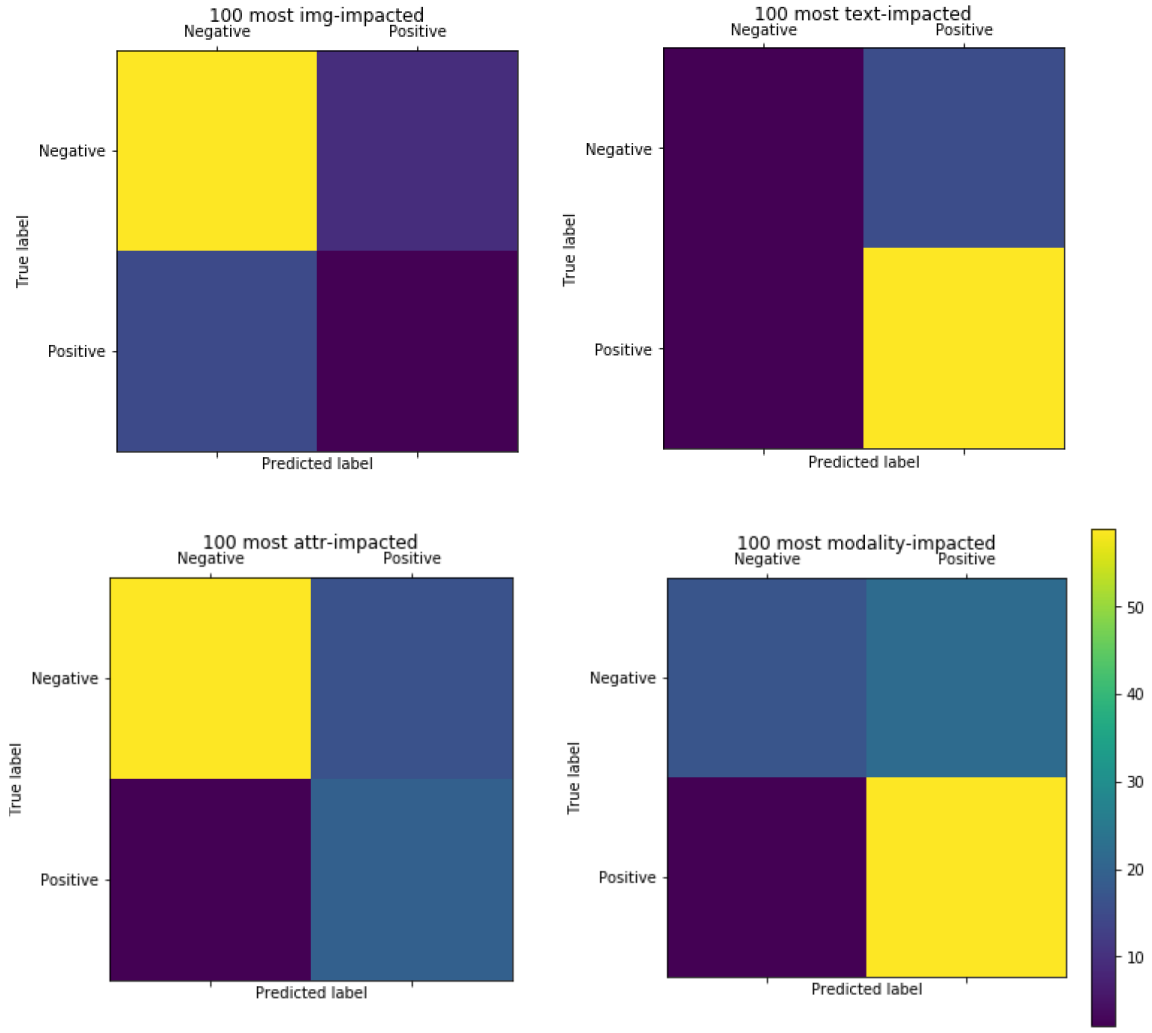

Figure 8: Confusion matrices for most impacted examples.