Indian Institute of Information Technology Kottayam

MULTIMODAL SENTIMENTAL ANALYSIS

BY

2020BCS0069 CHRISTO SOJAN

2020BCS0096 ANGATI BALA MURALI

2020BCS0184 KUMAR AMITANSHU

2020BCS0024 SYAM SIVADAS

Guided By,

DR. MANU MADHAVAN

INTRODUCTION

- Multimodal sentiment analysis simply deals with unlocking emotions Across Text, Audio, and Visual Data.

- **Definition**: Multimodal Sentiment Analysis is an advanced AI technique that goes beyond analysing text alone. It encompasses the interpretation of emotions expressed in multiple modes of data, such as text, audio, and visual content.

Significance of multimodal sentiment analysis

- Humans communicate with each other and express their emotions through various modalities, which mainly include text, visual and audio content.
- Combining modalities offers a richer emotional context for enhanced understanding.
- Some popular sentiment analysis applications include social media monitoring, customer support management, and analysing customer feedback. Using natural language processing techniques, machine learning software is able to sort unstructured text by emotion and opinion.

Key Components

- **Text Analysis**: Extracting sentiment from written content.
- **Audio Analysis**: Deciphering emotions conveyed through speech.
- **Visual Analysis**: Detecting emotions from images and videos.
- **Fusion Techniques**: Integrating insights from multiple modalities.
- **Machine Learning Models**: Algorithms for accurate sentiment analysis

Challenges and Considerations

- **Data Integration**: Unifying disparate data sources.
- **Data Quality**: Dealing with noise and inaccuracies in audio and visual data.
- **Complex Models**: Developing models that handle multiple modalities effectively.
- **Annotation and Labelling**: Labour-intensive process for training data.
- **Ethical Implications**: Ensuring privacy and mitigating bias.

Multimodal Sentiment Analysis: Addressing Key Issues and Setting up the Baselines

Indian Institute of Information Technology Kottayam

Modalities:

- The paper considers three modalities for sentiment analysis: text, audio, and visual.
- These modalities provide important cues for understanding emotions and sentiments in videos.

Indian Institute of Information Technology Kottayam

DATASET:

- The paper uses several datasets for evaluation. These include the MOUD dataset, which consists of product review and recommendation videos from YouTube, the MOSI dataset, which contains opinionated utterances from people reviewing various products, and the IEMOCAP dataset, which focuses on multimodal expressive dyadic interactions.

- MOUD: The Multimodal Opinion Utterances Dataset (MOUD) is a dataset that focuses on multimodal sentiment analysis.

- It consists of 80 product review videos in Spanish. Videos were found on the YouTube search page using the following keywords (translated to English): my favourite products, non recommended products, my favourite perfumes, recommended movies), non recommended movies, recommended books and non recommended books. Each video consists of multiple segments labelled to display positive, negative or neutral sentiment.

- It contains opinion utterances from online reviews, along with their corresponding text, images, and acoustic features. The dataset is designed to facilitate research in understanding sentiment in a multimodal context.

- IEMOCAP (Interactive Emotional Dyadic Motion Capture) is a widely used multimodal dataset for emotion recognition research. It consists of audio visual recordings of naturalistic dyadic conversations, where actors were instructed to engage in scripted conversations while expressing specific emotions.

- IEMOCAP provides several modalities of data, including audio, video, and text transcriptions. It contains approximately 12 hours of recordings, with a total of 10039 utterances, and is annotated with emotion labels using categorical and dimensional approaches.

Indian Institute of Information Technology Kottayam

ML/DEEP LEARNING MODELS USED

- The paper employs various models for feature extraction and sentiment classification.
- For textual feature extraction, convolutional neural networks (CNN) are used.
- For audio and visual feature extraction, 3D-CNN and openSMILE are utilized.
- The authors also employ a bidiredectional LSTM (bc-LSTM) to capture context from surrounding utterances. SVM and bc-LSTM fusion methods are used for sentiment classification.

CONVOLUTIONAL NEURAL NETWORK

- A Convolutional Neural Network (CNN) is a type of deep learning algorithm specifically designed for analysing visual data.
- It is widely used in computer vision tasks such as image classification, object detection, and image recognition.
- CNNs are inspired by the organization of the animal visual cortex, where individual neurons respond to specific regions of the visual field.

- CNNs consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers.

- Convolutional layers apply filters to input images to extract features, while pooling layers down sample the feature maps to reduce computational complexity.

- Fully connected layers are used for classification or regression tasks.

OPENSMILE

- OpenSMILE is an open-source toolkit for audio and speech processing.

- It provides various features for extracting low-level descriptors (LLDs) from audio signals, which can be used in applications such as speech emotion recognition, speaker recognition, and speech synthesis.

LSTM

- LSTM stands for Long Short-Term Memory, which is a type of recurrent neural network (RNN) architecture. It is designed to overcome the limitations of traditional RNNs in capturing and remembering long-term dependencies in sequential data.

- LSTMs are widely used in various applications such as natural language processing, speech recognition, time series analysis, and more. They have the ability to retain information over long periods of time, making them well-suited for tasks involving sequential data.

RESULTS OBTAINED

- The results show that the bc-LSTM fusion method consistently outperforms the SVM fusion method across all experiments.
- The audio modality performs better than the visual modality in both the MOSI and IEMOCAP datasets.
- The text modality plays a crucial role in both emotion recognition and sentiment analysis, with its unimodal performance being substantially better than the other modalities.
- The fusion of modalities shows more impact in emotion recognition than in sentiment analysis. The paper also discusses the performance of the models in speaker-exclusive experiments, where the models have to classify emotions and sentiments from speakers they have never seen before. The results show that the speaker-exclusive setting yields inferior results compared to the speaker-inclusive setting.

Indian Institute of Information Technology Kottayam

CURRENT STATUS OF MSA AND FUTURE SCOPE

- The paper serves as a benchmark for future research in multimodal sentiment analysis. It highlights the importance of considering context in classification, the impact of different modalities, and the generalizability of multimodal sentiment classifiers.

- The authors suggest future work in extracting semantics from visual features, exploring relatedness of cross-modal features and their fusion, and incorporating contextual dependency learning in the models to overcome limitations. The paper also emphasizes the need for benchmark datasets and further studies on generalizability across datasets of the same language.

SUMMARY

- The research paper explores multimodal sentiment analysis using deep-learning architectures for sentiment classification. It considers text, audio, and visual modalities for understanding emotions in videos. The authors use CNN, 3D-CNN, openSMILE, and bc-LSTM models. The bc-LSTM fusion method outperforms SVM in accuracy, while audio and text modalities play crucial roles. The paper emphasizes the need for context, different modalities, and generalizability of multimodal sentiment classifiers. Future work should focus on extracting semantics from visual features and incorporating contextual dependency learning.

Multi-Level Attention Map Network for Multimodal Sentiment Analysis

Indian Institute of Information Technology Kottayam

MODALITIES:

- The research paper focuses on MSA tasks with texts and images from user-generated content.

- It considers the challenges of aligning multimodal data that may be posted separately by users.

DATASET:

- The research paper evaluates the proposed model on three public datasets:
- MVSA-Single, MVSA-Multi, and Multi-ZOL.
- These datasets contain multimodal samples with sentiment labels.

- **MVSA SINGLE:**

- The mvsa-single dataset is a benchmark dataset designed for the task of Multi-View Sentiment Analysis (MVSA).
- It consists of a collection of sentences from different domains, such as product reviews, movie reviews, and news articles. Each sentence is annotated with sentiment labels indicating whether the sentiment expressed in the sentence is positive, negative, or neutral.

MULTI-ZOL

- Multi-ZOL is a dataset designed for bimodal sentiment classification of images and text.

ML/DEEP LEARNING MODELS USED

- The proposed model is the Multi-Level Attention Map Network (MAMN), which consists of three modules: multi-granularity feature extraction, multi-level attention map generation, and attention map fusion.
- The model utilizes techniques such as attention mechanisms, gated mechanisms, and multi-task learning.

MULTI-LEVEL ATTENTION MAP NETWORK

- The Multi-Level Attention Map Network is a deep learning model that incorporates attention mechanisms to enhance object detection in images.
- It aims to improve the accuracy of object localization by leveraging multi-level features and attention maps.

- **Multi-granularity feature extraction (MANN):**

- It is a technique used in machine learning and computer vision to extract features from images or data at multiple scales or levels of granularity. It involves analysing and extracting features at different resolutions or spatial levels to capture both local and global information.

- MANN can be useful in various tasks such as object recognition, image classification, and segmentation, where features at different scales are important for accurate analysis.

- By considering multiple granularities, MANN can capture fine details as well as overall context, leading to improved performance in various applications.

- **Multi-level attention map generation:**

- It refers to the process of generating attention maps at different levels or scales within an image. These attention maps highlight the salient regions or areas of interest in an image, allowing models to focus on important features during tasks such as object detection, image segmentation, or image captioning.

- One popular approach for multi-level attention map generation is the use of convolutional neural networks (CNNs) combined with spatial pyramid pooling. This technique involves extracting features at different scales using CNNs and then pooling the features at multiple levels to generate attention maps.

- **Attention map fusion:**

- It is a technique used in computer vision and image processing that combines multiple attention maps to generate a single fused attention map. Attention maps represent the saliency or importance of different regions in an image. By fusing multiple attention maps, the resulting fused attention map can capture the salient regions from different perspectives or modalities.

- The fusion of attention maps can be done using various methods such as averaging, max pooling, weighted averaging, or using deep learning-based approaches. These methods aim to leverage the complementary information from different attention maps to enhance the overall saliency detection or object localization performance.

Indian Institute of Information Technology Kottayam

RESULTS OBTAINED

- The experimental results show that the proposed MAMN model outperforms state-of-the-art methods in terms of accuracy and effectiveness for both document-based and aspect-based MSA tasks.

- The model achieves significant improvements in sentiment classification performance on the evaluated datasets.

CURRENT STATUS OF MSA:

- The research paper contributes to the field of multimodal sentiment analysis by proposing a novel model that addresses the challenges of noise reduction, feature extraction, and correlation capture. However, the current status of MSA is an ongoing research area, and there is still room for further advancements and improvements in the field.

FUTURE SCOPE OF MSA:

- The research paper suggests several future directions for MSA. These include exploring more advanced fusion methods, investigating the impact of different modalities on sentiment analysis, and exploring other application areas for MSA such as product marketing and public opinion monitoring. Additionally, the paper highlights the importance of addressing the challenges of noise reduction and feature extraction in MSA tasks.

SUMMARY

Indian Institute of Information Technology Kottayam

- This research paper presents a Multi-Level Attention Map Network (MAMN) for multimodal sentiment analysis (MSA).
- The paper addresses the challenges of noise reduction, feature extraction, and correlation capture in MSA tasks. The proposed MAMN model consists of three modules: multi-granularity feature extraction, multi-level attention map generation, and attention map fusion.
- The model utilizes techniques such as attention mechanisms, gated mechanisms, and multi-task learning. Experimental results on three public datasets demonstrate that the MAMN model outperforms existing methods in terms of accuracy and effectiveness for document-based and aspect-based MSA tasks.

Multimodal Sentiment Analysis via RNN variants

Indian Institute of Information Technology Kottayam

MODALITIES:

- The modalities considered in the research are text, audio, and video.
- The authors analyse the utterances in videos to extract sentiment information.

Indian Institute of Information Technology Kottayam

DATASETS:

- The authors use the CMU-MOSI dataset.

- The Multimodal Corpus of Sentiment Intensity (CMU-MOSI) dataset is a collection of 2199 opinion video clips. Each opinion video is annotated with sentiment in the range [-3,3].

- It is a widely used dataset for multimodal sentiment analysis and emotion recognition tasks. It consists of video clips from the TED Talks dataset, where each clip is accompanied by various modalities such as audio, video, and text.

- The dataset is rigorously annotated with labels for subjectivity, sentiment intensity, per-frame and per-opinion annotated visual features, and per-milliseconds annotated audio features.

- The dataset is split into training and testing sets, with 62 and 31 videos, respectively.

ML/Deep Learning Models used

- The authors propose four variants of Recurrent Neural Networks (RNNs) for sentiment analysis: GRNN, LRNN, GLRNN, and UGRNN.

- These models are based on LSTM and GRU architectures. Attention networks are also used for fusing the modalities.

- GRNN: GRNN stands for Generalized Regression Neural Network. It is a type of neural network that is primarily used for regression tasks. GRNN is known for its simplicity and ability to approximate complex non-linear functions.

- In GRNN, the training data is used to create a network of nodes, where each node represents a training sample.

- During the testing phase, the input is compared to the training samples, and the output is calculated based on the similarity between the input and the training samples.

- LRNN: LRNN stands for Long-Range Neural Networks. It is a type of neural network architecture that is specifically designed to handle long-range dependencies in sequential data.

- Traditional recurrent neural networks (RNNs) often struggle with capturing long-term dependencies.

- LRNNs address this issue by incorporating mechanisms to propagate information over longer distances in the sequence.

- They achieve this by introducing additional connections or memory cells that can store and retrieve information from previous time steps.

- GLRNN: GLRNN stands for Gated Linear Recurrent Neural Network. It is a type of recurrent neural network architecture that incorporates gating mechanisms to control the flow of information within the network.

- GLRNN has been used in various applications such as natural language processing, speech recognition, and time series analysis. It has been shown to achieve competitive performance in tasks that involve modeling sequential data.

- UGRNN: UGRNN stands for Unbounded Gated Recurrent Neural Network. It is a type of recurrent neural network (RNN) architecture that is designed to model sequential data.

- UGRNNs have been shown to perform well in various tasks, such as language modeling, machine translation, and speech recognition.

Indian Institute of Information Technology Kottayam

RESULTS OBTAINED:

- The research paper reports the accuracy achieved by the different RNN variants on the CMU-MOSI dataset.

- GRNN performs best for text, GRNN and GLRNN perform best for audio, and UGRNN performs best for video.

- After fusing the modalities, LRNN and GLRNN achieve the best results for multimodal sentiment analysis, with an accuracy of 78.05%.

Indian Institute of Information Technology Kottayam

CURRENT STATUS OF MSA

- The research paper contributes to the field of multimodal sentiment analysis by proposing novel RNN variants and evaluating their performance on the CMU-MOSI dataset.

- It demonstrates the effectiveness of using multimodal data for sentiment classification.

Indian Institute of Information Technology Kottayam

FUTURE SCOPE OF MSA:

- The authors suggest that the proposed approach can be further explored for other applications such as credibility analysis.
- They also mention the potential for using other types of data, such as contextual, crowd-source, and relationship information, to improve the analysis.
- The paper highlights the broader context of the research, including multimedia information processing, web mining, and artificial intelligence, indicating the potential for further advancements in multimodal sentiment analysis.

Summary:

- This research paper focuses on multimodal sentiment analysis, which involves classifying sentiment using different forms of data such as text, audio, and video.

- The authors propose four variants of recurrent neural networks (RNNs) - GRNN, LRNN, GLRNN, and UGRNN - for analysing the utterances in videos and classifying sentiment.

- They conduct experiments on the CMU-MOSI dataset and show that their approach achieves better sentiment classification accuracy than existing methods on individual modalities and also after fusing the modalities using attention networks.

Dense Fusion Network with Multimodal Residual for Sentiment Classification

Indian Institute of Information Technology Kottayam

MODALITIES:

- The DFMR framework integrates three modalities: language, acoustic speeches, and visual images. Each modality is processed separately in the modality-specific module to capture modality-specific interactions. The features extracted from each modality are then fused together in the dense multimodal fusion module to capture cross-modal interactions.

Indian Institute of Information Technology Kottayam

DATASETS:

- The authors conduct experiments on four public benchmark datasets: CMU-MOSI, ICT-MMMO, YouTube, and IEMOCAP.

Indian Institute of Information Technology Kottayam

ML/DEEP LEARNING MODELS USED

- The DFMR framework consists of four modules:
- The modality-specific module, dense multimodal fusion module, multimodal residual module, and sentiment classification module.
- The modality-specific module uses a bidirectional gated recurrent unit (Bi-GRU) to model modality-specific interactions.
- The dense multimodal fusion module fuses the modality-specific features using dense fusion (DF) blocks.
- The multimodal residual module integrates the fused features from the DF blocks to capture hierarchical multimodal fusion information. Finally, the sentiment classification module uses fully-connected layers to predict the sentiment category.

Indian Institute of Information Technology Kottayam

CURRENT STATUS

- The paper presents the DFMR framework as a solution for multimodal sentiment analysis. The current status of the work is that it has been evaluated and shown to outperform existing approaches.

Indian Institute of Information Technology Kottayam

FUTURE SCOPE:

- As for future scope, further research can explore the application of the DFMR framework in other domains and datasets. Additionally, improvements and extensions to the framework can be explored to enhance its performance and address specific challenges in sentiment analysis.

Indian Institute of Information Technology Kottayam

SUMMARY:

- The paper proposes the DFMR framework for multimodal sentiment analysis. It integrates language, acoustic speeches, and visual images using dense fusion and multimodal residual modules. The framework achieves state-of-the-art performance on benchmark datasets and provides a promising solution for sentiment analysis using multimodal data.

Fusion-Extraction Network for Multimodal Sentiment Analysis

Indian Institute of Information Technology Kottayam

MODALITIES:

- The research focuses on multimodal sentiment analysis, which involves analysing both visual and textual information in social media data.

DATASET:

- The authors use two public multimodal sentiment datasets, namely MVSA-Single and MVSA-Multiple. MVSA-Single contains 5129 text-image pairs from Twitter, while MVSA-Multiple has 19600 text-image pairs labelled by three annotators.

MVSA-Single dataset:

- MVSA (Multiple Viewpoint Semantic Annotation) is a dataset that is commonly used for evaluating the performance of semantic annotation algorithms.
- It consists of multiple datasets, each representing a different viewpoint or perspective on the same concept.
- The purpose of MVSA is to assess the ability of algorithms to handle different viewpoints and generate accurate annotations.

ML/DEEP LEARNING MODELS

- The proposed model is called Fusion-Extraction Network (FENet). It incorporates an interactive information fusion mechanism and an information extraction mechanism. It aims to extract and fuse information from multiple input modalities such as images, text, and audio to perform tasks like classification, generation, or retrieval.

- The interactive information fusion mechanism uses fine-grained attention to learn cross-modality fused representations.

- The information extraction mechanism employs gated convolutional layers to extract informative features and generate expressive representations.

RESULTS

- The experimental results show that FENet outperforms existing state-of-the-art methods on the two multimodal sentiment datasets. The model achieves higher accuracy and F1 scores compared to baseline methods such as SentiBank & SentiStrength, CNN-Multi, DNN-LR, MultiSentiNet, and CoMN.

CURRENT STATUS OF MSA

- The research paper contributes to the field of multimodal sentiment analysis by proposing a novel model that effectively utilizes the relationship between visual and textual information. It demonstrates improved performance compared to existing methods, indicating progress in the field.

FUTURE SCOPE OF MSA

- The research opens up possibilities for further advancements in multimodal sentiment analysis. Future work could explore more sophisticated fusion mechanisms, extraction techniques, and attention mechanisms to enhance the understanding of multimodal data and improve sentiment analysis performance. Additionally, the proposed model could be extended to other domains and datasets to evaluate its generalizability.

SUMMARY

- The document proposes a Fusion-Extraction Network (FENet) for multimodal sentiment analysis. The network utilizes an interactive information fusion mechanism to learn visual-specific textual representations and textual-specific visual representations. It also incorporates an information extraction mechanism to filter redundant parts and extract valid information from the multimodal representations. Experimental results on two public multimodal sentiment datasets show that FENet outperforms existing state-of-the-art methods.

CONCLUSION

- In wrapping up our exploration of Multimodal Sentiment Analysis, we've journeyed through an exciting realm where emotions transcend the boundaries of text, audio, and visual data. As we conclude this presentation, let's reflect on the key takeaways:

- **Holistic Understanding**: Multimodal Sentiment Analysis empowers us to gain a more complete understanding of human emotions, capturing nuances that single-modal analysis cannot.

- **Real-World Impact**: The applications are vast, ranging from improving customer experiences and content optimization to enhancing mental health support and human-computer interactions.

- **Complex Challenges**: While promising, this field comes with its share of challenges, including data integration, quality, model complexity, annotation efforts, and ethical considerations.

- **Continual Evolution**: With advanced AI models, innovative data collection methods, and a heightened focus on ethics, the future of Multimodal Sentiment Analysis is bound to bring breakthroughs and innovations.