International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I3P126

Volume 74 Issue 3, 376-387, March 2026

Original Article

Tourism MATE-LLM: Multimodal Cross-Attention Fusion of Skip-Gram and ConvNext Embeddings for Tourism Destination Recommendation

V Indumathy[1], K Shantha Kumari[2]

[1,2]*Data Science and Business Systems, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.*

[1]*Corresponding Author : im3688@srmist.edu.in*

Received: 06 December 2025 Revised: 27 January 2026 Accepted: 06 February 2026 Published: 28 March 2026

Abstract *- Multimodal data in the Tourism domain requires a precise representation with useful content. Tourism requires data from huge sources to satisfy the tourists' demands. From the basic needs to luxury things, tourists demand information about the tourist area for planning. This requires the integrated representation of the tourist area that covers the various factors demanded by tourists. Based on this objective, the proposed work uses a cross-attention model to fuse the multimodal data using the ConvNeXt extracted image features and text feature extraction using the Skip Gram model. The cross-attention model takes the important correlation factors between the different data inputs based on the weights and feature values. The proposed work attained 58% data fusion using two modal type datasets based on tourism and produced =an accuracy of 85.28% using the India Tourism dataset from Kaggle and the India Tourist destination dataset from Mendeley.*

Keywords *- ConvNeXt, Cross attention model. Data fusion, Multimodal, Tourism.*

1. Introduction

By 2024, India stood at the 39th rank among 119 countries for tourism as reported by the World Economic Forum (WEF). The tourism in India is growing at a larger scale, where the tourists from foreign countries have increased in millions from 2020 to 2024. Tourism in India is rising mainly due to factors such as culture, heritage, mysticism, spirituality, and professional developments. The activity of tourists is the movement of people from their home or homeland to other places, for leisure or economic purposes. Based on the tourism activity, the domains like transport, healthcare, food, restaurants, etc, are gaining more profits. The demand for tourism has increased the interest of humans to know the areas before they visit those places, book the mode of transport, places to dwell, choose the foods of their interest, and hospital facilities in the location, etc. All these are obtained using the information that is available on the internet. The data required to represent all this information is of a multi-modal type that needs to be represented to the visitors to gain their interest in the tourist place. The past visits of the person in the locality are also used to consider the visits.

Tourism in recent years requires multi-modal information and huge resources that collect the information, recent trends, activities, and important places to visit in the locality. Technologies like media, IoT (Internet of Things), Smart tourism, E-Tourism, Large Language models, etc, are playing greater roles in the collection of information, interpretation of the collected knowledge of the places, and responses to the queries raised by the tourists. Based on queries and reviews from tourists, demand for information about places increases exponentially. It is also observed that tourists require complete information about the tourist place in a concrete way, rather than providing redundant and unwanted information. This kind of data integration is possible with the data fusion techniques that combine the multi-modal information into a single representation that helps tourists understand the places and activities. Large Language Models (LLMs) provide a transformative boon in the tourism department by enabling intelligent services to tourists by being context-aware and providing human-like responses based on the information trained using the datasets. The Intelligent Travel Assistants help the tourist by means of language translation, chatbots that provide customer service at all times, tourism planning based on the customer's preferences, and real-time travel advice. Thus, the LLM intervention in tourism makes it a Smart AI-based travel system that highly satisfies the tourist expectations.

The unimodal system dependency reduces the accuracy of the prediction of the recommendation systems for tourism. Involving heterogeneous information provides accurate insight into the expectations of the tourists, and hence, fusion of multimodal data stands as an effective strategy in handling

enormous data. Hence, this work addresses the image-text fusion using a cross-attention model by a feature-level fusion model with the help of the Skip Gram model and ConvNeXt architecture in the LLM-assisted scenario for the Tourism domain. The objective of the work is to obtain an efficient data fusion algorithm for LLM that should contain useful information from heterogeneous data of multimodal tourism-based information. The paper is organised as follows: i) Introduction, ii) Related work, iii) Data fusion in Tourism, iv) Cross Attention Model, v) Proposed work, vi) Results and Discussions, and vii) Conclusion.

2. Related Works

The areas like point of interest, travel plan, and demand forecasting are important in the Tourism domain, where multimodal data fusion helps to represent the whole data in an integrated form. In [1], various data fusion systems operating in the tourism domain based on hierarchical, hybrid, early, and late fundamentals, datasets, and their statistics, and the involvement of federated learning applications in the tourism domain are surveyed. The ensemble learning [2] method can be used to fuse the multimodal data that shows good learning ability with high accuracy for the Ctrip dataset, which helps in the future prediction for tourism support systems. The real-time processing of online information and reviews based on the word vector obtained from the seed value of emotional sentiment analysis is used for the data fusion of multimodal information from various sources at the feature level, as proposed in [3], which gives the overall performance of the product providers in the online mode. Sentimental analysis of the multimodal data input is done using Weibo in [4], where the event awareness-based framework is designed using the multimodal data using the cross-modal process. The sentiment analysis is combined with the data fusion techniques in the multi-modal data, which are reviewed in [5], where the design using various architectures and designs is formulated, and provides the future scope.

The multimodal data for the demand of tourism constraints on the factors of pricing of transport, fluctuation in the economy, economic crisis, and the demand are fused based on the multivariate time series model that presents a complete form following the feature extraction, classification, and decision making of the ensemble process developed in [6]. The tourism recommendation requires huge amounts of multimodal data like reviews, images, and videos, which makes it practically impossible to present the vast data, making the system inefficient. The Belts (Bidirectional Long Short-Term Memory) is used to extract the text data, and Eine (External Attention Transformer) helps to find the relationship between the features, which are used to represent the vast data into integrated information, is proposed [7]. The complete landscape representation of the tourist spot is an efficient way of portraying the location information, as explained in [8], where the various locality-based data are added and combined into a single landscape design using Computer-Aided Design. In the tourism domain, important parameters like tourism resilience [9] are used to find the location availability based on area, health, and landscape. Providing multimodal data about the location paves the way for an efficient way of representing the tourism system. The characteristics like text, number, and images are used for computing the tourism domain based on the linguistic approach of the Spanish language, as done in [10], where the visualisation output is obtained after processes like feature extraction, classification, and data fusion.

In recent trends, the tourist spots are becoming famous based on the reviews from those who have previously visited them. In [11], the review-based approach is used to find accurate information about the spot using the LLAVA algorithm, and retrieval-based fine-tuning is used to represent the knowledge of reviews based on sentiment analysis, review strength, rating, data, and year. A Sentimental Aware Topic model algorithm is proposed in [12] for the fusion of data obtained from the reviews, landscapes, and availabilities, based on which the customer sentiment analysis is used to select the spot for tourism. The transfer learning of the information about the travels is studied in [13], where the datasets used before and after the pandemic are analysed using the multi-modal 2-step floating catchment area, showing better results compared to conventional methods of not using the multimodal dataset. The hyperspectral image classification is done using the transformer that uses the self-attention model outputs to find the cross relationship between the features of the images, which is proposed in [14], where the proposed system performs classification with higher accuracy compared to the self-attention model. From the survey, it is seen that the

Convolutional Neural Networks are efficient in the feature extraction processes. This feature extraction is modified to find the relationship between the features and represent it in the latent space, which is proposed in [15], where the hyperspectral image produces higher efficiency. Pedestrian detection [16] is used in a surveillance system that needs multiple data sources to detect a cause, in which one data source cannot dominate the other. Every modal data contains a certain amount of information required for detection. The Cross-modal attention transformer is involved in pedestrian detection that fuses multiple data sources from the dataset, and it has a good representation of features compared to the other systems. Deep Fusion Transformer [17], based on the cross-modal analysis, is used to represent the visual data using the multi-feature analysis from the multimodal datasets. The visual and semantic features are fused using the cross-attention model transformer. Urban Computing [18] is the obtaining of demographic information representing the urban area in a simple form that uses multiple form of data from different sources like media, traffic, datasets, etc. is analyzed and performed cross-domain data fusion that helps the multiple modals to represent the urban area in a simplistic visual representation that helps in various

sectors like tourism, vehicular networks, geolocation applications, etc. The customer support services require the sentiment analysis system to produce an accurate output that meets the expectations of the clients, which is analysed using multi-modal data in the decision level processing using the conversation dataset between the client and the support centres.

From the survey of the literature, it is clear that the fusion of tourist information is done by the concatenation and lacks the processing of multimodal fusion, like tourist text queries, travelling spot images, and other statistical details. Hence, the queries of the tourist, along with the location images, provide great insight for the tourist recommendation processes. Hence, this work undergoes feature-level fusion of text and image using the cross-attention model, where the features are extracted from the Skin Gram model and ConvNeXt model.

3. Data Fusion in Tourism

Tourism is the travel of people from one locality to another, known or unknown places, for vacation, leisure, or professional purposes. The domain of tourism is an integral part of various factors that are involved in it. The planning of tourism requires important information like its geographical area, location, and sites to visit in that locality. In terms of staying purposes, the people require enhanced health care centres, food, and restaurants. The main criteria in tourism are to analyse the budget based on the transport fare, transport facilities in the locality, expenses, etc. These factors play a major role in the fixation of tourism, which helps the tourist to have a successful trip based on the digital information provided. Figure 1 depicts the criteria that the tourism department of any country or locality must fulfil to gain more tourists, which, in turn, increases the revenue of that locality.

The collection of these data is crucial and needs more data and time to represent it. The higher the amount of data from various factors, the better the tourism predictions and forecasts are. Handling huge amounts of data requires high storage, and it makes the system more complex. Figure 2 is the data fusion process that fuses the multimodal data from various sources of information, comprising various factors for tourism. The data collected is of a multi-modal data type from heterogeneous sources that need to be fused to obtain the integrated data output that must be retrievable and should contain all important information from the input sources based on the tourists' demand. The application-based outputs also needed to be integrated rather than the complete tourism information alone. The modal like text, image, audio, and video are present to represent the tourist places. This enhances the places and makes the tourist decide based on their requirements. Data fusion not only integrates the data but also helps the communication system to process less data, which reduces the payload and handles data traffic efficiently.

Fig. 1 Factors of tourism

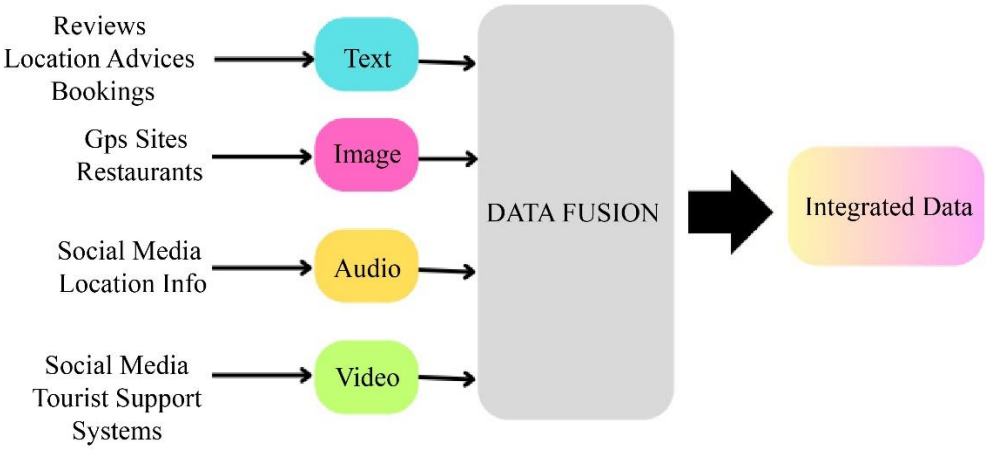

Fig. 2 Data fusion in tourism

4. Cross Attention Model

The cross attention model works on the features of different input data, obtains the relation, and produces the correlation output. It works the same as in the self-attention model, obtaining the parameters like query, key, and value. The description of the parameters is described below,

- A query is an embedding-based vector that concentrates on the relevant information that the user is seeking.
- The key is the label or identifier that obtains the feature that is more relevant to the query.
- Value forwarding element allows the attention from the query and key to the next level.

The query and key are parameters that gain the attention of the system to obtain the valid result of the user's request. The similarity score matrix is obtained with weights based on the biases of the relevant information, which is multiplied by the query and the key. The final scalar product is done with the value and attention score matrix. All the above-described function takes place for the conventional attention model. In the cross-attention model, the query from one type of modal is multiplied by the key of the other multimodal data type to obtain the attention score matrix. This cross-correlated attention score matrix is scalar multiplied by the value of another multi-modal data. Thus, this model helps correlate one feature to another and provides a single representation of multimodal data.

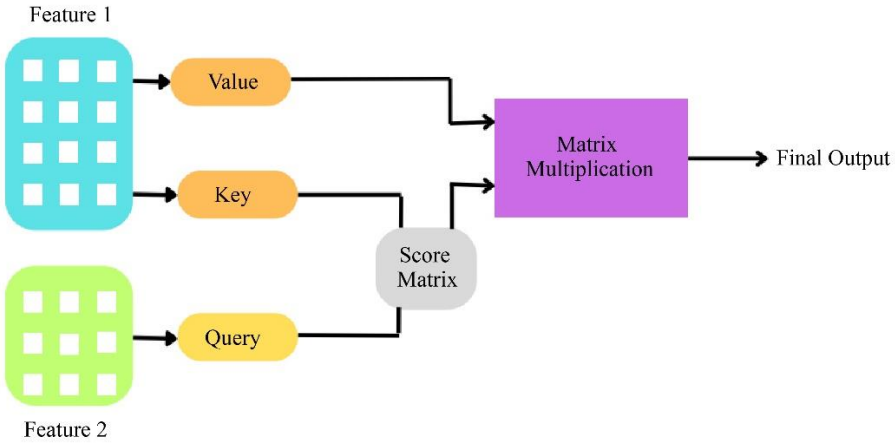

Fig. 3 Cross-attention model based on features

5. Proposed Work

The data fusion takes place at three levels: raw data fusion, feature level fusion, and decision level fusion. In the raw data fusion, the complete information obtained from the sources is fused, which increases the complexity of the system, as the raw data does not undergo data preprocessing. In the decision level fusion, the probabilities of occurrence of data are taken for fusion, which has less computation complexity but produces less accurate output. The middle one is the feature level fusion, where the input information is preprocessed, from which the features are extracted. These extracted features undergo fusion and are represented in the integrated form. This integrated data is used for decision-making processes. Figure 4 depicts the proposed method of

LLM based on Multimodal data fusion of text and image input using the ConvNext and Word2vec feature extraction model. Two different inputs based on the tourism information are taken for processing.

The tourism image-based information is fed into the ConvNext [20] feature extraction model. The architecture of the ConvNext model is illustrated in Figure 5, which is the pure form of CNN (Convolution Neural Network) that narrows the gap between the CNN and the transformers. It provides higher accuracy compared to ViT transformers. It processes the data from tiny to a scalable manner, utilising the advantages of ResNet and transformer design. Once the input images are fed into the ConvNext model, it makes the tiny patches, followed by four stages consisting of down-sampling and 2×2 convolution. Layer normalisation is used instead of batch normalisation in CNN. The expansion of the channel is done from C to 4C. GELU (Gaussian Error Linear Unit) is used in ConvNext, which provides a smooth and nonlinear activation function that helps in preserving the partial negatives, which are absent in the ReLU activation.

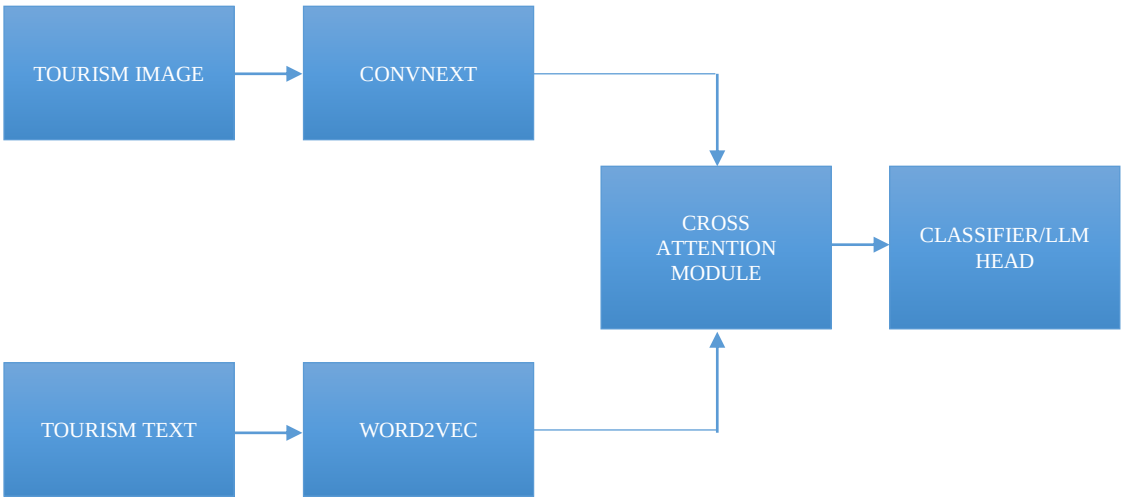

Fig. 4 Proposed LLM-based data fusion system for tourism

Figure 6 depicts the Skip-Gram model for the Word2Vec model, which is used for extracting the features from the input text information regarding tourism. Using LLM, the tourist or the customer uses the target word to obtain information about the location in such a case.

The Skip-Gram model acts as the best option, as it produces the context-based output from the target words. The learning happens by observing the words in the fixed window size that are nearby, and the embeddings are obtained.

The input target words are tokenized, and indices are assigned for the unique words from the vocabulary. In one-hot encoding, the binary vectors corresponding to the vocabulary are created. Using the weight matrix, the one-hot encoder vectors are converted to dense embeddings.

The hidden layers help to find the context words without an activation function. SoftMax is used to obtain the probabilities for each unique word and its one-hot encoder vector. Negative sampling plays a major role in the skip-gram model by reducing the dissimilar content in the context word based on the target word. In the optimisation process, the negative sampling is removed, which produces the real context words using the binary classification task. At the end, the features are obtained from the embedding matrix that contains the semantic relationship among the target word, which contains the maximised similarity rather than negative samples that have dissimilar words.

The output of the feature vectors from the ConvNext and skip-gram word2vec model is fed into the cross-attention model. As the proposed work is based on feature fusion, the cross-attention model fuses both the feature vector outputs and produces a single latent representation that indicates both features. As discussed above, the input image is considered as feature 1 that extracts the vectors using the ConvNext, which is used to obtain the key and value in the cross-attention model. Using the text-based feature vectors, the query for the cross-attention model is obtained.

Query and key parameters help to extract the relevant information from the given input, whereas the value parameter forwards the obtained query and key parameters to the next range. The attention scores are obtained by the dot product of the query and key parameter, which produces the matrix multiplication output, and the SoftMax function produces the attention weights that infer the vector values representing the query and its corresponding key parameter. These outputs are fed into the LLM head or classifier that produces the required output that represents the feature fusion of the given input image and the text.

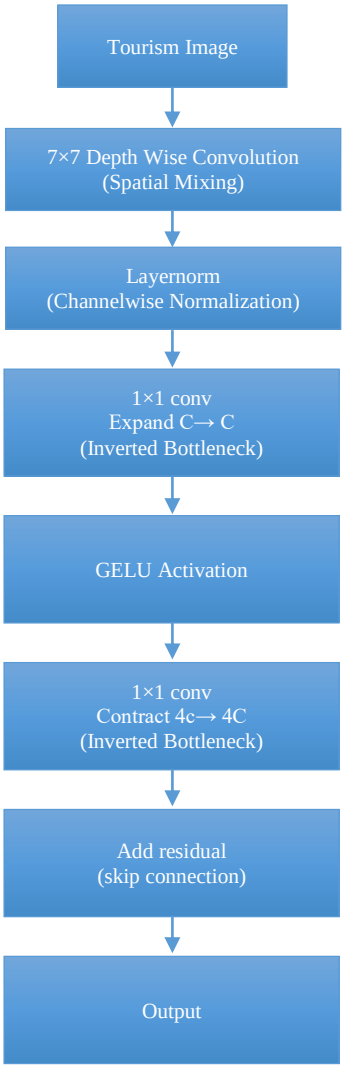

Fig. 5 ConvNeXt architecture

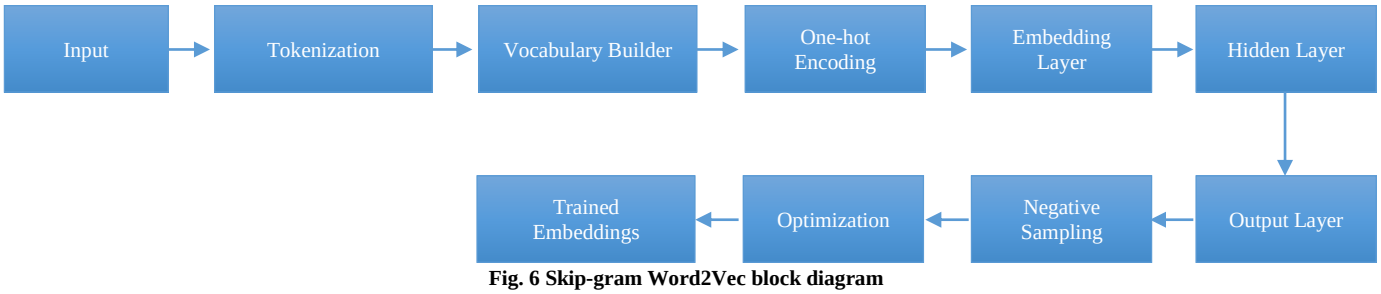

Fig. 6 Skip-gram Word2Vec block diagram

5.1. Mathematical Formulation

Let us consider that I be the input image based on tourism, and T be the Tourism Text based on the query, therefore T={w_1,w_2,……..w_n}. Let N_i represent the number of patch-wise tokens; the features based on ConvNext are represented by:

$$F_i = ConvNeXt(I) \quad (1)$$

Here in the equation (1), F_i is the image features extracted using the ConvNext operation of dimension N_i×d, where d is the feature dimension obtained after the linear projection. The Skip Gram model-based feature extraction of text input is given in equation (2),

$$x_i = Embeddings\ (w_i) \quad (2)$$

$$F_T = [x_1, x_2, \ldots\ldots\ldots\ldots\ldots x_T] \quad (3)$$

In equation (2), the embeddings based on the skip-gram model are obtained and are stacked in equation (3), where the F_T is the feature extracted output of the given input text of dimension $N_T \times d$. Now these feature vectors are given as input to the cross-attention model. The parameters of the cross-attention model are represented in equations (4), (5), and (6).

$$Q = F_T . W^q \quad (4)$$

$$K = F_i . W^k \quad (5)$$

$$V = F_i . W^v \quad (6)$$

The functions W^q, W^k, and W^V are the weight coefficients for query, key, and value of dimensions $N_T \times d_k$, $N_I \times d_k$, and $N_T \times d_v$. The attention score using the dot product of query and key is obtained by equation (7),

$$A = softmax\left(\frac{Q.\ K^T}{\sqrt{d_k}}\right)V \quad (7)$$

Thus, the attention scores are obtained using the above formulation, which is the feature-fused output of the image and the text. The fused output is provided in equation (8).

$$F_{fused} = \{A_1, A_2, \dots \dots A_{d_v}\} \quad (8)$$

Algorithm

i. Obtain the input image (I) and text (T).
ii. //Image Feature Extraction
iii. Load the pretrained ConvNext model for feature extraction of the image
iv. Preprocess the Input Image.
v. Extract the features using the ConvNeXt model of the preprocessed image
vi. F_i=convNeXt($I^{'}$)
vii. //Text Feature Extraction
viii. Tokenize T into words, T={w_1, w_2,……w_n}
ix. Retrieve skip-gram embedding for each word, e_i
x. Obtain the features of text F_t= {e_1, e_2……..e_n}
xi. Concatenated the features of text and image using the cross-attention model using the attention matrix, F= {A_1, A_2…. An}
xii. Aggregate the features using the mean pooling
xiii. F=meanpooling (F_i)
xiv. Feed the fused features into the classification head
xv. Prediction using, y=argmax (SoftMax (MLP(F)))
xvi. Obtain the evaluation metrics and loss function

6. Results and Discussions

6.1. Dataset

The India Tourism 2014-2020 from Kaggle is taken as text-based input to the data fusion model. This dataset contains information on foreign visits to the country in a quarterly manner from 2014 to 2020. This dataset comprises the foreigners' interest in visiting India. It mainly focuses on the economic and marketing policies for the development of Indian tourism. It comprises information like gender, quarterly visits by foreigners, places, states, currency, dollars, and visitors' way. Indian Tourist destination dataset [22] from Mendeley comprised of 5000 images of the tourist places in all the states of India, based on the classes like beaches, hill stations, monuments, temples, etc. These images help the tourist to fix the spot for travel. It helps the tourist to select the destination based on adventure, eco, and spirituality.

6.2. Implementation

The text feature extraction of the India Tourism dataset from 2014-2020 has been undergone using the PyTorch library to obtain the semantic features of the tourist interest in the 6 years. The embedding dimension of 300 of a window size of 5-10, using 10 negative samples with a minimum word frequency of 10, is used for feature extraction for 1-20 epochs. The image extraction is implemented using the ConvNeXt model using Torch Vision. Models. Convnext_tiny to obtain the features from the Indian Tourist Destination dataset from Mendeley. The input resolution of 224×224 pixels at the last global average pooling layer produces the output dimension of 768D, considering the tiny variant. Missing values are handled by using only the available features for the fusion processes, either the image or the text. Table 1 represents the simulation parameters used for the fusion of image and text features using the cross-attention model with the classifier output.

Table 1. Simulation parameters

Cross Attention Module	
PARAMETERS	VALUES
HEADS	8
QUERY	TEXT
KEY/VALUE	IMAGE
OUTPUT	512D fused vector
DROPOUT	0.1-0.3
LAYER NORMALIZATION	POST-ATTENTION MODEL
Skip Gram Module	
EMBEDDING DIMENSION	300
CONTEXT WINDOW SIE	50

WORD FREQUENCY	10
Convnext Module	
MODEL	TINY
INPUT RESOLUTION	224×224
EMBEDDING DIMENSION	768
Classifier	
INPUT	512D FUSED VECTOR
HIDDEN LAYER	1-2FC
ACTIVATION FUNCTION	ReLU
LOSS FUNCTION	CROSS ENTROPY
OPTIMIZER	ADAM
LEARNING RATE	3e-5
WEIGHT_DECAY	0.01
BATCH SIZE	32
TRAINING/VAL/ TEST	75%/15%/10%

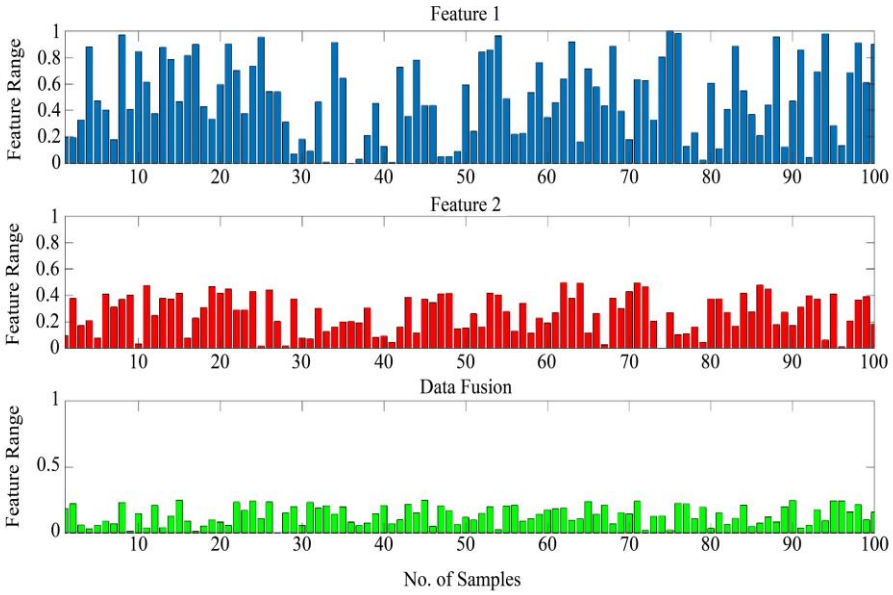

Fig. 7 Feature ranges after extraction and data fusion

Figure 7 is the illustration of the feature values obtained after extraction using the Skip Gram model and ConvNeXt model. It is observed that the image features are higher compared to the text features. These features are given as input to the cross-attention model. The text feature extracted is given as input to the query of the cross-attention model, and the key and value are assigned to the image features input.

These features are then combined via matrix multiplication, using the dot product of the query and key, which is applied to the attention mechanism by multiplying the value parameter. The output of the cross-attention model is represented in the plot as the data fusion output for 100 samples. These samples are a fused representation of image and text features that help the LLM-based approach obtain meaningful information from multimodal inputs.

Table 2. Sample outputs

S. No	Place	State	Fused feature	Prediction	Output
1	Jaipur	Rajasthan	196×768	Culture	"Jaipur is famous for its forts, palaces, and vibrant heritage, also called the pink city of India.
2	Allepey	Kerala	196×768	Nature	"Allepey, surrounded by coconut trees and paddy fields, offers peaceful backwater cruises."

3	Hampi	Karnataka	196×768	Heritage	"Hampi is the historical location featuring ancient stone temples."
4	Jim Corbett National Park	Uttarakhand	196×768	Wildlife	"Jim Corbett National Park is a premier wildlife reserve and is rich in biodiversity."
5	Havelock Island	Andaman and Nicobar	196×768	Leisure	"Havelock Island contains coral reefs and a pristine beach, which is a Serene destination."

Table 3. Confusion matrix for 5 classes of tourism

Actual/ Predicted	Heritage	Nature	Wildlife	Leisure	Culture
Heritage	91%	0%	0%	9%	0%
Nature	0%	80%	8%	12%	0%
Wildlife	0%	4%	96%	0%	0%
Leisure	0%	19%	0%	81%	0%
Culture	19%	0%	0%	0%	81%

Table 2 is the example obtained using the fusion of two datasets before the LLM, and the output obtained for the data fused input using the cross-attention model. Five classes of output are obtained, where the samples are provided for each class. It is seen that culture, heritage, nature, wildlife, and leisure represent the classes of the places present in the states of India. The dimension of the fused vector after the cross attention model and the LLM output for the given input of the fused data is present in the output column. Table 3 is the confusion matrix obtained for the proposed work of data fusion using the cross-attention model before the LLM-based approach. The 5 different classes are obtained based on the given inputs that are fused using the cross-attention model. The feature values of the text and image help in the improved accuracy of the proposed work. It is seen that the Heritage class and Wildlife class produced more accurate output compared with other classes. It is seen that the proposed work misclassifies the wildlife and nature classes, but there is a major misclassification that happens in the culture and heritage classes of the places in India. Table 4 shows the evaluation metrics value obtained using the confusion matrix.

The Precision helps to find the true positive output from the total predicted outcomes. In the confusion matrix, the nature and leisure classes have high false positives compared to other classes. The culture class obtained 100% precision because it had no false positives, followed by the wildlife class, which had fewer false positives and attained a high precision accuracy of 92.3%. Recall is the parameter that reflects the positive instances of the classes; thus, the wildlife shows 96% of positive inference of low chances of a few false negatives. Heritage has 91% recall percentage, with the second class having fewer false negatives. The F1 score helps in the evaluation of the overall performance of the proposed work, which signifies the importance of true positives and false positive rates caused by misclassification. It is seen that the F1 score is very low for the Nature class and Leisure class, which are misclassified by the other classes. The proposed work approaches 85.28% of accuracy, and the evaluation metrics for classes are represented in Figure 8 for comparison. The Wildlife class showed more appropriate results for all the metrics compared to the other classes. The Culture class has a high precision percentage, but it has high false negatives in the Heritage class.

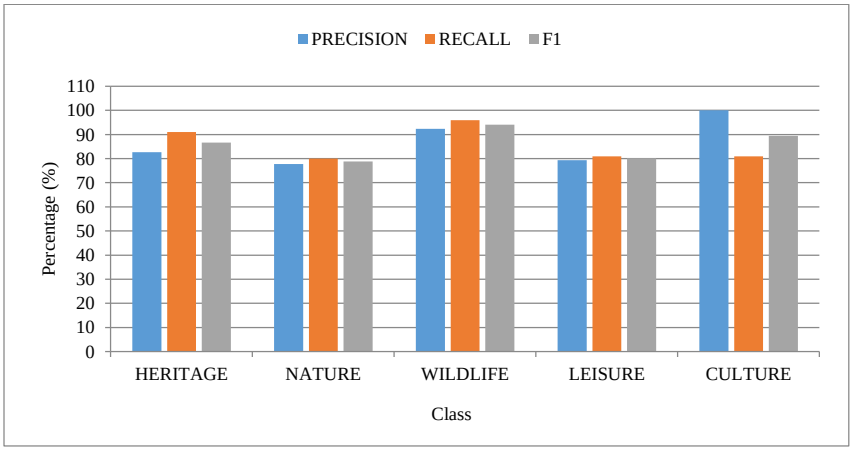

Fig. 8 Comparison of 5 classes for evaluation metrics

Table 4. Evaluation metrics

Class/metrics	Precision	Recall	F1
Heritage	82.7%	91%	86.6%
Nature	77.7%	80%	78.8%
Wildlife	92.3%	96%	94.1%
Leisure	79.4%	81%	80.2%
Culture	100%	81%	89.5%

Figure 9 is the error rate obtained for the proposed work using the cross-attention model for data fusion for the LLM-based approach. There is a fall from 35 to 15% of error rate for 12 epochs, and it tries to saturate at 14.7% of the error rate till 20 epochs. The error rate is quite high, mainly because the data fusion percentage of the proposed work attained only 58% using the feature vectors. The loss in the feature fusion leads to an increase in the error rate of the proposed work.

Table 5. Ablation study

Ref	Contribution	Accuracy Percentage
[3]	Text only	71.6%
[7]	Image only	74.3%
[2]	Feature concatenation	79.8%
[12]	Decision level fusion	82.7%
Proposed work		85.28%

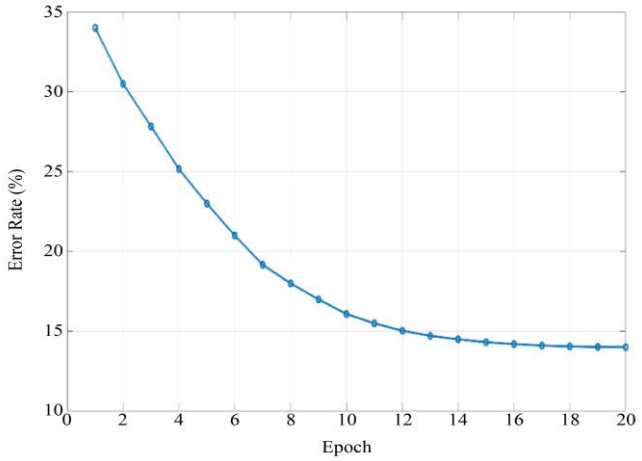

Fig. 9 Error rate of the proposed work

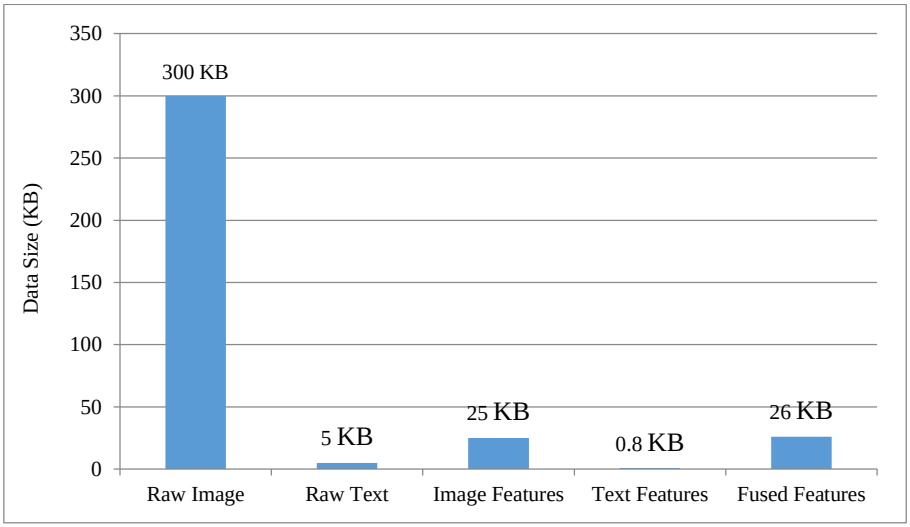

Fig. 10 Data size representation of the proposed work

Figure 10 is the data size comparison of the raw image and text with the extracted and fused feature data size. The proposed work reduces the data size for the interpretation of the information obtained from the datasets, which reduces the computational complexity of the system at a larger scale. The features extracted are of a smaller data size with high information content, which produces an accuracy of about 85.28% using the data fusion percentage of 58%. Table 5 provides the ablation study of the proposed work with the unimodal, concatenate fusion, and decision-level fusion. The features from the Skip Gram model obtain the text embeddings, and the ConvNeXt model, using the higher level depthwise convolution, produces the image-based embeddings that effectively work on the Query, Key, and Value network in the cross attention model, which effectively fuses the information, which remains the reason for its high accuracy compared to other models.

7. Conclusion

The data fusion technique is applied to the tourism data of multimodality from heterogeneous data sources using the ConvNext for image feature extraction, text feature extraction

using the Skip Gram model, and a cross-attention model for correlating and fusing the data. The proposed work is simulated and presents the output showing the data fusion levels, fusion percentage, accuracy, and error percentage plot. From the output, it is seen that the proposed work performs data fusion of 58% on average for testing with two input datasets of text and image. Evaluation metrics are calculated using the confusion matrix, which shows there is less misclassification among the 5 classes of prediction. The data size reduction due to the data fusion process reduces the computation complexity of the system. The extension of the proposed work will be carried out for sentiment analysis of the fused data using the LLM approach to improve the Tourist experience based on their expectation.

References

[1] Qazi Waqas Khan et al., "Multi-Modal Fusion Approaches for Tourism: A Comprehensive Survey of Data-Sets, Fusion Techniques, Recent Architectures, and Future Directions," *Computers and Electrical Engineering*, vol. 116, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] YaoGuang Li, and HeChi Gan, "Tourism Information Data Processing Method based on Multi-Source Data Fusion," *Journal of Sensors*, vol. 2021, no. 1, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Meng Li, "Research on Extraction of Useful Tourism Online Reviews based on Multimodal Feature Fusion," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Lijuan Wang et al., "Multimodal Event-Aware Network for Sentiment Analysis in Tourism," *IEEE MultiMedia*, vol. 28, no. 2, pp. 49-58, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ankita Gandhi et al., "Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions," *Information Fusion*, vol. 91, pp. 424-444, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Hongwei Wang, and Wenzheng Liu, "Forecasting Tourism Demand by a Novel Multi-Factor Fusion Approach," *IEEE Access*, vol. 10, pp. 125972-125991, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yuhang Cui, Shengbin Liang, and YuYing Zhang, "Multimodal Representation Learning for Tourism Recommendation with Two-Tower Architecture," *Plos one*, vol. 19, no. 2, pp. 1-23, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Zhongyuan Yang, and Jiaping Chen, "Optimisation of Tourism Scenic Area View Planning and Design based on Multimodal Fusion," *Computer-Aided Design and Applications*, vol. 22, pp. 201-214, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yi Liu, Yougen Jiang, and Qiuju Luo, "Advancing Tourism Resilience and Data Science using Multimodal Data," *Journal of Policy Research in Tourism, Leisure and Events*, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Julian Monsalve-Pulido, Carlos Alberto Parra, and Jose Aguilar, "Multimodal Model for the Spanish Sentiment Analysis in a Tourism Domain," *Social Network Analysis and Mining*, vol. 14, no. 1, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Hiromasa Yamanishi, Ling Xiao, and Toshihiko Yamasaki, "A Multimodal Dataset and Benchmark for Tourism Review Generation," *18th ACM Conference on Recommender Systems*, Bari, Italy, pp. 1-19, 2024. [Google Scholar]

[12] Xi Shao, Guijin Tang, and Bing-Kun Bao, "Personalised Travel Recommendation based on Sentiment-Aware Multimodal Topic Model," *IEEE Access*, vol. 7, pp. 113043-113052, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yongqi Zhang et al., "How Does Multi-Modal Travel Enhance Tourist Attraction Accessibility? A Refined Two-Step Floating Catchment Area Method using Multi-Source Data," *Transactions in GIS*, vol. 28, no. 2, pp. 278-302, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Jinghui Yang et al., "A Cross-Attention-based Multi-Information Fusion Transformer for Hyperspectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 13358-13375, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Fulin Xu et al., "Bridging CNN and Transformer with Cross Attention Fusion Network for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips, "Cross-Modality Attention and Multimodal Fusion Transformer for Pedestrian Detection," *Computer Vision - ECCV 2022 Workshops Tel Aviv, Israel*, Tel Aviv, Israel, pp. 608-623, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Jing Zhang et al., "Cross on Cross Attention: Deep Fusion Transformer for Image Captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4257-4268, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Xingchen Zou et al., "Deep Learning for Cross-Domain Data Fusion in Urban Computing: Taxonomy, Advances, and Outlook," *Information Fusion*, vol. 113, pp. 1-38, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jinhu Qi et al., "Research on Tibetan Tourism: Viewpoints Information Generation System based on LLM," *2024 12th International Conference on Intelligent Computing and Wireless Optical Communications (ICWOC)*, Chongqing, China, pp. 35-41, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Zhuang Liu et al., "A Convnet for the 2020s," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976-11986. 2022. [Google Scholar] [Publisher Link]

[21] Tomas Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint*, pp. 1-12, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[22] Tejaswini Bhosale, "Indian Tourist Destination," *Mendeley Data*, vol. 1, 2024. [CrossRef] [Publisher Link]