PreCog

MEMEX: Detecting Explanatory Evidence for Memes via Knowledge-Enriched Contextualization

Research Paper Summary

Introduction

Meme is disseminated with sarcasm or humour and acts as mean of propagating complex ideas through visual-lingual semiotics. Context of meme is critical to facilitate it's holistic assimilation camouflaging their intended meaning. MEMEX - on given a meme and context can predict the part of context which explains the meme. The idea also has multimodal applications for detecting pieces of evidences in a document providing context. Emotion analysis , visual-semantic role , and detection of hatred, sarcasm, trolling struggle when applied to memes due to contextual dependency.Memes are generally misinterpreted by the uninitialized due to their context

MEMEX - performs an 'evidence detection task' by learning cross-modal analogies and injecting contextual signals. Multimodal-abstraction and detection of contextual evidence forms the novel task. For cross-modal retrieval and vision-language pretraining, accurate measurement of cross-modal similarity is imperative.

MCC : Meme Context Corpus

3 stages for data collection due to unavailability of memes+context type datasets:
1) **Meme Collection** : Political, historical and English language ,also movies and geo-politics memes were focused due to availability of systematically documented information. Google Images and Reddit were used for their extensive and diverse search and multimedia presence.
2) **Context Document Curation** : Wikipedia to curate contextual corpus and also use community-based forums like Quora

3) **Annotation Process :** Annotators (urban social media vernacular) were given guidelines to annotate the evidence part in the context of meme called 'evidence sentences'. Cohen's Kappa was used to assess and got a score of 0.55 which finally came up to 0.72

Dataset Description : Max token length of ground truth evidence was 312 and max threshold for context was 512 tokens. An 80:10:10 splits leads to 3003 train , 200 validation and test sets. Each sample has a meme image , Context , OCR extracted meme text and ground-truth evidence sentence

MIME (MultImodal Meme Explainer)

MIME consists of a pretrained BERT encoder and pooled individual sentence representation to get unified context to encode the context and a multimodal encoder to encode the meme (image and text).

To address the linguistic abstractions besides factual knowledge , we design a Knowledge-enriched Meme Encoder (KME) that augments the joint multimodal representation of the meme.The paper proposes use of ConceptNet, a semantic network, and a pre-trained Graph Convolutional Network (GCN) trained on ConceptNet, to enrich meme representations with semantic characteristics and external common sense knowledge to enhance the comprehension and contextual mapping of memes.

Also describes a process to encode memes using a pre-trained MMBT model, resulting in a multimodal representation. External knowledge representation is obtained using Graph Convolutional Network (GCN) node representations from the meme text, which are then averaged to create a unified knowledge representation. A Gated Multimodal Fusion (GMF) block is utilized, employing meme and knowledge gates to modulate and fuse corresponding representations using trainable parameters.

Inspired by context-aware self-attention methods, the paper proposes a meme-aware multi-headed attention (MHA) mechanism to integrate multimodal meme information during self-attention computation. This results in a meme-aware Transformer (MAT) encoder, which computes cross-modal affinity for context representations conditioned on knowledge-enriched meme representations. Unlike conventional self-attention, meme-aware MHA generates key and value vectors conditioned on meme information

before using them for multi-headed attention-based aggregation.For detailed working of the Meme-Aware Transformer one needs to read the research paper.

MIME incorporates a recurrent neural network (RNN), specifically a Meme-Aware LSTM (MA-LSTM), a model that integrates meme information into sequential learning. MA-LSTM is inspired by previous work and incorporates meme representations during cell and hidden state computation using a gating mechanism. The architecture of MA-LSTM includes conventional steps for input, forget, output, and gate value computation, with additional input from meme representations for enhanced contextual understanding.

The final model concatenates enriched meme and context representations, passing them through a feed-forward layer to predict the likelihood of a sentence being valid evidence for a meme, with optimization performed using cross-entropy loss.

Bert , ViT (Unimodal) and Early-fusion , MMBT , CLIP , BAN , VisualBERT (Multimodal) are some of the baseline models for encoding memes and context representations

Experimental Results

Five independent runs on a thematically diversified test-set, followed by a comparison using standard metrics including accuracy, macro-averaged F1, precision, recall, and exact match score were performed.For partial match scenarios, precision, recall, and F1 are computed separately for each case before averaging across the test set. Additionally, basic image-editing operations are performed on meme images in MCC for optimal OCR extraction and noise-resistant feature learning.

The performance analysis reveals that unimodal systems, particularly text-based models like BERT, outperform image-based models such as ViT. This suggests that textual cues play a crucial role in modeling associations when the target modality is text-based, while purely image-based conditioning may not capture fine-grained correlations effectively. Multimodal models, on the other hand, either compete strongly or outperform unimodal ones, with MMBT achieving the highest F1 score of 0.7725. Models like MMBT and VisualBERT leverage pre-trained unimodal encoders and employ joint-modeling schemes for multiple modalities. However, MIME, a contextualization-based approach, demonstrates significant

improvements over the best baseline (MMBT) across various metrics, indicating its potential for enhancing performance through optimal contextualization.

After analysing detected evidences , we see that MIME correctly predicts relevant evidence in cases where MMBT fails to fully explain the meme, suggesting MIME's improved performance in providing more fitting evidence

Incremental assessment over the base model MMBT shows enhancements across various metrics when adding external knowledge-based cues through KME, MAT, and MA-LSTM. Removing MA-LSTM or MAT leads to performance drops across metrics, highlighting their significant contributions to MIME's efficacy. Replacing MAT with a standard Transformer-based encoder or MA-LSTM with a BiLSTM layer results in performance degradation, emphasizing the importance of systematic memetic contextualization for addressing MEMEX..

Noticeable challenges with abstract concepts and novel facts in ground-truth evidence and as partial predictions are observed, influenced by inductive biases related to concepts like presidential race and Jimmy Carter, with some instances of the model relying solely on embedded text without considering visual context. MIME achieves an exact match for 58.50% of cases, lacks predictions for 12.5%, provides partial matches in 14%, and makes incorrect predictions in the remaining 14%.

Conclusive Note

Challenges start from the MEMEX model's inability to grasp the complex abstractions inherent in memes, particularly when critical yet cryptic visual information lacks systematic integration with factual knowledge. Additionally, MIME struggles with insufficient textual cues for contextual associativity and may pick up spurious evidence due to lexical biases within the related context.

The above work introduces the MEMEX task and the MCC dataset, then proposes MIME, a novel modeling framework integrating knowledge-enriched meme representation with context via a multi-layered fusion mechanism, demonstrating efficacy through empirical examination and ablation studies while highlighting areas for improvement.