Multi-modal Sentiment Analysis

Emil Dalgaard Moroder
Department of Computer Science at Aalborg University
emorod19@student.aau.dk

Abstract

This paper addresses multi-modal sentiment analysis using text and image pairs, aiming to develop methods for efficiently capturing a shared compact latent space to improve sentiment classification. We propose two methods: An Autoencoder (AE)-based method and an Attention-based (AB) method, designed to integrate multi-modal data. Experiments on the MVSA and Meme datasets show that our methods achieve performance comparable to state-of-the-art models like VisualBERT and CLIP, but all models struggled with overfitting, likely due to noisy, low-quality data. This highlights the need for higher-quality datasets and advanced noise reduction techniques. Future work will explore hybrid approaches combining AE and AB methods with sophisticated fusion strategies to enhance multi-modal representations.

1 Introduction

In recent years, numerous fields have seen a rapid increase in recorded data, encompassing a wide range of different modalities for a given task (Sui et al., 2023). However, typically machine learning models are developed with a single modality, which leaves additional data unused. Consequently, it would be beneficial to have a model that can utilize multiple modalities to capture a more complete understanding of the data and the relationships between the modalities (Chen and Luo, 2019).

This is relevant in several domains, e.g. in the medical context, where patient data consists of MRIs, X-rays (images), bloodwork, ECGs (time series) and clinical notes (text). Using a single modality for a patient only provides limited information about their state. As a result, models utilizing multiple modalities have been applied to get a better understanding of the patient, where it has been used to predict breast cancer (Liu et al., 2022) and more general disease diagnosis and prognosis. It typically performed better than the uni-modal variant, however, introducing multiple modalities also came with a cost of bias and noise (Cui et al., 2023). Data has inherent noise, and thus if an additional modality only contains redundant information then adding the new modality to the pipeline will introduce noise which could lead to a worse performance.

Another critical domain is *Sentiment Analysis (SA)* which this paper will *focus* on due to data availability. The goal of SA is to predict the emotional tone of the data, e.g. the sentiment a user feels towards a product, which can be a user review of the product (Gandhi et al., 2023). In addition, it can be a social media post containing images, text, video and sound. In this use case, it is important to use all the modalities in combination to gauge the sentiment, since a hateful post could be written with normal text but a sexist image (Chen and Luo, 2019).

In this paper, we study the task of multi-modal sentiment analysis with text and image pairs due to their prevalence. However, the proposed methods can be applied to N modalities with any type of data. Despite the potential advantages, effectively integrating multi-modalities remains a challenging problem, where the uni-modal counterpart often has superior performance. The primary goal of this project is to develop efficient methods that capture a compact and representative latent space for multi-modal data which can improve the performance of sentiment analysis.

These challenges are addressed by proposing two methods. A method that utilizes *Autoencoders (AEs)* (Said et al., 2017) and an *Attention-based* method (Yang et al., 2021) (Yu et al., 2020). They are designed to be efficient and to generate compact latent representations that contain information from all modalities.

We validate the methods on two multi-modal datasets: a Tweet dataset *MVSA* (Niu et al., 2016) and a Meme dataset (Javaid, 2024). Both data consist of image-text pairs with a sentiment. The methods are compared against uni-modal approaches and state-of-the-art image-text models such as *VisualBert* (Li et al., 2019) and *CLIP* (Radford et al., 2021), which demonstrates that proposed methods can result in competitive performance while being less complex.

By addressing the challenges of multi-modal data integration, the aim is to provide methods that can be used as a foundation for future research and application in the domain of multi-modal sentiment analysis. Moreover, the aim is to identify limitations and reasons for the lack of multi-modal methods. Which we hope will uncover insights that can aid the research to develop more suitable multi-modal approaches.

The remainder of the paper is organized as follows. Section 2 reviews multi-modality learning representation methods. Section 3 will introduce the proposed methods. Section 4 will describe datasets and the experimental setup. Section 5 will present the results from the experiments and analyze them. Lastly, Section 6 will conclude and suggest directions for future work.

2 Related Work

Multi-modal classification

Multi-modality classification focuses on integrating multiple modalities into a shared discriminatory space. The difficulty is to identify redundant and complementary information (Xu et al., 2019). The general framework can be seen in Figure 1, where we consider N modalities (e.g. images, text...). The datasets can be defined as:

$$\begin{aligned}\mathcal{D} = \{&(m_{1,1}, m_{1,2}, \ldots, m_{1,N}, y_1),\\ &(m_{2,1}, m_{2,2}, \ldots, m_{2,N}, y_2),\\ &\ldots,\\ &(m_{M,1}, m_{M,2}, \ldots, m_{M,N}, y_M)\}\end{aligned} \tag{1}$$

where we have M data points and $m_{M,N}$ denotes the Mth data point of modality N. In this definition, every data point contains all modalities which is not always the case. Each modality has its own model that is used to extract salient features. This model consists of an encoder, and an optional classifier, which is the difference between *early* and *late* fusion (sometimes the latter also appends the predicted class probabilities) (IV et al., 2021). The individual salient features are defined as $\phi_{m_1}, \phi_{m_2}, \ldots, \phi_{m_N}$, hereafter a fusion network combines the features into a shared space

$$z = \{\phi_{m_1} \oplus \phi_{m_2} \oplus \cdots \oplus \phi_{m_N}\} \tag{2}$$

where $\oplus$ can be any operator that joins the latent representations. The most common choice is *concatenation* and *addition*, however, it can be arbitrarily complex. Given two vectors $\mathbf{a} = [a_1, a_2, \ldots, a_n]$ and $\mathbf{b} = [b_1, b_2, \ldots, b_m]$. The concatenation of $\mathbf{a}$ and $\mathbf{b}$, denoted $\mathbf{a} \oplus \mathbf{b}$, is given by $\mathbf{c} = [a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m]$ where $\mathbf{c}$ is the concatenated vector. The addition operator is defined as traditional vector addition, which requires the vectors to be the same size (IV et al., 2021)

The entire pipeline can be split into the five following stages: *Preprocessing*, *Feature Extraction*, *Fusion*, *Primary Learning* and *Classification*, where each stage is an important part that plays a role in the final performance. Preprocessing should be considered for each modality and be performed with domain expertise. The feature extraction can be any model e.g. traditional approaches like Random Forests and dimensionality reduction, however, it is typically done by *Deep Neural Networks (DNNs)* such as *Convolutional Neural Networks (CNNS)* for images, and *Recurrent Neural Networks (RNNs)* for textual and time series data (IV et al., 2021). As mentioned, fusion combines the extracted features, where three strategies exist: *early*, *late* and *cross-modal*. Early combines the features before without a learning network, whereas late fusion has an individual classifier for its respective salient features before it is fused. Lastly, cross-modal has its individual learning model distribute information to the other learning networks (IV et al., 2021).

Uni-modal Sentiment Analysis

Standard methods for text SA typically embed the text with bag-of-word or TD-IDF, which is fed to a standard classifier such as *Support-Vector-Machines (SVMs)*, *Naive Bayes* or an MLP. More advanced *deep learning (DL)* networks have better performance but require more data (Nandwani and Verma, 2021).

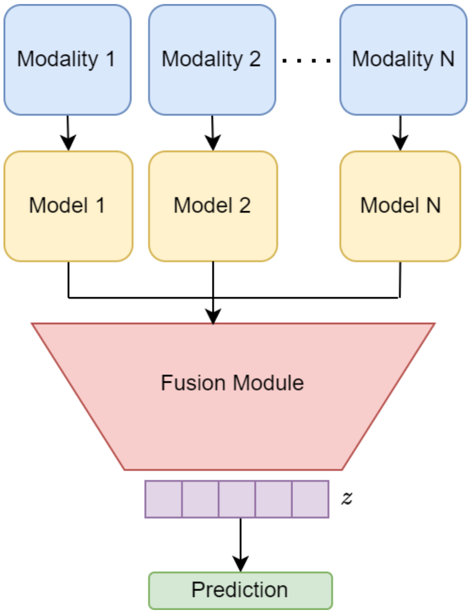

Figure 1: The general framework for multi-modality classification with N modalities. It shows how each modality has its own feature extractor that is fused into z which is the input to a classifier.

One such prominent DL method is *BERT* (Bidirectional Encoder Representations from Transformers), which achieves state-of-the-art performance in text classification. It uses a transformer-based architecture in a bidirectional manner, which captures the context of the text from both directions. This results in a context-rich text embedding (Devlin et al., 2018).

Image SA is typically performed with CNNs due to them easily learning the hierarchical features of raw images. A famous architecture that has achieved state-of-the-art performance and is often used as a baseline is the *ResNet* (He et al., 2015)

Multi-modal Sentiment Analysis

A method proposed by (Zhang et al., 2020) uses a Denoising AE (DAE) for the text as they argue text is very noisy. To extract image features they use a VAE-ATT, which is a Variational AE that uses attention. Hereafter, they propose a novel fusion method known as CFF-ATT, which calculates an attention matrix M that captures the correlation between the text embedding E and the image embedding G. This matrix M is then used to compute the attention-weighted features of the image which is denoted by J. Hereafter, E and J are concatenated and go through a fully connected layer to produce the final fused output (Zhang et al., 2020).

VisualBERT is another method for integrating text and images. It is mostly used for image captioning, where it locates regions of the images that are associated with words, however, it can also be applied to SA. The text input consists of the Bert text embedding and a pre-computed image embedding such as ResNet. These in combination are treated as sequences where it uses attention to learn how the image is contextually related to the text. This results in an information-rich shared representation that can be used for a classifier (Li et al., 2019).

CLIP (Contrastive Language-Image Pre-training) is a state-of-the-art image-text classifier that is designed to learn a joined representation. It revolutionized this by having a contrastive learning goal, where it takes a batch of image-text pairs and learns to minimize the distance between the correct pairs and maximize the distance for incorrect pairings. CLIP is pre-trained on a large and diverse dataset that allows it to easily be fine-tuned to various downstream tasks such as SA (Radford et al., 2021).

In summary, there exist various methods for multi-modal SA that can generate a powerful embedding, however, it remains a challenge to generate a representative embedding efficiently.

3 Methodology

Before dwelling into the proposed method architectures, we will first investigate the overall pipeline which is shown in Figure 2. The first grey box shows how a multi-modal dataset consisting of text, images and labels is split into training, test and validation (70-20-10). The individual modalities are preprocessed accordingly. The next box model architecture, where the text is embedded by a Bert-model followed by a text-encoder. The image is only embedded by an image cnn-encoder (however, one could also use a ResNet for an initial embedding). The encoded modalities are fused and fed to a classifier. This model is fit on the training set while validating on a separate set to ensure that it is not overfitting. The trained model is lastly evaluated on the unseen test set. Hereafter, we will explore the individual steps and the proposed methods.

Data preprocessing

The preprocessing for the different modalities remains important and is performed in the following manner. The text is cleaned by:

1. Converting to lowercase to ensure uniformity
2. Convert URLs to a unique placeholder
3. Replace user mentions (e.g @username is converted to $< USER >$)
4. Expand hashtags (converting #MachineLearning into machine learning).

The desired goal is to have a contextual text embedding, wherein there exist various strategies such as *GloVe* or BERT. We choose to convert the text into a sequence of tokens using a BERT-tokenizer. The sequence of tokens is fed to a pre-trained BERT model. It outputs a set of contextualized embedding for each token in the input sequence. Hereafter, we use [CLS] token embedding, which is used for aggregation of information from the whole sequence, as this results in a contextual full-sentence embedding.

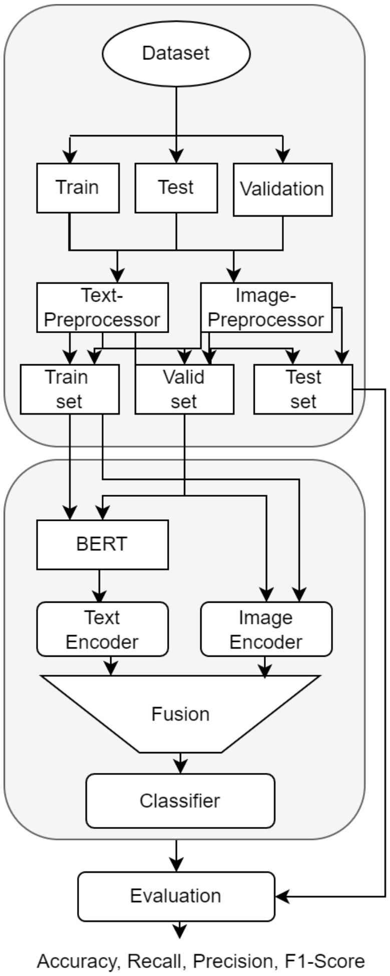

Figure 2: The overall architecture is illustrated through two boxes. The first box illustrates how the data is split and preprocessed. The second box shows the model architecture which is fitted on the training set. Lastly, the trained model is evaluated on the test set.

The image is processed in a simple manner, where it is resized to a predefined dimension $(3, 224, 224)$ and each pixel value is normalized. Additionally, one could use a pre-trained *ResNet* to get a high-quality image embedding. This ResNet embedding was only used for VisualBert and thus was not included in the pipeline shown previously.

Autoencoder-Based Multi-Modal Learning

In this method, the goal is to use AEs to integrate multiple modalities into a shared latent representation, that is compact, discriminatory of representative of all modalities. The main idea is to have modality-specific AEs that learn to produce compact features. The individual features are fused into a shared representation z, where the separate decoders have to use the fused features to reconstruct their original representation. Since z is constrained to be small, it ensures that representation is compact while containing information from all modalities. This method has a dual-objective training strategy, where it optimizes the reconstruction loss and classification loss.

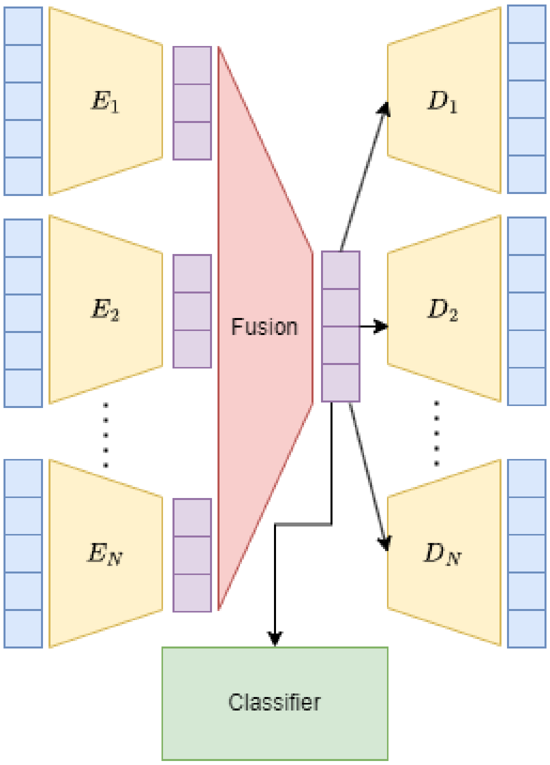

Figure 3: The multi-modal Autoencoder architecture shows the overall structure of the network. Each modality is embedded into a latent representation that is fused into a shared representation, hereafter each modality has its own decoder that uses the shared representation to reconstruct the original input. The fused input is also used by a classifier

The overall architecture can be seen in Figure 3 which illustrates the aforementioned strategy. Each modality i (e.g. text, images...) has its own encoder E_i. We can define the reconstruction loss for the jth data point for the ith modality accordingly:

$$MSE_i = (x_j - \hat{x_j})^2 = (E_i(m_{i,j}) - G(z_j))^2 \tag{3}$$

which is the *Mean Squared Error (MSE)* between the original input x and the reconstruction $\hat{x}$, and z is the fused representation. Moreover, the MSE for the jth data point for all modalities is:

$$MSE = \sum_{i=1}^{N} (x_{i,j} - \hat{x}_{i,j})^2 \tag{4}$$

The AE is crafted to handle the specific type of data (e.g. CNN for images). The encoder maps the original modality input to a latent representation x of size D_ϕ, which is defined to be the same for each modality to ease the fusions step.

Furthermore, it has to be mentioned that the AEs can be arbitrarily complex (VAEs, DAEs...), however, we kept them simple for efficiency. At first, individual classifiers for the latent representation of modality i were trained before the fusion step, meaning that $(N + 1)$ classifiers were trained accordingly. C_i is the ith classifier for modality i, which was trained to minimize the cross-entropy between $\mathcal{L}_{C_i} = \text{CE}(C_i(x), y) = \text{CE}(\hat{y}, y)$, but training these classifiers before the fusion step did not result in accuracy gain and was thus excluded.

We implement two fusion methods, namely, *concat* and *addition*. The first method concatenates each modality's latent representation and is the input to a linear layer. This can be described as:

$$h_{concat} = [\phi_{m_1} \oplus \ldots \oplus \phi_{m_N}]$$

and

$$z = Wh_{concat} + b$$

The second fusion method is the *addition*, which works as follows:

$$h_{add} = [\phi_{m_1} + \ldots + \phi_{m_N}]$$

and

$$z = Wh_{add} + b$$

This ensures that the fused representation z has the same latent size as the individual latent modalities and can be decoded.

The training procedure uses the dual-objective (reconstruction and classification), where the overall loss can be defined as

$$\mathcal{L} = \sum_{i=1}^{N} \lambda_i MSE_i + \lambda_0 CE(\hat{y}, y) \tag{5}$$

where $\lambda = [\lambda_0, \lambda_1, \ldots, \lambda_N]$ are hyperparameters that can be used to weigh the importance of each reconstruction loss for the modality and the classification loss. The classification loss is the typical cross-entropy defined as $CE = -\frac{1}{m} \sum_{j=1}^{m} y_j \log(\hat{y}_j)$, where y is the ground-truth label and $\hat{y}$ is the predicted label. This loss is optimized using the Adam optimizer.

Attention-Based Multi-Modal Learning

The goal of this method is to use attention to learn inter-modal relationships, which should create a latent space that focuses on the most essential parts of each modality in the context of the other modalities.

Similarly, the network architecture has a unique encoder for each type of modality that learns compact features of size D_ϕ. The latent space is forced to be the same size for each modality in order to easily apply attention to find the inter-modal relationships. The goal is to learn compact features for each modality while learning the relationships and having a discriminatory latent space.

This method will be introduced using two modalities and afterwards extended to N modalities. The full architecture with two modalities is illustrated in Figure 4. The individual encoders learn the compact features for each modality, and the attention between the individual latent features is calculated, where we learn the attention weights. This can be described mathematically as follows.

Given two modalities X_1 and X_2 with their encoded representations $h_1 \in \mathbb{R}^{D_\phi}$ and $h_2 \in \mathbb{R}^{D_\phi}$, where D_ϕ is the size of the latent space:

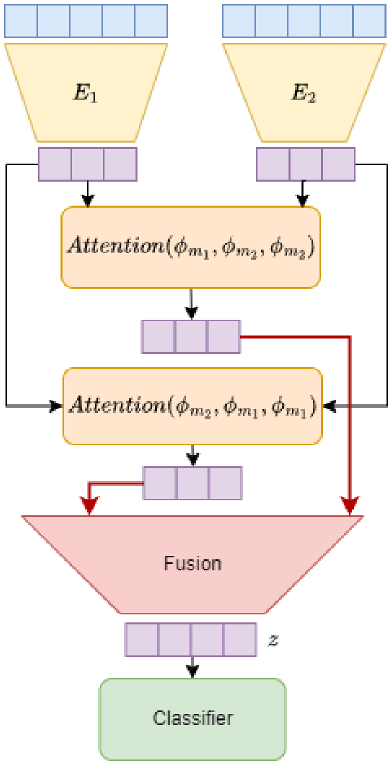
E_1
E_2
$Attention(\phi_{m_1}, \phi_{m_2}, \phi_{m_2})$
$Attention(\phi_{m_2}, \phi_{m_1}, \phi_{m_1})$
Fusion
z
Classifier

Figure 4: The architecture of the multi-modal attention-based method that is shown for two modalities, where it has a network that learns how modality 1 is related to modality 2, and also in the opposite direction. The attended features are afterwards fused and used for a classification network.

- Compute the attention score:

  $$a_{1,2} = \text{softmax}\left(\frac{h_1 W_Q (h_2 W_K)^T}{\sqrt{D_\phi}}\right)$$

  where W_Q and W_K are learnable weight matrices.

- Apply the attention score to get the attended feature:

  $$h_{1,2} = a_{1,2}\, h_2$$

  The attended feature $h_{1,2}$ captures the relevant information from h_2 in the context of h_1.

Similarly, $h_{2,1}$ is calculated which is afterwards fused into $z = \text{fusion}(h_{1,2}, h_{2,1})$. This method implements the same fusion methods as defined in the prior network. The fused representation is likewise the input to a classification network where the latent representation before the classification can be used as the common compact representation.

The objective of this method is to optimize the classification loss (cross-entropy), which is defined as:

$$\mathcal{L} = CE(\hat{y}, y) \tag{6}$$

which is identical to the classification loss in the previous method.

This method can be extended given N modalities with their encoded representation $h_i \in \mathbb{R}^{D_\phi}$ for $i = 1, 2, \ldots, N$ we would:

1. Compute the pairwise attention scores for each pair of modalities (i, j)

   $$a_{i,j} = \text{softmax}\left(\frac{h_i W_Q (h_j W_K)^T}{\sqrt{D_\phi}}\right)$$

2. Compute the attended features for each pair

   $$h_{i,j} = a_{i,j}\, h_j$$

3. Fuse the attended features

   $$z = \text{fusion}(h_{1,2}, h_{1,3}, \ldots, h_{i,j}, \ldots, h_{N-1,N})$$

however, this would require somewhat expensive calculations and thus the first method might be more useful for a larger number of modalities.

4 Experiments

Implementation Details

The models were implemented using PyTorch with the following hardware and software. It was trained on a low-performing *GPU*: NVIDIA GeForce GTX 980. With the following software: *Operating System*: Windows 10, PyTorch 2.2.2+cu121 and Python 3.11

Experimental Setup

1. **Hyperparemeters**:
   - Batch size: $[32, 64]$
   - Learning Rate: $\eta : 0.001$
   - Epochs: 5 to 15 (early stopping if model overfits)
   - Latent encoding size D_ϕ
   - λ parameters (AE method)

2. **Training Procedure**:
   - Optimized using the *Adam* optimizer
   - Dataset split into training, validation and test set (70-10-20 split)
   - Early stopping to prevent overfitting

3. **Evaluation Metrics**:
   - *Accuracy*, *precision*, *recall* and *F1-score* to evaluate the performance
   - Since the datasets have multiple classes, we evaluate the macro-metrics (the mean across classes), which can be calculated as follows:

     $$\text{Macro-Metric} = \frac{\text{Metric}_1 + \ldots, +\text{Metric}_N}{N} \tag{7}$$

     where N is the number of classes in the dataset, and the Metric is the evaluation metrics mentioned above.

Datasets

MVSA: Sentiment Analysis on Multi-view Social Data (Niu et al., 2016). This consists of around 20.000 samples of image-text pairs with a sentiment. Each sample is a tweet (social media post) which was *negative*, *neutral* or *positive*. It was annotated by three reviewers. The most frequent sentiment was used as the overall sentiment. The following IDs were removed $['3151','5995','3910']$ due to corrupted files.

The Meme dataset (Javaid, 2024). The dataset consists of 6992 samples of memes. Each meme contains an image and the text written on the meme (OCR). E It contains 6992 pairs of images and text of a meme (an image with text that has a humorous or sarcastic element). Each meme is *very-negative*, *negative*, *neutral*, *positive* or *very-positive*. This dataset only had a single annotated and was thus used for the overall sentiment. Only one sample was removed, which was ID: 5119 due to the image being corrupted.

Baseline Models

The following models are used as baseline:

- *Uni-Text*: Consists of the CLS BERT embedding which is fed to a classifier
- *Uni-Image*: is a CNN for image classification of similar complexity to the CNN encoder in the multi-modal methods
- *ResNet*: a more complex state-of-the-art image classification network (He et al., 2015)
- *CLIP*: learns joint text-image representations through contrastive learning, which is fine-tuned to our datasets (Radford et al., 2021)
- *VisualBERT*: uses transformers to associate the visual ResNet embedding with text tokens, which is also fine-tuned to our datasets (Li et al., 2019).

5 Results

The following tables are used to compare the proposed methodologies with the baseline models. There are two tables for each dataset and the difference is the batch size. Due to model-complexity *ResNet*, *CLIP* and *VisualBert* were only evaluated on batch-size $4-8$. AE-CAT(32) means that it is an Autoencoder-based method, that uses concatenation for fusion, and the latent size is 32. AB refers to the attention-based method.

Table 1 and Table 2 show the performance metrics of AE method for the MVSA dataset. At first glance, it seems most models perform similarly and no real gain is achieved by integrating multiple modalities. By viewing D as the entire domain of all data points, and C_{text} and set of data points correctly identified by the uni-text, and C_{image} being the correctly identified by the image-classifier. Thus the maximum theoretical accuracy can be defined as:

$$\frac{|C_{\text{text}} \cup C_{\text{image}}|}{|D|}$$

and it showed that by integrating both modalities we could potentially get an accuracy of 73.65% and that the multi-modal methods only capture around 81% of the theoretical accuracy. However, it was seen that the proposed multi-modal methods correctly predicted around 150 samples (around 4% of the test set) that neither uni-modal could predict, which could indicate that the shared latent representation results in new information that was not present in the uni-modal perspective.

Model	Accuracy	Recall	Precision	F1 Score
Uni-Text	**0.6432**	0.4797	0.5524	0.4996
Uni-Image	0.5917	0.3481	0.5042	0.3196
ResNet	0.5978	0.401	0.4269	0.3829
AE-Cat(32)	0.6361	0.4365	0.5340	0.4309
AE-Cat(64)	0.6340	0.4774	0.5244	0.4875
AE-Cat(128)	0.6366	0.3917	0.6166	0.3801
AE-Add(32)	0.6136	0.3382	**0.6549**	0.2647
AE-Add(64)	0.6121	0.4378	0.5057	0.4507
AE-Add(128)	0.6251	0.4131	0.4939	0.4151
AB-CAT(32)	0.6131	0.4287	0.5177	0.4433
AB-CAT(64)	0.6310	0.3940	0.5224	0.3860
AB-CAT(128)	0.6220	0.3618	0.5566	0.3213
AB-Add(32)	0.6338	0.4732	0.5430	0.4918
AB-Add(64)	0.6358	0.3762	0.3832	0.3489
AB-Add(128)	0.6327	0.3706	0.3822	0.3395
CLIP	0.6399	**0.5238**	0.5270	**0.5162**
Visual-Bert	0.6126	0.3333	0.2042	0.2533

Table 1: Performance comparison of models with the MVSA dataset with batch size 32. The best value is in bold and second best is underlined.

Model	Accuracy	Recall	Precision	F1 Score
Uni-Text	**0.6470**	0.4549	**0.6272**	0.4717
Uni-Image	0.5715	0.3709	0.4253	0.3703
ResNet	0.5978	0.401	0.4269	0.3829
AE-Cat(32)	0.6468	0.4604	0.5505	0.4683
AE-Cat(64)	0.6202	0.4498	0.5140	0.4654
AE-Cat(128)	0.6157	0.4554	0.5292	0.4734
AE-Add(32)	0.6297	0.4544	0.5420	0.4733
AE-Add(64)	0.6368	0.4481	0.5269	0.4630
AE-Add(128)	0.6412	0.4307	0.5492	0.4379
AB-Cat(32)	0.6378	0.5195	0.5245	0.5140
AB-Cat(64)	0.6386	0.4773	0.5243	0.4736
AB-Cat(128)	0.6338	0.3976	0.5366	0.3884
AB-Add(32)	0.6407	0.4939	0.5301	0.4782
AB-Add(64)	0.6409	0.4452	0.5634	0.4589
AB-Add(128)	0.6378	0.4069	0.5735	0.4046
CLIP	0.6399	**0.5238**	0.5270	**0.5162**
Visual-Bert	0.6126	0.3333	0.2042	0.2533

Table 2: Performance comparison of models with the MVSA dataset with batch size 64. The best value is in bold and second best is underlined.

From these performance metrics, the overall best-performing models were uni-text and CLIP, but the highest macro-precision was achieved by the proposed AE-ADD(32) method. VisualBERT resulted in surprisingly low scores and only predicted a single class (the majority class, which is positive tweets) for the entire dataset, which resulted in poor scores due to the ill-defined macro precision, recall and F1-score. This is problematic due to the interest lies in identifying hateful/negative tweets, and as a result, I tried random oversampling methods and a class-weighted loss function to penalize misclassifications of the minority classes. This resulted in better classification of the negative tweets but overall worse metrics. One should consider the desired goal and accordingly choose a strategy, e.g. by using a different sampling method or a class-weighted loss function.

Model	Accuracy	Recall	Precision	F1 Score
Uni-Text	0.4496	0.2026	0.1734	0.1324
Uni-Image	0.4356	0.2000	0.0871	0.1214
ResNet	**0.4574**	0.1997	0.2124	0.1321
AE-Cat(32)	0.4378	0.2015	**0.2473**	0.1251
AE-Cat(64)	0.4235	0.1959	0.1220	0.1284
AE-Cat(128)	0.4263	**0.2037**	0.1559	0.1567
AE-Add(32)	0.4328	0.2006	0.1499	0.1329
AE-Add(64)	0.4399	0.2033	0.2303	0.1303
AE-Add(128)	0.4328	0.1989	0.1269	0.1226
AB-Cat(32)	0.4356	0.2000	0.0871	0.1214
AB-Cat(64)	0.4356	0.2000	0.0871	0.1214
AB-Cat(128)	0.4356	0.2000	0.0871	0.1214
AB-Add(32)	0.4356	0.2000	0.0871	0.1214
AB-Add(64)	0.4356	0.2000	0.0871	0.1214
AB-Add(128)	0.4356	0.2000	0.0871	0.1214
CLIP	0.4027	0.2013	0.1508	**0.1693**
Visual-Bert	0.4356	0.2000	0.0871	0.1214

Table 3: Performance comparison of models with the Meme dataset with batch size 32

Model	Accuracy	Recall	Precision	F1 Score
Uni-Text	0.4346	**0.2104**	0.2016	**0.1698**
Uni-Image	0.4356	0.2000	0.0871	0.1214
ResNet	**0.4574**	0.1997	0.2124	0.1321
AE-Cat(32)	0.4020	0.1935	0.1713	0.1524
AE-Cat(64)	0.4278	0.2006	0.1803	0.1425
AE-Cat(128)	0.4256	0.1979	**0.2345**	0.1330
AE-Add(32)	0.4292	0.2053	0.1966	0.1588
AE-Add(64)	0.4313	0.2031	0.1597	0.1475
AE-Add(128)	0.4192	0.1973	0.1451	0.1431
AB-Cat(32)	0.4356	0.2000	0.0871	0.1214
AB-Cat(64)	0.4356	0.2000	0.0871	0.1214
AB-Cat(128)	0.4356	0.2000	0.0871	0.1214
AB-Add(32)	0.4356	0.2000	0.0871	0.1214
AB-Add(64)	0.4356	0.2000	0.0871	0.1214
AB-Add(128)	0.4356	0.2000	0.0871	0.1214
CLIP	0.4027	0.2013	0.1508	0.1693
Visual-Bert	0.4356	0.2000	0.0871	0.1214

Table 4: Performance comparison of models with the Meme dataset with batch size 64

Table 3 and Table 4 showcase a similar pattern as the previous dataset, but a surprise was ResNet that had 2% better accuracy than all the other models. It can also be observed that many models had the same performance which is an accuracy of 0.4356 and an F1-score of 0.1214. This was due to these models only predicting the majority class. Every AB-method could only predict a single class and seemed to not capture any signal, and as a result, additional experiments were conducted. I created an additional network which had the goal of creating a common representation of the encoded features. Thus there were two main networks: one network that fused the attention features and one network that fused the encoded features. This method did not result in any performance increase. The AE-based method did not suffer the same fate. This method had competitive results with the highest macro-precision score of 0.2345.

In summary, all models including the proposed methods, performed comparably in terms of the metrics, and the choice of fusion method nor the size of the shared embedding z did not significantly impact the result.

Approximately 4% of samples were only correctly predicted by the proposed multi-modal models but not by the maximum theoretical uni-modal approach, which highlights the potential of the multi-modal approach to capture a more nuanced embedding.

Despite the competitive results, it seems that introducing multiple modalities introduces a bunch of noise into the dataset, which leads to overfitting the training set. This is especially problematic in the real world since datasets tend to be messy and noisy, and thus injecting multiple modalities might make it more likely that the model captures more noise.

The proposed method has a straightforward and non-complex architecture that can easily be applied to any domain with any modalities. It was demonstrated that the proposed methods could be competitive with a non-overly complicated model architecture.

6 Discussion

The experiments revealed that the proposed methods: AE and AB achieved performance comparable to state-of-the-art models like VisualBert and CLIP. This demonstrates the potential for the simplistic approach to multi-modal sentiment analysis. Despite the competitive results, the proposed methods were simple and maintained low complexity compared to other methods, which highlights the efficiency. Moreover, it was shown that the models overfitted to the data without capturing much signal, which had an impact on all the models. This could indicate that the data was of low quality with inherent noise. This further highlights, the problem of multiple modalities, wherein each modality brings noise into the equation.

Overfitting

One critical observation was the fact that all models (even the uni-modal) quickly overfit the datasets without capturing much signal. This could be the result of various factors, e.g. the dataset being small, low quality and noisy. If this is true, it could explain why the multi-modal approach did not gain any significant performance increase by adding a new modality.

This presence of noise can significantly hinder the learning process of neural networks, meaning the model struggles to generalize and makes it difficult for a model to learn meaningful patterns, which is only exaggerated in the multi-modal view.

Limitations

The proposed multi-modal methods performed comparably to state-of-the-art models like CLIP and VisualBERT. VisualBERT was limited due to the nature of the model architecture as it associated certain tokens with regions in the images. The issue is that the input of the model is a sequence of text tokens and a sequence of regions in the images. However, there entire image was only treated as a single sequence, and thus it might have been difficult to create a high-quality integrated embedding. The fusion strategy did not influence model performance, however, more advanced fusion strategies were not implemented. One critical limitation is the quality of the datasets. The fact that if each modality is very noisy, then combining them makes it even more difficult for a model to learn to generalize. Another factor was the class imbalance which could be combated with other sampling techniques and a weighted loss function, but this was tested and did not result in better performance.

Future Work

The research suggests that a large high-quality dataset is required to evaluate whether or not the model architecture can capture more signal than the uni-modal approaches. There is a need for improved data quality or methods that can combat noisy data. Moreover, more advanced fusion strategies, AEs and hyperparameter settings (the λ parameters from the AE method) need to be investigated.

In addition, a new method that integrates both of the proposed methods with more advanced fusion strategies could be interesting. This would consist of AE networks that independently process each modality, learning compact and representative latent features and another network that uses attention to learn the interactions between the modalities. There would afterwards be two fusion networks. One that fuses the latent features and one that fuses the interaction features. Each would be optimised for a classification task. Afterwards, these embeddings are fused by a third network in combination with a classifier. This hybrid approach is much more complex, however, it leverages both strengths of the proposed methods which could result in a more comprehensive multi-modal representation, by having multiple networks focus on different elements. Adding this complexity with low-quality datasets would probably not result in any performance increase, and thus the need for higher-quality multi-modal datasets is pivotal.

7 Conclusion

In conclusion, this paper presents two methods for multimodal learning representation that can be used for sentiment analysis: an Autoencoder-based method and an Attention-based method. The primary goal was to develop efficient techniques that can capture compact and representative latent spaces for N modalities. Through extensive experiments on two text-image sentiment datasets: MVSA and meme, we demonstrated that the proposed methods can achieve competitive performance compared to state-of-the-art multi-modal approaches such as VisualBert and CLIP while maintaining lower complexity. It also has similar performance to unimodal approaches, where it was shown that the multi-modal approach revealed new unique correct classifications which highlight the potential of using multiple modalities. However, it was only able to achieve around 80% of the theoretical maximum of both uni-modal models.

Our findings highlight several significant challenges that need to be investigated further for future research: Data Quality and Noise in Multimodal Integration. The datasets appear to be of low quality with noise present in both modalities. This noise overwhelmed the models, which made it incredibly difficult to identify meaningful signal. Moreover, each modality brings its own noise, and once combined this noise is amplified, which makes it even harder for multi-modal approaches to perform effectively. This highlights the need for higher-quality datasets and potentially advanced noise reduction techniques.

References

Feiyang Chen and Ziqian Luo. 2019. Sentiment analysis using deep robust complementary fusion of multi-features and multi-modalities. *CoRR*, abs/1904.08138.

Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori Coburn, Keith Wilson, Bennett Landman, and Yuankai Huo. 2023. Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: A review. *Progress in Biomedical Engineering*, 5, 03.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

William C. Sleeman IV, Rishabh Kapoor, and Preetam Ghosh. 2021. Multimodal classification: Current landscape, taxonomy and future directions. *CoRR*, abs/2109.09020.

Hammad Javaid. 2024. 6992 labeled meme images dataset. `https://www.kaggle.com/datasets/hammadjavaid/6992-labeled-meme-images-dataset`. Accessed: 2024-06-09.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, and Y. Peng. 2022. A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. *IRBM*, 43(1):62–74.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling*, page 15–27.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Ahmed Ben Said, Amr Mohamed, Tarek Elfouly, Khaled A. Harras, and Z. Jane Wang. 2017. Multimodal deep learning approach for joint EEG-EMG data compression and classification. *CoRR*, abs/1703.08970.

Jing Sui, Dongmei Zhi, and Vince D Calhoun. 2023. Data-driven multimodal fusion: approaches and applications in psychiatric research. *Psychoradiology*, 3:kkad026, 11.

Naixin Xu, Wenliang Mao, and Guiguang Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 371–378.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2021. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026.

Jianfei Yu, Jing Jiang, and Rui Xia. 2020. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Kang Zhang, Yushui Geng, Jing Zhao, Jianxin Liu, and Wenxiao Li. 2020. Sentiment analysis of social media via multimodal feature fusion. *Symmetry*, 12(12).