Cross Modal Classification with Text and Image Data

Dongri Siddarth
Dept of CSE NIT Warangal

***Abstract*—In today's digital landscape, the coexistence of text and images is pervasive, particularly evident on the internet. Whether browsing through social media feeds, online articles, or e-commerce platforms, it's common to encounter textual descriptions accompanied by visual representations. This integration of text and images presents a fertile ground for classification tasks, where the goal is to categorize items based on their attributes or characteristics. In this study, we embark on a journey to explore the synergy between textual descriptions and visual representations for classification purposes. Recognizing the inherent value in leveraging both modalities, we investigate two primary methods of combining text and images to enhance classification accuracy. These methods involve early fusion, where textual and visual information are integrated at the outset of the classification process, and late fusion, where the modalities are processed sepa rately before their outputs are combined. To further refine our approaches and improve performance, we employ stacking techniques, which involve the integration of multiple classifiers to leverage diverse perspectives and enhance robustness. By stacking classifiers trained on different combinations of features or representations, we aim to capture a more comprehensive understanding of the data and im prove classification accuracy. My experimentation unfolds on the UPMCFood-101 dataset, a challenging benchmark renowned for its complexity and noise. This dataset, which comprises diverse food images with associated textual descriptions, mirrors real-world scenarios where multimodal data is prevalent and classification tasks are inherently challenging. The results of our study reveal the efficacy of our approach, particularly the early fusion technique combined with stacking. By integrating information from both modalities early in the classification pipeline and leveraging stacked classifiers, we achieve superior performance compared to previous methods. This finding underscores the importance of multimodal fu sion in addressing the nuances and complexities of real-world classification tasks, where textual and visual information often complement each other.**

I. Introduction

In the realm of classification tasks, the primary objective is to categorize items into distinct groups, facilitating the organization and understanding of vast amounts of data. Image classification, a prominent area within this field, involves train ing computers to recognize the contents of images, such as identifying whether a picture depicts a cat or a dog. Convolutional Neural Networks (CNNs) have emerged as powerful tools for this purpose, as they excel at extracting features and understanding the visual content of images.However, real-world data often comprises more than just images; it frequently includes accompanying text de scriptions or tags. This blend of images and text, known as multimodal data, presents a richer and more complex source of information. Consider the sce nario of online shopping, where product listings not only feature images but also contain textual descriptions detailing features, specifications, and other rele vant information.my work delves into the realm of multimodal data using the UPMC Food-101 dataset as a focal point. Initially, our efforts concentrate on developing robust image classification models capable of accurately identifying the contents of images within this dataset. Leveraging the capabilities of CNNs, we train these models to recognize various food items depicted in the images with high accuracy. Subsequently, we turn our attention to the textual component of the dataset, aiming to extract meaningful information from the associated text descriptions. This involves natural language processing techniques to understand and categorize the textual content related to each image accurately. Here we integrate both the image and text information to create a unified multimodal classification system. By combining insights from both modalities, we aim to enhance the overall accuracy and effectiveness of our classification model. This integrated approach allows us to leverage the complementary strengths of image and text data, leading to more nuanced and precise categorization of items within the dataset. 1. Image Classification: Initially, the researchers develop and test various techniques to classify images within the dataset accurately. This step aims to achieve the highest level of accuracy in classifying images into predefined cate gories. 2. Text Classification: Subsequently, attention is directed towards the textual part of the dataset. The researchers aim to create an algorithm capable of accurately classifying text descriptions associated with each image into the correct class. 3. Multimodal Fusion: Finally, the researchers propose a technique to com bine the outputs of the image classifier and text classifier into a new multi modal classifier. By leveraging information from both modalities, this combined approach aims to achieve higher levels of accuracy compared to using each classifier individually.

II. Data set Description

The UPMC Food-101 dataset, chosen for its complexity and real-world relevance, presents a formidable challenge in the realm of multimodal classification. Un like simpler datasets such as ETHZ Food-101, UPMC Food-101 contains images and accompanying text descriptions collected in uncontrolled environments. This uncontrolled nature introduces noise and variability, reflecting the complexities of real-world data. The proposed model architecture is tailored to address the chal-

Identify applicable funding agency here. If none, delete this.

lenges posed by UPMC Food-101. It comprises distinct pathways for image and text inputs, leveraging powerful techniques such as Inception V3 for image processing and BERT for text understanding. These pathways are integrated through concate nation, allowing the model to fuse information from both modalities and make informed classification decisions. dataset

(a) Real Sashimi

(b) Multiple plates

(c) Random image

(d) Wrong class

Fig. 1. Data set has different challenges

challenges, showcasing examples from the 'sashimi' class. While some images accurately depict sashimi plates, others contain noise, such as multiple plates or unrelated content. This noise, estimated at approximately 5 percent age of the dataset, underscores the importance of robust classification methods capable of handling such variability. Comparing UPMC Food-101 with ETHZ Food-101 further highlights the scale and complexity of the former. With nearly 91,000 images and text documents spanning 101 classes, UPMC Food-101 offers a rich and diverse dataset for experimentation. The decision to utilize only the title of each text document for classification underscores the focus on leveraging textual information in a concise and informative manner. Overall, the exploration of UPMC Food-101 underscores the challenges and opportunities inherent in multimodal classification tasks. By addressing noise and variability through robust model architectures and preprocessing techniques, 3 researchers can

UPMC and ETHZ Food-101 comparison			
Dataset	Number of Images per class	Data Type	Source Environment
ETHZ	1000	Images	Controlled
UPMC	790 - 956	Images + Text	Not Controlled

Fig. 2. Comparision of datasets

unlock valuable insights from complex real-world datasets like UPMC Food-101, advancing the field of multimodal classification.

III. Related Work

In this project, various preprocessing techniques were applied to both text and image data to enhance their quality and suitability for analysis. For text data, preprocessing involved removing punctuation, converting to lowercase, and elimi nating stop words to clean the text and reduce noise. Similarly, image data underwent resizing, noise reduction filtering, and pixel value normalization to standardize the images and improve consistency. Data augmentation techniques, such as rotation and flipping, were also utilized to enhance diversity and in crease the robustness of the dataset. Additionally, outliers and anomalies were identified and removed based on data exploration and visualization. The project leveraged the power of BERT (Bidirectional Encoder Representa tions from Transformers) for word embedding, enabling a deeper understanding of text data by analyzing each word in the context of its surroundings. This contextual understanding facilitated more accurate and informative word repre sentations, thereby enhancing the quality of text analysis and improving the performance of downstream natural language processing tasks. Furthermore, Hi erarchical Attention Networks (HANs) were employed to model hierarchical text structures, such as documents, by dynamically focusing on different levels of text granularity using attention mechanisms. By combining BERT's contextual embed dings with HANs' hierarchical modeling capabilities, the project achieved improved performance across various text processing tasks, offering a robust solution for natural language understanding tasks. In addition to BERT and HANs were incorporated to enhance sequence modeling and hierarchical text un derstanding. improving the model's ability to capture sequential dependencies and contextual information in both forward and backward directions. This integration further enhanced the project's capability to understand both individual words and broader contextual relationships within hierarchical text, contributing to its overall effectiveness in natural language understanding tasks.

IV. Proposed Approach

In designing a model for text classification, the focus is on categorizing tex tual data into predefined classes or categories. This process involves analyzing the content of text documents and assigning them to relevant classes based on their characteristics. Techniques such as preprocessing are commonly employed to clean and standardize the text data, which may include steps like removing punctuation, converting text to lowercase, and eliminating stop words to focus on meaningful content. The model learns to understand the contextual relation ships between words and phrases, enabling it to accurately categorize text into relevant classes. On the other hand, in image classification, the goal is to identify and cate gorize the contents of images. This involves analyzing the visual features present in images and assigning them to appropriate classes or categories. Preprocess ing techniques are applied to standardize the images and improve consistency, which may include resizing images, applying noise reduction filters, and nor malizing pixel values. The model learns to extract hierarchical

features from images through techniques like convolutional neural networks (CNNs), enabling it to recognize patterns and objects within images and make accurate classifica tions. while text classification focuses on understanding and categorizing textual data based on its content, image classification involves analyzing visual features in images to assign them to relevant categories. Both tasks utilize preprocessing techniques and machine learning algorithms to learn from labeled data and make accurate predictions.

A. Text classification using BERT and HANs

In addition to the advanced techniques like BERT, HANs, the model's training process incorporated optimization strategies and regularization techniques to further enhance its performance in text classification tasks. The Adam optimizer was employed to efficiently update the model's parameters during training, leveraging adaptive learning rates for faster convergence and improved convergence quality. Moreover, a learning rate of 0.01 was selected to strike a balance between fast convergence and stability, ensuring smooth progress through out the training process. To prevent overfitting and promote generalization, the model implemented early stopping and learning rate reduction techniques. Early stopping allowed the train ing process to halt when the model's performance on a validation dataset ceased to improve, thus preventing unnecessary training and mitigating the risk of over f itting to the training data. Additionally, learning rate reduction techniques, such as ReduceLROnPlateau, dynamically adjusted the learning rate during training based on the model's performance on the validation set. This adaptive learning rate scheduling ensured that the model continued to make progress even as it approached convergence, ultimately leading to more robust and generalizable text classification outcomes. By integrating these optimization strategies and regularization techniques into the training process, the model not only leveraged cutting-edge methodologies like BERT, HANs also optimized its learning process for maximum effectiveness. This holistic approach facilitated notable improvements in accuracy and performance, enabling the model to navigate through the complexities of textual data with precision and insight

B. Image classification with InceptionResNetV2

In the context of image classification, the model employed the Inception ResNet V2 architecture, renowned for its effectiveness in recognizing diverse objects and patterns within images. This pre-trained model was selected due to its robust performance in handling various image classification tasks, leveraging pre-trained weights on large datasets to extract meaningful features from input images. The choice of Inception ResNet V2 was motivated by its ability to handle variations in input images, including changes in lighting conditions, orientations, and back grounds, making it suitable for real-world applications where images may exhibit diverse characteristics. Typically, the Inception ResNet V2 model consists of 1,000 output classes representing a wide range of objects, animals, and scenes such as dogs, birds, bottles, cars, beaches, and mountains. Its architecture comprises hundreds of lay ers, allowing it to capture complex hierarchical features present in images. The utilization of ResNet architecture within Inception ResNet V2 addresses the van ishing gradient problem commonly encountered in deep neural networks, enabling more efficient training and improved performance. To adapt the pre-trained Inception ResNet V2 model for specific classification tasks, an additional Dense layer was added. This Dense layer incorporated the Rectified Linear Unit (ReLU) activation function to introduce non-linearity and enhance the model's capacity to learn complex relationships within the data. Furthermore, a dropout rate of 30 percentage was applied to the Dense layer to prevent overfitting by randomly dropping out a fraction of the neurons during training, promoting better generalization and robustness of the model. In summary, the model architecture for image classification leveraged the pow erful capabilities of the Inception ResNet V2 pre-trained model, augmented with a customized Dense layer. This approach enabled the model to effectively recog nize and categorize objects and scenes within images, overcoming challenges posed by variations in input data while maintaining high accuracy and generalization performance.

C. Fusion of Text and images

First, dataset of images and text for fusion, we're integrating two modalities of data that are inherently related to each other. In this context, the images depict visual representations of objects, scenes, or concepts, while the accompa nying text provides descriptive information or context about those images. For example, an image of a food item may be accompanied by text describing its make, model, and features. The fusion process aims to leverage the comple mentary information from both modalities to enhance the overall understanding of the data. By analyzing the visual content of images alongside the textual descriptions, the model can capture richer semantic meaning and context. This combined approach enables more accurate and insightful classification or analysis tasks, as the model can draw upon both visual and textual cues to make in formed decisions The concatenation step in the model architecture represents a crucial phase where the outputs from the individual text and image models are merged to form a cohesive representation that captures the combined information from both modalities. This merged representation serves as the foundation for subsequent layers to extract higher-level features and make classification predic tions. With the Concatenate layer, the outputs from the text and image sides of the model are fused together, enabling the model to consider both textual descriptions and visual features simultaneously. This integration facilitates a more comprehensive understanding of the data, leveraging the inherent relationships be tween text and images to enhance classification accuracy and robustness. Following the concatenation, the merged representation undergoes further pro cessing through additional Dense layers. The first Dense layer with 256 units and ReLU activation function refines the combined features, introducing non linearity to capture complex patterns and relationships

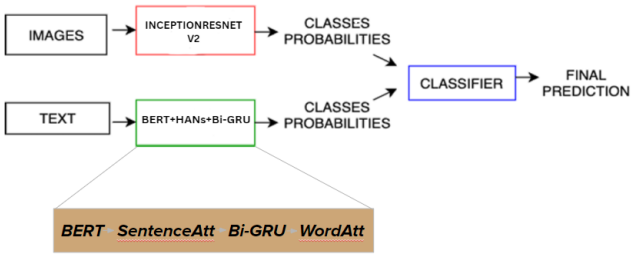

Fig. 3. Proposed Model Architecture

within the data. The subsequent Dropout layer with a dropout rate of 0.2 helps prevent overfitting by randomly deactivating neurons during training, promoting better generalization of the model. Continuing the refinement process, another Dense layer with 128 units and ReLU activation function further enhances the extracted features, facilitating the model's ability to capture and represent the intricate characteristics of the com bined data. Once again, a Dropout layer with a dropout rate of 0.2 is incorporated to improve regularization and prevent the model from memorizing noise in the training data. At the output layer with a softmax activation function generates the final classification probabilities across the defined classes. By combining the information from both text and

TABLE I
PROPOSED MODEL ARCHITECTURE

Image Input	Text Input
INCEPTIONRESNET	BERT Model
AVG POOLING2D (8X8)	Sentence Attention (128)
Dropout (30%)	Word Attention (128)
Dense ReLU (256)	Dense ReLU (256)
Dropout (30%)	Dropout (10%)
Dense Softmax (101 classes)	Dense Softmax (101 classes)
Concatenation	
Dense ReLU (128)	
Dropout (20%)	
Dense ReLU (128)	
Dropout (20%)	
Dense Softmax (101 classes)	

images through concatenation and subsequent processing, the model is equipped to make accurate predictions based on the comprehensive understanding of the input data. This approach harnesses the complementary nature of text and image data, enabling the model to exploit rich semantic information and achieve superior performance in classification tasks. These methodologies complement each other's strengths, allowing the model to harness the rich information present in both textual and visual modalities. By integrating features extracted from text and images through concatenation, the model gains a comprehensive understanding of the input data, leading to more accurate and insightful analysis and classification outcomes. This holistic approach to multi-modal learning empowers the model to tackle complex real-world problems with enhanced accuracy and effectiveness, making it a powerful tool for a wide range of applications in natural language processing and computer vision.

V. RESULTS

The experiment achieved a text classification accuracy of 85.38%, indicating the model's ability to accurately categorize textual information. This high accuracy suggests that the model effectively understands the meaning of text, capturing subtle nuances and relationships within sentences. The use of advanced techniques like BERT, HANs contributed to this success, demonstrating the power of combining different methodologies for text analysis. Overall, this result highlights the model's proficiency in text classification and its potential for various real-world applications. The obtained image classification accuracy of 69.38% signifies the model's ca pability to correctly identify objects and patterns within images. While this ac curacy demonstrates some effectiveness in categorizing visual data, it also suggests room for improvement. Possible factors contributing to this moderate accuracy could include the complexity of the dataset, variations in lighting conditions, orientations, and backgrounds, as well as the need for further optimization of the model architecture and training parameters. Despite the modest accuracy, this result

TABLE II
TEST ACCURACY OF PROPOSED MODEL

Model	Accuracy
Text Model	73
Image Model	66
Fusion(cross) Model	85

lays a foundation for future refinement and enhancement of the model's performance in image classification tasks. The fusion results of almost 85% indicate the effectiveness of combining information from both text and image modalities in improving overall classification accuracy. By integrating features extracted from both modalities, the fusion model achieves a higher level of understanding and context, leading to more accurate classi f ication outcomes. This demonstrates the synergy between textual descriptions and visual representations, allowing the model to leverage the strengths of both modalities to achieve superior performance. The fusion approach enhances the model's ability to discern patterns and relationships across diverse datasets, un derscoring its potential for a wide range of applications in multimodal learning and classification tasks.