ACL Paper Summary: Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks
Xiaocui Yang, Shi Feng, Yifei Zhang, Daling Wang School of Computer Science and Engineering, Northeastern University, China

Problem Description:
Due to the rapid increase in popularity of smartphones, "multi-modal" social media posts have seen a rise in population. The term "mutli-modal posts" refers to posts that have both a text caption and an image to express the poster's intent. The goal of this paper is to determine the sentiment of multi-modal posts using Multi-channel Graph Neural Networks with Sentiment-awareness (MGNNS). Previous research on the topic saw models that only considered the sentiment of single posts but were not able to determine the global sentiment using co-occurrence characteristics from the dataset.

Prior Work:
In the field of multimodal sentiment analysis, multimodal polarity analysis and emotion analysis are often combined. In the past, traditional machine learning techniques have been able to tackle this task, and recently deep learning have seen good results. Specifically for videos, Wang et al implemented a TransModality technique which fused multimodal features with translation techniques. Also, Zhang et al mined information from unlabeled data using semi-supervised autoencoders. Additionally, Hazarika et al developed the MISA framework which makes a projection of modality to modality-invariant and modality-specific subspaces.

In the field of Graph Neural Networks (GNN), most of the achievement has come from text classification, recognition of multiple labels, and multimodal tasks. Traditionally, Text GCN, TensorGCN and TextLevelGNN have been used for text classification, but a baseline GNN has seen better performance and thus has seen rapid developments being made for new variants. Recently, Graph Convolutional Networks have been popular for tasks such as multimodal fake news detection from Visual Dialog and Visual Question Answering performed by Guo et al and Wang et al respectively. Also, Jiang et al modeled the visual dialogue cross-modal information with fine granularity using a Knowledge-Bridge Graph Network.

Unique Contributions:
The paper described three main areas of focus in terms of the development stages: encoding, multi-channel graph neural networks, multimodal interaction, and sentiment detection.

Starting with encoding, the paper obtains a text memory bank by using BiGRU by first generating an embedding vector using GloVe. Then, in the case of image modality, features are extracted from the objects and scenes because there are believed to be interdependencies between objects and scenes within a single image. The object extraction is done using YOLOv3, while the scene extraction is done using VGG-Place. Then, the memory banks for the objects and scenes are created using ResNet.

In terms of multi-channel graph neural networks, this is where the aper proposes its multi GNN module which consists of a Text GNN (TG) channel, Image GCN (IGO) channel and IGS channel. The Text GNN learns, through the Text Level GNN, text representations. The vocabulary for the graph was built by constructing edges between words when the number of co-occurrences of two words was greater than or equal to 2. Then, using globally shared matrices built around the vocabulary, edge weights are calculated. These weights are initialized using pointwise mutual information (PMI) and are adjusted/learned during training. This adjustment is done using message passing mechanism (MPM) based on its original representations and neighboring nodes.

In terms of the Image GCN, the paper uses this module to represent scene-object interdependence using IGX. The edges of IGX were built by first constructing a co-occurrence matrix which is also globally shared. Then the conditional probability for each node is calculated, and to account for simple correlation, a binary co-occurrence matrix is built. This is constructed as 1 if the conditional probability of the co-occurrence of two nodes is greater than or equal to some hyperparameter beta which can be tuned, and 0 otherwise. Then, it is worth noting that the central node has higher importance than that of the neighboring nodes, so the edge weights are calculated using the weighted co-occurrence matrix and the importance of the neighboring nodes. Finally, the nodes and edges are placed on the graph convolutional network. The main insight from this module is that stacking various GCN layers allows us to learn complex interdependence between nodes. Then, the paper uses multi-head attention to learn sentiment-awareness image representation because it is not possible to capture the relationship between nodes and sentiments.

In terms of multimodal interaction, the paper used a Multimodal Mutihead Attention Interaction module to learn the text modality and image modality interaction using multiple channels. This involves multiple models being fused together by concatenation.

Finally, in terms of sentiment detection, the SoftMax function is applied to the fused models into the fully connected layer of the graph network. So, the final equation used id $L^m = softmax(w^s R^m + b^s)$, where w^s, b^s are parameters of the fully connected layer.

Evaluation:
The paper compared their work to multimodal sentiment models which used the same modalities as well as unimodal models. In terms of the unimodal baselines, the text modalities were compared against CNN and Bi-LSTM. Also, BiACNN. Which combines the CNN and BiLSTM with an attention mechanism was used for comparison for text sentiment analysis. For image modality, OSDA was used for comparison, which is a model based on various views. SGN and OGN were also used for image sentiment analysis comparisons are they are graph convolutional neural networks based on scenes and objects.

In terms of the multimodal baselines, HSAN, a hierarchical semantic attentional network that is based around captions for images was used alongside MDSN, a deep semantic network with attention for multimodal sentiment analysis. Co-Mem and MVAN were also used as baseline models for comparison.

The paper used metrics such as Acc, and F1 score to compare the performance of the MGNNS against the baselines, and on all the data sets tested (MVSA-Single, MVSA-Multiple and TumEmo), the MGNNS had the best performance.

Conclusion:
I think this work is important because in the new age of technology and social media, a great deal of information is transferred to though multimodal posts. Whether it be for entertainment, education, or persuasion, most of the information we consume is from posts with both a picture and caption. With this increase in popularity of multimodal posts, there has also been an increase in the number of posts which could be classified as troll or just have an ambiguous meaning. The use of the model proposed by the paper can be very helpful in understanding the meaning of such posts and can be extended to social media and news platforms to limit the number of spam posts. This paper has received 20 citations on google scholar.