Paper Reading

Debangan Mishra

Summary

1. MemeX model tries to find the context of a meme given the meme itself (image and text components), a reference contextual document from which evidences are generated.
2. Model involves the Knowledge Enriched Meme Encoder to get a combined representation of the meme and context document, Meme Aware Transfer to get a combined representation of meme with the context document, Meme Aware LSTM to process output of MAT, followed by fully connected layers for final outputs.
3. Sentence embeddings are obtained from BERT Model and they are concatenated to get the embedding of the entire context.
4. External knowledge encoded via GCN and GMF

5. Pre-trained Graph Convolutional Network is used to represent the meme text which embeds common sense. Gated Multimodal Fusion block is combining these representations.

6. Meme aware transformer designed, where keys and values are conditioned with meme information and regulated by a sigmoidal gating mechanism. Query is by scaled dot-product-based attention. Multihead attention then generates meme representation.

7. Meme Aware LSTM have conventional features but have the values to the input and the gate are calculated with additional meme information. Hidden states of the LSTM are concatenated to get a combined representation.

8. The final content and meme representations are concatenated and passed to a feed-forward network that gives the likelihood of a sentence being an evidence for the meme.

9. We see that this model performs well in the task of identifying the evidence text and has metrics comparable and exceeding many models. Additionally, they observe that the multimodal models tend to perform better as compared to the unimodal ones in the task.

Strengths

1. The model is comprehensively creating a metric for assessing the models, professionally annotated dataset which enables more effective training and evaluation.
2. The models are modified to have meme representation encoded with them including the Meme Aware transformer and LSTM which enables multimodality.
3. All the models seem to be important to the performance, Advantage that authors have included the effect of removing the model in their paper

Weaknesses

1. Their models will not be able to detect memes outside of the context in which they are trained very well. The scope of the models seem to be fairly narrow.
2. There does not seem to be a method of mitigating textual inaccuracies or erraticness. This is also one of the flaws recognised by the authors.
3. An extremely expansive architecture, which might cause high computational burden.

Improvements

1. The models rely extensively on concatenation of input data which makes the process less intuitive as we are combining more spaces. A possibility is to experiment with non-concatenating methods (ie, reducing the usage of early fusion) such as dot products, or neural network based methods.
2. The model is not interpretable due to its complexity and high levels of abstraction as a result of which it is not explainable.
3. The model can be trained on more themes and on weaker evidences to increase its scope