SUMMARY

The paper describes how to detect prominent part of the context related to meem which explain the meme , not just image and text correlation, as meem is hiding a lot of meaning which is a lot of abstraction so unless the models tries to understand the context behind the text and how it is combined with the image to give off a meaning. The novelty lies in how the multimodal architecture detects the abstraction between the context and the visual image which is being shown .

NOTE :

- The input to that model are a text which describes the context and meme(image and text associated with image)
- Context is encoded with pretrained BERT and meme is encoded with multimodal encoder called KME(Knowledge enriched meme encoder)
    - Here KME consists of two part which are MMBT and GCN node
    - MMBT fused the image and text to get a single embedding
    - GCN trained over the ConceptNet to get a common sense enriched representation and incorporate external knowledge associated with them meme Text
- We use transformer based models called MAT(Meme aware transformer) to combine both the contextual and meem embeddings to get a single embedding
- The output obtained from MAT is then filtered specific to meme related context using MA-LSTM(Meme Aware LSTM) which is capturing the correlation between context and meme and decide what is more important/prominent part of obtained output from MAT
- Finally concatenating the output from MA-LSTM and KME will have completed contextual knowledge associated with meme and also the correlation of it with meme

The major advantages can be described as follows

- The important aspect observed here is that meaning conveyed by the meme is very much dependent on "***in which context it is used***" and what is actual meaning of the meme and how it is "***correlated to the context***" . this scenario is tackled here by taking in the both the meme and context as inputs to decrypt , understand and learn how the relation between the context and meem should be identified
- In Order to understand the hidden meaning in the memes we have make the models learn the different context along various aspects like sarcasm , harmfulness, trolling, abusiveness(it can be like gender, race or on historicals events etc) and for this to happen we need to have a data specifically dealing with such scenarios - (MCC, containing 3400 memes and related context, along with gold-standard human annotated evidence sentence-subset.
- The model to understand the contextual meaning deeply, it need's a method to represent the meaning inside the context efficiently and also need to ensure that there is no loss in meaning while converting the text to embedding and here that is taken care of by pretrained models called BERT which is very good in understanding the contextual meaning
- There should be a way to link the context and meme used over it to learn the correlation between them which is fulfilled by MAT & MA-LSTM

This paper has tackled the root cause where many multimodality based models for Q&A based analysis like VQA or contextual analysis based models were unable explain or decrypt the meme related data however I feel that there few drawbacks to this model on certain aspect which are follows

- Here for first steps stars from using pre trained BERT for contextual embedding and then MMBT for multimodal fusing embedding and then GCN to get common sense enriched text associated with image and then these two are again passed into MAT(transformer based) which is again connected to MA-LSTM to filter the embedding , so over all two Three transformer based models and one Graph convolutional model are use which is making the model too bulky

- MMBT itself learns to map image embeddings to an appropriate subspace of the text encoder’s input space, so why not use MMBT itself again for encoding image and the context together rather than just using another Pretrained BERT for contextual embeddings .
- I feel that MAT lacks explainability like how exactly its fusing the contextual embedding with the multimodal embedding for example Transformer based models like bert have several methods to interpret and explain the outputs, which aim to make their predictions more interpretable and transparent like Attention Visualization or Layer-wise Relevance Propagation (LRP) assigns relevance scores to each input token based on its contribution to the final prediction.