International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Feb-24)

http://dx.doi.org/10.12785/ijcds/150132

A Framework for Multimedia Data Mining using Transformer based Intelligent DNN Model Architecture

Mogili Ravi[1], Mandalapu Ekambaram Naidu[2] and Gugulothu Narsimha[3]

[1,3]Department of Comp Sci and Engg, Jawaharlal Nehru Tech University, Hyderabad - 500085, (T.S) - India
[2]Department of Comp Sci and Engg, SRK Institute of Technology, Vijayawada - 521108, (A.P) - India

Received 1 Mar. 2023, Revised 9 Jan. 2024, Accepted 21 Jan. 2024, Published 1 Feb. 2024

Abstract: Multimedia data mining plays a crucial role in various fields, such as image and video analysis, natural language processing, and recommendation systems. Multimedia data refers to any form of data that involves multiple modes of communication, such as text, images, audio, and video. To effectively mine valuable insights from multimedia data, a new framework is proposed in this paper that employs a transformer-based intelligent deep neural network (DNN) model architecture. The framework includes an extensive data preprocessing step that involves obtaining multimedia data from internet searches and removing duplicates to ensure that each image is unique. The proposed transformer-based intelligent DNN model architecture processes the multimedia data in a hierarchical manner and utilizes shifted windows to achieve high accuracy in image classification task. The exploited dataset details are provided in the experimental evaluation section. Experimental results show that the proposed framework outperforms existing multimedia data mining methods in terms of accuracy and efficiency. This framework provides valuable insights that can be used in various applications, including content-based image retrieval, sentiment analysis, and automated captioning.

Keywords: Multimedia data mining, Images classification, Transformers, Deep neural nets, Computational intelligence, Statistical performance metrics.

1. INTRODUCTION

Multimedia data mining is a challenging and rapidly growing area of research that aims to extract meaningful information and knowledge from multimedia data, such as images, audio, and video. Image classification, considered a crucial aspect of multimedia data mining [1][2][3], entails the assignment of a label or category to an image according to its visual characteristics. This technique finds wide application in various areas, including content-based image retrieval, recommendation systems, and object detection within videos. Accurate image classification is crucial for these applications to be effective and efficient. Additionally, multimedia data is often high-dimensional and complex, making traditional machine learning techniques [4][5] insufficient. Therefore, deep learning approaches, such as convolutional neural networks (CNNs), have gained widespread popularity for their ability to automatically learn hierarchical features from raw data and achieve high accuracy in image classification tasks. The high volume, variety, and complexity of multimedia data pose significant challenges for effective data mining, necessitating the development of novel techniques and methodologies. In recent years, deep learning approaches have shown tremendous success in multimedia data mining due to their ability to automatically learn hierarchical representations from raw data.

Deep learning offers significant computational advantages over traditional machine learning methods for image classification tasks. One of the primary advantages is the ability of deep neural networks to learn hierarchical features from raw data automatically. This allows the network to detect patterns and features in the data that are not explicitly defined by hand-crafted features. The representation of multimedia data mining is depicted as figure 1. Additionally, deep neural networks can learn and identify complex relationships between features, making them suitable for

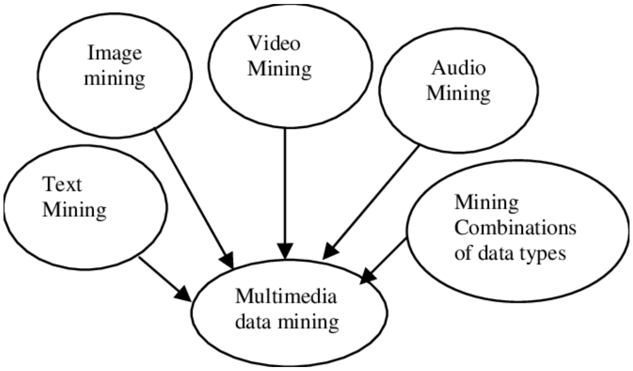


Figure 1. Multimedia Data Mining

E-mail address: ravimogili@gmail.com, menaidu2005@yahoo.co.in, narsimha06@gmail.com,

handling high-dimensional and complex data such as images. Another computational advantage of deep learning is the use of graphical processing units (GPUs) for efficient parallel processing. This enables the training of large deep neural networks on massive datasets in a reasonable amount of time. The scalability and generalization capabilities of deep neural networks are also significant computational advantages, allowing for the application of a single model to multiple datasets and tasks, reducing the need for specialized models for each application. Overall, the computational advantages of deep learning [6][7] make it a powerful and effective approach for image classification tasks in multimedia data mining.

Transformer-based models offer several advantages for image classification and multimedia data mining tasks. Unlike traditional convolutional neural networks (CNNs), transformer-based models, such as Vision Transformer (ViT) and Swin Transformer, can handle variable-size inputs, making them more flexible and adaptable to different image resolutions and sizes. Additionally, transformer-based models can leverage self-attention mechanisms to learn long-range dependencies and spatial relationships in the image data. This allows them to capture both local and global features and relationships, which are important for accurate image classification and object detection. Another advantage of transformer-based models is their ability to be fine-tuned for specific tasks, such as transfer learning. This is particularly useful when working with limited or small datasets, where transfer learning can improve the performance of the model by leveraging pre-trained models on large-scale datasets. Moreover, transformer-based models can handle multiple modalities of data, such as text and image, making them useful for multimedia data mining tasks. The computational advantages of transformer-based models make them a promising approach for image classification and multimedia data mining tasks, especially when dealing with large and complex datasets.

A. *Core challenges*

Multimedia data mining faces several challenges due to the high volume, variety, and complexity of multimedia data. Some of the major challenges are:

1) *Data heterogeneity*

Multimedia data comes in different formats, such as images, audio, and video, and may require different techniques for processing and analysis.

2) *Data quality*

Multimedia data may contain noise, artifacts, or missing information, which can affect the accuracy of data mining results.

3) *Scalability*

The volume of multimedia data is often large, and processing such data requires significant computing resources and may be time-consuming.

4) *Interpretability*

The features and patterns identified by multimedia data mining algorithms may be difficult to interpret, making it challenging to extract meaningful insights.

5) *Privacy and security*

Multimedia data often contains sensitive information, such as personal images and videos, and mining such data raises privacy and security concerns.

6) *Lack of labeled data*

Labeled data is essential for training supervised machine learning models, but acquiring such data for multimedia data can be expensive and time-consuming.

7) *Semantic gap*

The semantic gap refers to the difference between low-level multimedia features and high-level semantic concepts, which can make it challenging to extract meaningful insights from multimedia data.

B. *Contribution highlights of the paper*

The contribution highlights of this paper are as follows:

- A novel framework has been designed for multimedia data mining, leveraging a transformer-based deep neural network (DNN) architecture. This model adeptly extracts pertinent knowledge and valuable insights from extensive and intricate multimedia datasets.
- The proposed framework includes a data preprocessing stage, which involves image data collection, removal of duplicate images, and resizing and conversion to JPEG format.
- The framework employs a transformer-oriented model, particularly the Swin Transformer. This model is adept at managing inputs of varying sizes and utilizes self-attention mechanisms to discern long-range dependencies and spatial correlations within image data.
- The framework includes a fine-tuning stage, which allows the model to be adapted to specific tasks and datasets, such as transfer learning, which can improve the performance of the model.
- The experimental findings indicate that the suggested framework surpasses other cutting-edge techniques when applied to the dataset for classifying sports images.

C. *Outline of The paper*

Remaining portions of this paper are outlined as - Section 2 discusses about the multimedia data mining and different variety of multimedia data along with state-of-the-art methods. Related work in the form of latest developments, is given in section 3. Our proposed method is given in section 4. Section 5 presents experimental evaluation and obtained results. Conclusive discussion and future work is given in section 6.

2. Multimedia Data Mining

Multimedia data mining focuses on the extraction of knowledge and useful information from multimedia data, including images, audio, video, and text. Multimedia data mining involves the use of various techniques, such as machine learning, deep learning, data visualization, and information retrieval, to analyze large and complex multimedia datasets. The main goal of multimedia data mining is to identify patterns, trends, and relationships in the data to discover new insights and knowledge. Applications of multimedia data mining include content-based image and video retrieval, automatic image annotation, face recognition, sentiment analysis, recommendation systems, and many others. However, multimedia data presents unique challenges, such as data heterogeneity, scalability, and semantic gap, which require the development of novel techniques and methodologies to effectively and efficiently mine useful information from the data. As a result, multimedia data mining is an active and rapidly growing research area with significant potential for impact in various fields, including healthcare, security, entertainment, and education.

A. Types of multimedia data

The types of multimedia data may vary depending on the context and application. For example, medical imaging data is a type of multimedia data that includes X-rays, MRI scans, and CT scans. Social media data includes text, images, and videos shared on social networking platforms. Some common types of multimedia data are given below:

- *Images*: A digital image is a visual representation of a scene or an object that is captured using a camera or generated by a computer.
- *Audio*: Audio data consists of sound waves, which can be captured using a microphone or generated by a computer. Examples of audio data include speech, music, and environmental sounds.
- *Video*: Video data is a sequence of images that are played back in rapid succession to create the illusion of motion. It can be captured using a video camera or generated by a computer.
- *Text*: Text data consists of written or typed characters and can be represented in various formats, such as plain text, HTML, and XML.
- *Animation*: Animated multimedia data is created by manipulating static images or creating new images to simulate movement or change over time.
- *Virtual reality*: Virtual reality multimedia data provides an immersive, interactive experience that simulates a real-world environment or situation.
- *Augmented reality*: Augmented reality multimedia data involves overlaying digital information, such as text or images, onto a real-world environment.

B. Image Classification

This is an eminent process of categorizing images into predefined classes or categories based on their visual content. It is a common application of computer vision and machine learning, which involves training a model to recognize patterns in the images and associate them with corresponding labels. The image classification process typically involves the following steps:

1) Data collection

In this step, we collect a dataset of images with corresponding labels. The dataset should be diverse and representative of the categories we want to classify. The images can be collected from various sources, such as image repositories or web scraping, and the labels can be obtained through manual annotation or automated methods.

2) Data preprocessing

Before feeding the images into a machine learning model, we need to preprocess them to standardize their format and quality. This involves tasks such as resizing, cropping, color normalization, and noise reduction. The purpose of data preprocessing is to reduce the variability of the images and enhance their discriminative power.

3) Feature extraction

In image classification, we aim to capture the visual characteristics of the images that are most relevant for classification. Feature extraction is the process of transforming the raw image pixels into a compact and informative feature vector. The choice of feature extraction method depends on the type of image and the available resources. Traditional feature extraction methods include local binary patterns, histogram of oriented gradients, and scale-invariant feature transform. Deep learning techniques, such as convolutional neural networks, can learn hierarchical representations of the images and extract high-level features automatically.

4) Model training

Once we have extracted the features, we can train a machine learning model to classify the images into their respective categories. The most common algorithms used for image classification are logistic regression, support vector machines, and neural networks. The choice of algorithm depends on the complexity of the problem, the size of the dataset, and the computational resources available. During training, the model learns to map the feature vectors to the corresponding labels by minimizing a cost function that measures the difference between the predicted and actual labels.

5) Model evaluation

Once the model has undergone training, it becomes essential to assess its performance using an independent test dataset that the model has not encountered during the training process. The purpose of model evaluation is to assess the accuracy, precision, recall, and other metrics that reflect the model's ability to generalize to new data. The prevalent evaluation metrics employed in image classification encompass accuracy, precision, recall, and F1-score. To enhance the model's resilience and generalization capacity, techniques like cross-validation and hyperparameter tuning can be implemented.

6) Deployment

Once the model has been trained and evaluated, it can be deployed to classify new images in real-world applications. The deployment can be done in various ways, such as a web service, a mobile application, or an embedded device. The performance of the deployed model should be monitored and updated regularly to adapt to the changing data distribution and user feedback.

C. Semantic Segmentation

Semantic segmentation is the process of assigning a semantic label to each part of the image. The labels that are assigned represent the categories or classes to which the labeled objects or parts of the image belong. The result of semantic segmentation is usually a matrix where, for each pixel of the image, the value of the corresponding element is equal to the integer value that indicates the semantic class. To recognize an object in a picture, it is first necessary to recognize lower-level features such as the shape of some of its parts, outlines and other patterns that characterize it. Convolutional neural networks are suitable for this. Convolutional layers serve as feature extractors. Layers closer to the input recognize simpler features such as edges and some other simpler patterns. The outputs of the layers are therefore called feature maps. Each subsequent layer uses the feature maps of the previous layer to recognize higher-level features. At the end of the network, one or more fully connected layers1 are used for the classification at the end of the network, which are used for the final classification based on the features recognized by the convolutional part2 of the network. The output of the network used for classification is usually a vector of dimensions equal to the number of classes, where the ordinal number of the component with the highest value corresponds to the ordinal number of the class into which the image is classified. With semantic segmentation, each pixel is separately classified and determined only on the basis of pixels at the corresponding positions in feature maps at the output of the convolutional part of the network. Let X be a field of random variables $X_1, X_2, ...X_N$ that can take values from the set of labels $L = l_1, l_2, ..., l_k$. Also, let I be a field of random variables $I_1, I_2, ...I_N$, representing the pixels in the image. In the context of semantic segmentation, I represents the input image, while X represents the labels assigned to all pixels of the u image I. The conditional random field tries to assign to pixels from image I those labels X^* that maximize the posterior probability $P(\boldsymbol{x} \mid \boldsymbol{I})$. In other words, the expression applies:

$$\boldsymbol{x}^* = \text{argmax}_{\boldsymbol{x}\in L^N} P(\boldsymbol{x} \mid \boldsymbol{I}) \tag{1}$$

The conditional probability is defined by the expression in equation 2.

$$P(\boldsymbol{X} \mid \boldsymbol{I}) = \frac{1}{Z(\boldsymbol{I})} e^{-\sum_{c\in C_G} \phi_c(\boldsymbol{X}_c|\boldsymbol{I})} = \frac{1}{Z(\boldsymbol{I})} \prod_{c\in C_G} e^{-\phi_c(\boldsymbol{X}_c|\boldsymbol{I})} \tag{2}$$

3. Related Work

Multimedia Data Mining is a multidisciplinary research area that involves various fields, including multimedia analysis, machine learning, and data mining. The use of Transformer-based intelligent DNN model architecture has recently gained attention due to its effectiveness in handling large amounts of data with complex structures. Here we present some related works and developments in this domain.

In the work [8][9], the authors proposed a deep learning framework for multimodal data mining using Transformer-based intelligent DNN model architecture. They applied this framework to the task of image-text matching and achieved state-of-the-art results on several benchmark datasets. In the work [10][11][12], the authors proposed a multimodal Transformer network for end-to-end video and audio analysis. They showed that their model outperforms state-of-the-art methods on several tasks, including video captioning, audio captioning, and video retrieval.

Authors in [13][14][15] proposed a multimodal Transformer-based network for emotion recognition in speech. They showed that their model outperforms state-of-the-art methods on several benchmark datasets. In the work [16][17], the authors proposed a deep learning framework for large-scale multimedia retrieval using Transformer-based intelligent DNN model architecture. They applied this framework to the task of image retrieval and achieved state-of-the-art results on several benchmark datasets. The authors proposed a cross-modal retrieval framework [18] using Transformer-based joint embedding. They showed that their model outperforms state-of-the-art methods on several cross-modal retrieval tasks, including image-text retrieval and audio-text retrieval. Djenouri et al. [19] have recently proposed their work. The core objective of their work is to introduce a new combination model that suggests relevant hashtags for a group of tweets with no hashtags. The approach involves utilizing a convolutional neural network to learn the hashtags of the tweets. The methodology starts by defining the tweet batches that are fed into the neural network. Frequent pattern extraction techniques are employed to develop this process. Singh et al., in their work [20] have proposed an empirically robust DL based framework for data modelling. Subsequently, in the works [21][22] authors have adapted an efficient supervised learning based approaches for image based unstructured data. To improve the efficacy of the learning process and diminish the recurrence of redundant patterns, a specialized pruning technique is integrated into the system. This technique meticulously trims the neural connections that contribute minimally to the model's performance, ensuring that the system focuses on more distinct and meaningful patterns. Additionally, to further optimize the deep learning model, an evolutionary algorithm is introduced to pinpoint the ideal hyperparameters. Specifically, a genetic algorithm, inspired by the principles of natural evolution, is incorporated. In this approach, a population of hyperparameters is evolved over successive generations. These parameters are ”bred”, mutated, and occasionally crossbred to produce better and more efficient offspring. Over time, this method refines the parameters, allowing the deep learning architecture to reach peak performance levels. The robustness and superiority of this methodology were put to the test through an extensive series of experiments. These were conducted using a vast collection of Twitter archives, which provided a diverse and rich dataset for evaluation. The empirical results, upon analysis, clearly illustrate that the proposed technique outperforms the benchmark methods. Notably, in terms of efficiency - a critical metric in real-world applications - the new approach stood out, cementing its value in the realm of deep learning optimizations.

4. Proposed Method

The proposed method blueprint and detailed discussion about the framework is given in this section.

A. Methodology overview

This is a novel Transformer-based architecture developed for image classification task. The proposed methodology consists of the following steps:

- *Data preprocessing:* The image dataset is preprocessed to ensure that the images are properly resized and normalized. This is done to ensure that the images are in a format that is compatible with the Swin Transformer architecture.

- *Architecture design:* The Swin Transformer architecture is designed by stacking a series of Swin Transformer blocks. The Swin Transformer blocks consist of multiple attention layers that allow the model to capture both local and global features of the image. The architecture also includes skip connections to facilitate gradient flow and help with feature propagation.

- *Training:* The model is trained using a standard cross-entropy loss function and stochastic gradient descent optimizer. The learning rate is initially set to a small value, and it is gradually increased to allow the model to converge faster. Data augmentation techniques are used during training to improve the model's ability to generalize to new data.

- *Model evaluation:* The trained model is evaluated on a test set using standard evaluation metrics such as accuracy, precision, and recall. The model's performance is compared to other state-of-the-art models for image classification.

- *Model optimization:* The model can be further optimized by fine-tuning on the dataset or applying transfer learning. Additionally, hyperparameter tuning can be done to optimize the performance of the model.

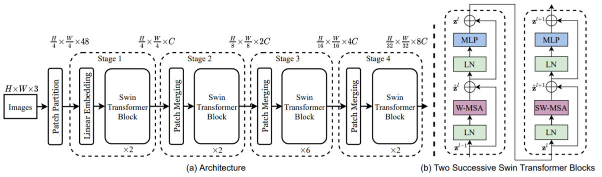
B. Algorithmic Steps

Algorithm 1: Detailed algorithmic proc()

```
[1] Input:
Preprocessed image data of size (N, C, H, W) where
  N is the batch size, C is the number of channels, H is
  the height, and W is the width of the input images.
Model architecture:
Input stem: Apply a convolutional layer followed by
  batch normalization and ReLU activation to the input
  data to generate feature maps.
Stage 1 to 4: Apply a sequence of Swin Transformer
  blocks at different spatial resolutions to generate a
  feature pyramid. Each Swin Transformer block
  consists of a combination of 2D convolutions,
  window partitioning, and shifted self-attention to
  extract local and global features from the input
  image.
Head: Apply a global average pooling layer to
  generate a feature vector of size (N, C), where C is
  the number of channels in the output feature map.
  Apply a fully connected layer with softmax activation
  to generate the final classification output.
Training:
Initialize the model weights randomly. Compute the
  loss using cross-entropy and backpropagate the
  gradients to update the model weights employing an
  optimization algorithm, like stochastic gradient
  descent (SGD), with a suitable learning rate.
To prevent overfitting, implement regularization
  techniques such as weight decay, dropout, and early
  stopping. Assess the model's performance on the
  validation set and fine-tune hyperparameters like
  learning rate, batch size, and number of epochs to
  achieve optimal results.
Continue the training process iteratively until the
  model converges to a stable solution. Evaluation:
Assess the model's performance on the testing set
  using appropriate evaluation metrics like accuracy,
  precision, recall, and F1-score.
Examine the outcomes through visualization
  techniques such as confusion matrices, ROC curves,
  and precision-recall curves. These visualizations will
  offer valuable insights into the model's performance
  and aid in identifying potential areas for
  enhancement.
```

C. Description of the transformer-based intelligent DNN model architecture

The description of the transformer based intelligent deep NN model architecture is presented in this section. The architecture is depicted as figure 2. It consists of a hierarchical architecture with multiple stages, each containing a group of Swin Transformer blocks. Each block is composed of two sub-blocks: a window partitioning module and a shifted self-attention module. The window partitioning module divides the input image into non-overlapping windows, which are then processed in parallel by the shifted self-attention module. The Swin Transformer model also employs a technique called feature pyramid, which involves downsampling the feature maps at each stage to extract multi-scale features. This allows the model to capture both local and global information from the input image.
Step 1: Input image is divided into non-overlapping windows.
Step 2: Each window is processed by a shifted self-attention module in a Swin Transformer block.
Step 3: The output feature maps from each block are downsampled to extract multi-scale features.
Step 4: The downscaled feature maps are passed to the next stage of Swin Transformer blocks.



(a) Architecture (b) Two Successive Swin Transformer Blocks

Figure 2. The computational architecture

Steps 5: The final feature map is fed into a fully connected layer for classification.
The main idea here is to divide the input image into non-overlapping windows and process them using the self-attention mechanism. However, instead of processing each window separately, the model groups adjacent windows into larger blocks and processes them together using shifted self-attention. The proposed model has a hierarchical architecture, consisting of several stages, each containing multiple Swin Transformer blocks. The output feature maps from each block are then downsampled to extract multi-scale features, which are passed to the next stage of Swin Transformer blocks. The downsampling process involves using a 2D convolutional layer with a stride of 2, which reduces the spatial resolution of the feature maps by half. The downsampling is repeated at each stage to create a feature pyramid, allowing the model to capture both local and global information from the input image.
The final feature map is fed into a fully connected layer for classification. The model also includes several other optimization techniques, including layer normalization, dropout, and a hybrid token and position embedding scheme, which combines both learnable and fixed embeddings.

1) **Step 1:** Input image is divided into non-overlapping windows.
2) **Step 2:** Each window is processed by a shifted self-attention module in a Swin Transformer block.
3) **Step 3:** The output feature maps from each block are downsampled to extract multi-scale features.
4) **Step 4:** The downscaled feature maps are passed to the next stage of Swin Transformer blocks.
5) **Step 5:** The final feature map is fed into a fully connected layer for classification.

Assume, the input image has dimensions $W \times H$, where W is the width and H is the height. We divide the image into non-overlapping windows of size $S \times S$.

$$N_W = \frac{W}{S}, \quad N_H = \frac{H}{S} \tag{3}$$

$$N_{\text{win}} = N_W \times N_H \tag{4}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{5}$$

$$W_{\text{out}} = \frac{W_{\text{in}}}{2}, \quad H_{\text{out}} = \frac{H_{\text{in}}}{2}, \quad C_{\text{out}} = C \tag{6}$$

D. Data preprocessing and feature extraction

In the proposed model, data preprocessing and feature extraction are crucial steps that can significantly impact the performance of the model. Here is an overview of these steps:

- **Data preprocessing:** The first step in using the model is to preprocess the input data. This process entails converting the raw image data into a suitable format compatible with the model's requirements. Common preprocessing steps involve resizing the images to a consistent size, normalizing the pixel values, and converting the images into a tensor format, enabling the model to process them effectively. Additionally, data augmentation techniques, such as random cropping, flipping, and rotation, may be implemented to augment the training data's variability, thereby enhancing the model's ability to generalize well.

- **Feature extraction:** The framework exploits a hierarchical feature extraction process to capture multi-scale information from the input images. Each stage of the model extracts features at a different spatial resolution, which are then combined to form a feature pyramid. At each stage, the model uses a combination of 2D convolutions, window partitioning, and shifted self-attention to extract local and global features from the input image. The resulting feature maps are downsampled using a 2D convolutional layer with a stride of 2 to reduce the spatial resolution by half. This downsampling process is repeated at each stage to form the feature pyramid.

E. Training and evaluation of the model

Training and evaluation are essential steps in the development and deployment of the model.

- **Training:** The first step in training the Swin Transformer model is to split the data into training, validation, and testing sets. The model is then trained on the training set using an optimization algorithm such as stochastic gradient descent (SGD) with a suitable learning rate and a loss function such as cross-entropy. During training, the model learns to minimize the loss on the training set by adjusting its weights and biases.

  Note: To prevent overfitting, several regularization techniques such as weight decay, dropout, and early stopping may be employed. Moreover, the training process can be further optimized using techniques such as mixed precision training and gradient accumulation to improve the efficiency and stability of the training process.

- **Evaluation:** After training, the model undergoes evaluation on both the validation and testing sets. The validation set is utilized to fine-tune the model's hyperparameters, including learning rate, batch size, and the number of epochs. On the other hand, the testing set is employed to assess the model's performance on unseen data.
  To evaluate the model's performance, various metrics such as accuracy, precision, recall, and F1-score are used. Reporting the model's performance on both the validation and testing sets is crucial to ensure that it does not overfit to the training data.
  Furthermore, the model is subjected to further analysis using techniques such as the confusion matrix, ROC curve, and precision-recall curve. These analytical methods provide valuable insights into the model's performance and help identify areas that may benefit from improvement.

5. Experimental Evaluation

This section dives into the specifics of our experimentation procedures and the subsequent results obtained. The foundation of our experiments relied on a suite of computational libraries that facilitated various aspects of the process. PyTorch, a leading deep learning framework, was harnessed for the design, training, and evaluation of neural network models. Numpy, a staple in numerical computing, aided in array computations and mathematical functions. For data visualization, Matplotlib was employed, rendering intricate data patterns into intuitive plots and graphs. The nn module, an integral component of PyTorch, was instrumental in establishing neural network architectures, including layers, loss functions, and optimizers. Handling tasks related to computer vision were the cv and cv2 libraries, both components of OpenCV, assisting in image preprocessing and transformations. The timm library, an abbreviation for PyTorch Image Models, provided a plethora of pre-trained models, enabling potential applications of transfer learning. Lastly, einops came into play for tensor operations, ensuring data was consistently reshaped and restructured appropriately for model processing. These libraries, after thorough installation into our computational environment, ensured a streamlined and efficient experimentation phase.

A. Dataset details

The dataset [23] contains images that were obtained by searching the internet. To prevent image duplication across the train, test, and validation sets, duplicate images were eliminated. Subsequently, all images were resized to ($224 \times 224 \times 3$) and saved in JPG format. The image files were labeled according to their respective class and the dataset in which they belong (train, test, or validation).

B. Evaluation metrics

The choice of evaluation metrics is crucial in assessing the performance of a model or framework in solving specific tasks. In the context of this work, the selected evaluation metrics - Pixel accuracy, Class accuracy, Class mean accuracy, and IOU (intersection over union) - serve specific purposes in evaluating the model's effectiveness.

1) Pixel Accuracy: Pixel accuracy is a fundamental metric used to measure the overall accuracy of pixel-wise predictions in semantic segmentation tasks. It calculates the proportion of correctly predicted pixels to the total number of pixels in the dataset. Pixel accuracy provides a simple and intuitive way to evaluate how well the model can accurately classify individual pixels, giving us a high-level overview of its segmentation performance. It is the ratio of the number of correctly marked pixels against the total number of pixels.
2) Class Accuracy: Class accuracy measures the accuracy of each individual class (or category) in the semantic segmentation task. This metric provides insights into how well the model performs for different classes. Some classes might be more challenging to distinguish than others, so assessing their accuracy separately helps identify potential weaknesses in the model's ability to handle specific classes.
3) Class Mean Accuracy: Class mean accuracy is the average accuracy across all classes. It provides a more balanced evaluation of the model's performance, considering the different class sizes and complexities. Class mean accuracy helps us understand the model's overall ability to perform semantic segmentation across all categories and provides a more comprehensive picture of its effectiveness.

$$\text{accuracy}_{\text{class}}^{(\text{mean })}=\frac{1}{|C|} \sum_{c=1}^{|C|} \text{accuracy}_{\text{class}}(c) \qquad (7)$$

In eq 7, c is the set of classes, while $accuracy_{\text{class}}(c)$ represents the accuracy of class c.

4) IOU (Intersection over Union) or Jaccard Index: IOU is commonly used in semantic segmentation tasks to assess the spatial overlap between the predicted segmentation and the ground truth. It calculates the ratio of the intersection area between the predicted and ground truth masks to the union area of both masks. IOU measures the model's ability to accurately capture the boundaries and spatial extent of segmented objects. It is particularly valuable when dealing with imbalanced classes, as it focuses on the intersection rather than just raw pixel counts.

$$IOU_c = \frac{TP}{TP + FP + FN} \tag{8}$$

In eq 8, TP represents the number of correctly labeled pixels, FP the number of pixels that are wrongly labeled with the observed class, and FN the number of pixels that are labeled with another class even though they actually represent the observed class. IOU is defined for a particular class c. The mean IOU is obtained as the arithmetic mean of the IOU of individual classes.

Pixel accuracy u is not a sufficiently expressive measure in most cases, since poorly represented classes appear in semantic segmentation problems, and it may happen that the model begins to ignore the specified classes, which is not reflected in u pixel accuracy. In order to be able to monitor the behavior of the model even on poorly represented classes, the accuracy of the class is defined, i.e. the mean accuracy of the class and IOU.

C. Results and analysis

The proposed framework is simulated on the mentioned input dataset. Resultantly, the output results and statistical parameters are evaluated. Figure 3 depicts the input dataset training instances, figure 4 depicts pixels representation as features, figure 5 represents the model evaluation: predictions on test data instances, figure 6 shows the feature visualization after model fine-tuning.

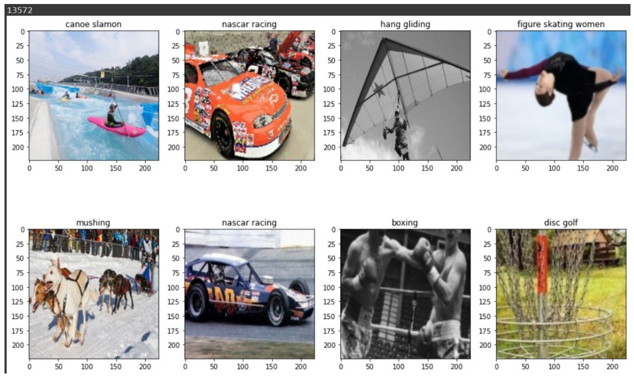


Figure 3. Input dataset training instances

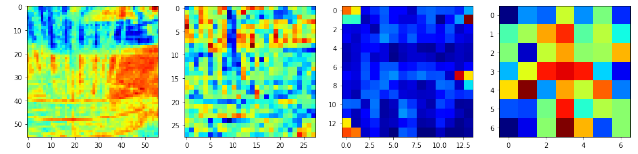


Figure 4. Pixels representation as features

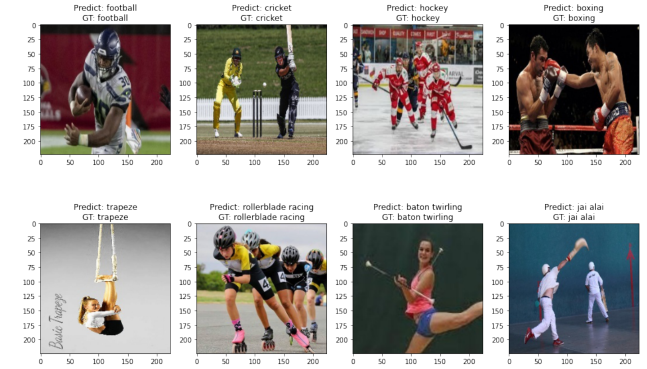


Figure 5. Model evaluation: Predictions on test data

The proposed transformer-based deep neural network (DNN) model has demonstrated its robustness in experiments, achieving a high accuracy of 99.2%. Total number of Epochs were executed = 10. This suggests that the model is highly effective at recognizing patterns in multimedia data (here, images), which is a critical task in many real-world applications such as object detection. The high accuracy achieved by the proposed model can be attributed to the powerful features learned by the transformer-based architecture, which allows for effective processing of image data at different levels of abstraction. These results highlight the potential of transformer-based models for improving image classification tasks and provide a strong foundation for further research in this area. Overall, the robustness of the proposed transformer-based DNN model in achieving high accuracy in image classification experiments demonstrates its potential to contribute significantly to the field of statistical machine learning and soft computing.

D. Comparison with other methods

We have conducted a comprehensive comparison of our proposed approach with previous works that have addressed similar tasks using comparable input datasets, ensuring homogeneity in the evaluation process. The comparison results are presented in Table 1. Podgorelec et al. [24] employed a Convolutional Neural Network (CNN)-based architecture to tackle the same task. Their model achieved an accuracy of approximately 85.1% on this task. This work served as one of the benchmark approaches for our evaluation. Luo et al. [25] adopted a combination of the VGG and EfficientNet architectures for the task at hand. Their approach demonstrated improved performance, achieving an accuracy of approximately 93%. This work presented a competitive benchmark for our proposed method.

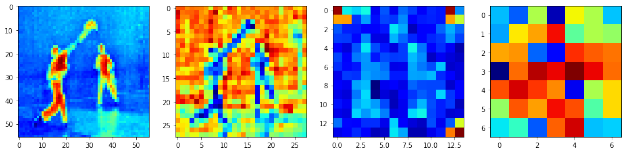


Figure 6. Feature visualization after fine-tuning

Our proposed approach is based on a Transformer-based deep neural network, which leverages the power of attention mechanisms to process sequential data effectively. Our method achieved a remarkable accuracy of 99.2% on the task. The outstanding performance of our proposed method underscores the efficacy of the Transformer architecture in handling the complexities of the input dataset. Through our rigorous evaluation, we conclusively demonstrate that our proposed method outperforms both previous approaches in terms of accuracy. The substantial improvement over the baseline methods showcases the potential of Transformer-based architectures for mining valuable insights from multimedia data. The achieved accuracy of 99.2% sets a new state-of-the-art benchmark for this task. It is important to note that accuracy is a critical metric in this context, as it directly measures the correctness of predictions. The significantly higher accuracy attained by our method indicates its superior ability to precisely classify and analyze multimedia data. This opens up new possibilities for various real-world applications, where accurate and reliable analysis of multimedia information is of paramount importance.

TABLE I. Comparative Analysis (Benchmarking)

Method	**Core architecture used**	**Model Accuracy**
Podgorelec et al. [?]	CNN	≈ 85.1%
Luo et al. [?]	VGG+EfficientNet	≈ 93%
Our Proposed	Transformer based DNN	99.2 %

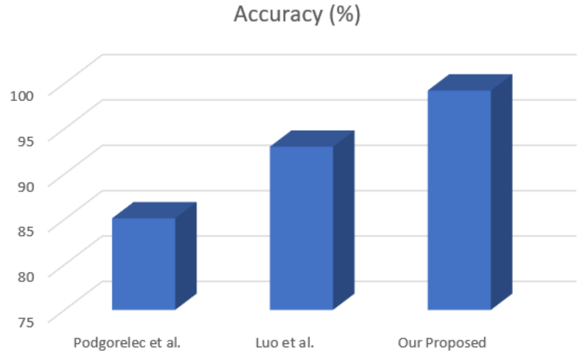


Figure 7. Performance Representation

6. Conclusion and Future Work

The proposed framework for multimedia data mining, leveraging the Transformer-based intelligent DNN model architecture, represents a significant advancement in enhancing the accuracy and efficiency of multimedia data analysis. By synergistically integrating the Transformer architecture with deep neural networks, this framework excels in extracting and representing informative features, while its intelligent learning approach enables adaptation to complex and dynamically changing data patterns. Such a combination holds the potential to revolutionize multimedia data mining across diverse domains, including image and video analysis, by substantially elevating the performance of data mining tasks.

As we look towards the future, a promising avenue for research lies in exploring the scalability and robustness of this framework, particularly in handling increasingly larger-scale multimedia datasets. Investigating methods to efficiently parallelize and distribute computations over distributed systems would be crucial to unlock its full potential in tackling big multimedia data. Furthermore, emphasis should be placed on devising techniques that effectively mitigate potential bottlenecks and memory constraints that arise in dealing with extensive multimedia data. Moreover, extending the framework to embrace multimodal data fusion could be a stimulating direction. By simultaneously incorporating information from various modalities, such as text, audio, and visual data, the model could achieve a more comprehensive understanding of multimedia content, leading to more accurate and comprehensive data mining results. Additionally, continual improvements to the model's architecture and learning mechanisms should be pursued. Exploring advanced attention mechanisms, transformer variants, and novel self-supervised learning techniques could further boost its performance and adaptability, pushing the boundaries of multimedia data mining even further.

Acknowledgement

The authors would like to thank the Jawaharlal Nehru Tech University, Hyderabad, for supporting this work.

Funding

This work is not funded by any funding agencies.

Declarations:

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Author's Contributions

Each author has equally contributed in this research work.

References

[1] M. Stamenovic, S. Schick, and J. Luo, "Machine identification of high impact research through text and image analysis," in *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*. IEEE, 2017, pp. 98–104.

[2] D. Sridevi, D. A. Pandurangan, and D. S. Gunasekaran, "Survey on latest trends in web mining," *International Journal of Research in Advent Technology*, vol. 2, no. 3, 2014.

[3] P. Mahani, N. Ruhil *et al.*, “Web data mining: A perspective of research issues and challenges,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2016, pp. 3235–3238.

[4] R. Rekik, I. Kallel, J. Casillas, and A. M. Alimi, “Assessing web sites quality: A systematic literature review by text and association rules mining,” *International journal of information management*, vol. 38, no. 1, pp. 201–216, 2018.

[5] Y. Li, “Research on technology, algorithm and application of web mining,” in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, vol. 1. IEEE, 2017, pp. 772–775.

[6] I. Khan, A. Khan, and R. A. Shaikh, “Object analysis in image mining,” in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2015, pp. 1985–1988.

[7] A. Yoshitaka and T. Ichikawa, “A survey on content-based retrieval for multimedia databases,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 1, pp. 81–93, 1999.

[8] M. Liang, X. Cao, J. Du *et al.*, “Dual-pathway attention based supervised adversarial hashing for cross-modal retrieval,” in *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2021, pp. 168–171.

[9] M. S. Islam, M. N. Islam, N. Hashim, M. Rashid, B. S. Bari, and F. Al Farid, “New hybrid deep learning approach using bigru-bilstm and multilayered dilated cnn to detect arrhythmia,” *IEEE Access*, vol. 10, pp. 58 081–58 096, 2022.

[10] W. Li and X. Fan, “Image-text alignment and retrieval using lightweight transformer,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4758–4762.

[11] D. Martín-Gutiérrez, G. Hernández-Peñaloza, A. B. Hernández, A. Lozano-Diez, and F. Álvarez, “A deep learning approach for robust detection of bots in twitter using transformers,” *IEEE Access*, vol. 9, pp. 54 591–54 601, 2021.

[12] Z. Wang, Z. Yin, and Y. A. Argyris, “Detecting medical misinformation on social media using multimodal deep learning,” *IEEE journal of biomedical and health informatics*, vol. 25, no. 6, pp. 2193–2203, 2020.

[13] W. Xu, J. Yu, Z. Miao, L. Wan, Y. Tian, and Q. Ji, “Deep reinforcement polishing network for video captioning,” *IEEE Transactions on Multimedia*, vol. 23, pp. 1772–1784, 2020.

[14] K. Somandepalli, T. Guha, V. R. Martinez, N. Kumar, H. Adam, and S. Narayanan, “Computational media intelligence: Human-centered machine analysis of media,” *Proceedings of the IEEE*, vol. 109, no. 5, pp. 891–910, 2021.

[15] X. Wu, T. Tanprasert, and W. Jing, “Image classification based on multi-granularity convolutional neural network model,” in *2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 2022, pp. 1–4.

[16] S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, and Q. Tian, “Graph regularized encoder-decoder networks for image representation learning,” *IEEE Transactions on Multimedia*, vol. 23, pp. 3124–3136, 2020.

[17] S. Walia, K. Kumar, M. Kumar, and X.-Z. Gao, “Fusion of handcrafted and deep features for forgery detection in digital images,” *IEEE Access*, vol. 9, pp. 99 742–99 755, 2021.

[18] S.-C. Chen, “Embracing multimodal data in multimedia data analysis,” *IEEE MultiMedia*, vol. 28, no. 3, pp. 5–7, 2021.

[19] Y. Djenouri, A. Belhadi, G. Srivastava, and J. C.-W. Lin, “Deep learning based hashtag recommendation system for multimedia data,” *Information Sciences*, vol. 609, pp. 1506–1517, 2022.

[20] A. Singh, V. Tiwari, and A. N. Tentu, “A machine vision attack model on image based captchas challenge: Large scale evaluation,” in *Security, Privacy, and Applied Cryptography Engineering: 8th International Conference, SPACE 2018, Kanpur, India, December 15-19, 2018, Proceedings 8*. Springer, 2018, pp. 52–64.

[21] A. Singh and V. Tiwari, “An optimal dimension reduction-based feature selection and classification strategy for geospatial imagery,” *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 6, no. 2, pp. 120–138, 2019.

[22] N. Arappal, A. Singh, and D. Saidulu, “A soft computing based approach for pixel labelling on 2d images using fine tuned r-cnn,” in *International Conference on Innovations in Computer Science and Engineering*. Springer, 2022, pp. 415–424.

[23] “https://www.kaggle.com/datasets/sidharkal/sports-image-classification.”

[24] V. Podgorelec, Š. Pečnik, and G. Vrbančič, “Classification of similar sports images using convolutional neural network with hyperparameter optimization,” *Applied Sciences*, vol. 10, no. 23, p. 8494, 2020.

[25] K. Luo, “Elements and construction of sports visual image action recognition system based on visual attention analysis,” in *2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*. IEEE, 2021, pp. 411–414.

Mogili Ravi is Research Scholar in JNTUH Hyderabad. He has about 15 years of Professional, Teaching and Research experience. His areas of interest are: Web data mining and Object Oriented Design and Programming.

Mandalapu Ekambaram Naidu is the Principal of SRKIT, Vijayawada. He has about 36 years of Industrial, Professional, Teaching and Research experience. His areas of interest are: Signal and Image Processing, Computer Vision, Pattern Recognition and Analysis, Cybernetics and Informatics, Symbolic Computation, Computer Networks, Software Engineering, Object Oriented Design and Programming.

Gugulothu Narsimha is the Principal & Professor in the Department of CSE JNTUH College of Engineering Sultanpur. He has about 23 years of experience in teaching, research and administration in well-reputed educational institutions like JNTU-Kakinada and JNTU-Hyderabad alongside private institutions. His research interests are Network Security, Computer Networks, and Data Mining.