# Memotion Analysis through the Lens of Joint Embedding (Student Abstract)

**Nethra Gunti**[1*], **Sathyanarayanan Ramamoorthy**[1*], **Parth Patwa**[2], **Amitava Das**[3, 4]

IIIT Sri City, India[1]
UCLA, USA[2]
Wipro AI Labs, India[3]
AI Institute, University of South Carolina, USA[4]
{nethra.g18, sathyanarayanan.r18}@iiits.in, parthpatwa@ucla.edu, amitava.das2@wipro.com

## Abstract

Joint embedding (JE) is a way to encode multi-modal data into a vector space where text remains as the grounding key and other modalities like image are to be anchored with such keys. Meme is typically an image with embedded text onto it. Although, memes are commonly used for fun, they could also be used to spread hate and fake information. That along with its growing ubiquity over several social platforms has caused automatic analysis of memes to become a widespread topic of research. In this paper, we report our initial experiments on Memotion Analysis problem through joint embeddings. Results are marginally yielding SOTA.

## Text with Image Joint Embeddings

Distribution based compositional word embeddings like Word2vec, GloVe are popular in modern NLP. Consider the *king, queen* word vector analogy (Fig.1.), which shows how good these word embeddings are at capturing syntactic and semantic regularities in language. However, CNN based image embeddings fail to capture such contextuality (Fig. 1), since they are only source image dependant and do not capture corpus level distributions. Our work is focused towards textual grounding of images and to achieve distributional representation quite similar to word vectors.

## Memotion Analysis - Data and SOTA

We use the Memotion data (Sharma et al. 2020), consisting of 10K annotated memes, and OCR extracted text. Memotion analysis has 3 tasks- Task A: Sentiment Analysis (*positive, negative, neutral*). Task B: Emotion Analysis (sarcasm, humour, offense, motivation). Task C: Semantic Class Prediction of emotional sub-classes. Reported SOTA (F1) results on these tasks are 35.47%, 51.84%, and 32.35% respectively. ((Sharma et al. 2020) reports details of SOTA).

## Learning Joint Embedding

Learning text+image JE has gained significant momentum recently. Majority of them are based on some form of

*Equal contribution.



(a) The king, queen analogy

(b) An expected joint embedding(JE) space: king$_{JE}$- queen$_{JE}$ $\approx$ *boy*$_{JE}$ - girl$_{JE}$
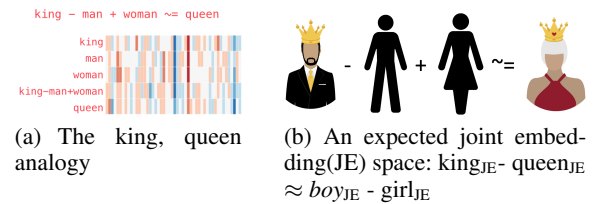
Figure 1: CNN based image embeddings are unable to capture contextuality like existing word embeddings.

Canonical Correlation Analysis (CCA), which finds similarities between two modalities by projecting them into one common vector space. We use two such existing models:

i) CLIP (Radford et al. 2021), uses a visual transformer as the image encoder and a transformer as text encoder, to generate JE embeddings. It was pre-trained on a large dataset consisting of over 400M images.

ii) Stanford's Joint Embedding (Kolluru 2019) (StanJE) uses VGG-19 and GLoVe to generate the image and text encodings respectively. They introduce a two-branch embedding network and a triplet loss function, which applies a margin based penalty, to obtain joint representation. It is pre-trained on Flickr30k (Plummer et al. 2016) dataset.

These models are trained with sentence-image pairs (SJM). So we re-train both the models with word-image pairs. Further, we fine tune the models using Memotion data+Flickr30k data to obtain JE.

### Word with Image Joint Embedding - (WJM)

The smallest meaningful unit language is word. So, anchored word representations along with image encoding should be the expected milestone for JE. We re-train both the models with word-image pairs from Flickr30K and Memotion data. To get meaningful words, we apply choose only those words belonging to specific Parts-of-Speech (POS) *noun, verb, adverb, adjective* categories. This results in a near-exponential increase in the dataset size, as every image is paired with multiple unique and significant words. Fig. **??**, shows that the obtained JE space is able to capture meaningful analogy: boy$_{JE}$- girl$_{JE}$ $\approx$ man$_{JE}$ - woman$_{JE}$.
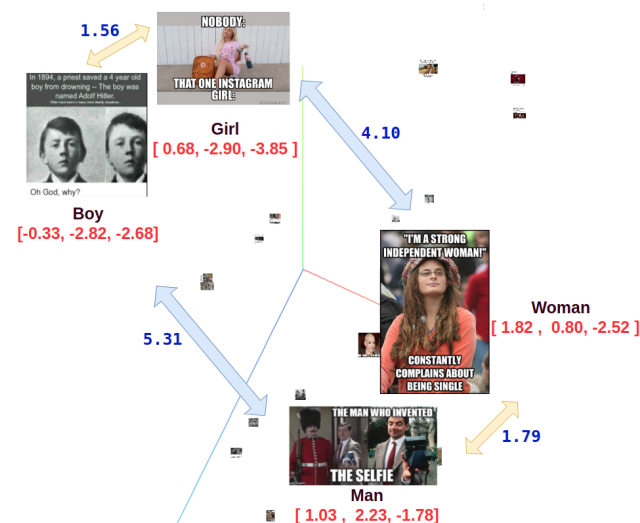
Figure 2: Visualising Analogy- emulating the king, queen analogy but instead using boy, girl, man, woman. boy$_{JE}$-girl$_{JE}$ $\approx$ man$_{JE}$ - woman$_{JE}$.

## Memotion Analysis using Joint Embeddings

Learning joint representations is a new research topic, similarly how to use such learned JEs for several downstream tasks - is an open territory for research exploration. CLIP allows zero-shot prediction and StanJE allows fine-tuning mechanism to use trained JEs for a particular task. We find that both the methods are not adequate for the Memotion task. Therefore, we propose new methods - Task A is a multi-class problem, so we use a Multi-Layer Perceptron (MLP) with 3 dense layers and a softmax output layer trained using categorical cross-entropy loss. For both the Task B and Task C, a Multi-Task Learning (MTL) framework, similar to (Samghabadi et al. 2020), is applied, which consists of four MLPs (one per sub-task). Binary cross-entropy loss is used for Task B and *motivation* emotion of Task C, since they are binary prediction problems. Rest of the sub-tasks in Task C are multi-class problems, so we use categorical cross-entropy loss. The code is available at https://github.com/NethraGunti/Memotion-Analysis-Through-The-Lens-Of-Joint-Embedding

## Results

Table 1 reports the performance of all the experiments. CLIP is a data hungry model, so it fails to perform well on datasets with less and diverse data. Hence, it performs poorly on Memotion Analysis tasks as compared to StanJE. WJM performs better than SJM in all the task. This validates our assumption that words are ideal choice for pairing with images for multimodal concept learning. The Word-Image Joint Embeddings obtained using the StanJE architecture yield the SOTA performance in all the categories.

| Models | Data | CLIP | StanJE |
|---|---|---|---|
| Sentiment Analysis- Macro-F1 Scores | | | |
| SJM | Flickr30k | 0.31 | 0.305 |
| | Memotion | 0.294 | 0.311 |
| | Combined | 0.3 | 0.284 |
| WJM | Flickr30k | 0.301 | 0.362 (+2%) |
| | Memotion | 0.31 | **0.373** (+5%) |
| | Combined | 0.307 | 0.354 |
| SOTA | 0.355 (Sharma et al. 2020) | | |
| Emotion Analysis- Macro-F1 Scores | | | |
| SJM | Flickr30k | 0.484 | 0.474 |
| | Memotion | 0.47 | 0.484 |
| | Combined | 0.476 | 0.423 |
| WJM | Flickr30k | 0.481 | 0.484 |
| | Memotion | 0.461 | **0.522** (+0.77%) |
| | Combined | 0.453 | 0.455 |
| SOTA | 0.518 (Sharma et al. 2020) | | |
| Semantic Classification- Macro-F1 Scores | | | |
| SJM | Flickr30k | 0.282 | 0.238 |
| | Memotion | 0.287 | 0.309 |
| | Combined | 0.236 | 0.309 |
| WJM | Flickr30k | 0.281 | 0.261 |
| | Memotion | 0.273 | **0.324**(+0.62%) |
| | Combined | 0.275 | 0.243 |
| SOTA | 0.322 (Sharma et al. 2020) | | |

Table 1: Results of Memotion Analysis using JE. Reported results are of 5-fold cross validation.

## Conclusion and Future Work

In this work we report our initial experiments to solve Memotion analysis through JE. Results are marginally yielding SOTA. Our major contribution could be summarised as: (i) To the best of our knowledge this is the first endeavor to solve Memotion analysis through JE. ii) JE techniques have so far only been used for either image captioning task or text-based image retrieval task. Memotion analysis is arguably a much more complex multi-modal task.(iii) We propose a different setup on how to use obtained JE to solve a multi-modal task. We carefully choose relatively simpler architectures like MLP and a simple version of MTL. The reason for such choices is to see and assess how useful JEs are - otherwise it could be argued that it is not the JE, but rather the architecture which is performing well. (iv) Results show word-image pairs are better alternatives than sentence-image pairs for JE training. There are many challenges in working with JEs, For Eg: Image and Text associated with a meme are not necessarily related yet they convey reasonable meaning. Future work should be towards such considerations.

## References

Kolluru, S. H. 2019. A neural architecture to learn image-text joint embedding. *Stanford web*.

Plummer, B.; et al. 2016. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *Transactions of the Association for Computational Linguistics*.

Radford, A.; et al. 2021. Learning transferable visual models from natural language supervision. *ICML*.

Samghabadi, N. S.; et al. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *TRAC2*.

Sharma, C.; et al. 2020. SemEval-2020 Task 8: Memotion Analysis–The Visuo-Lingual Metaphor! *SemEval*.