Science & Technology Development Journal 2026, 29(1):3911-3918

Open Access Full Text Article

Research Article

Multitask learning based on attention and transformer mechanism for event recognition and importance image prediction in photo albums

Vo Hoai Viet[*], Le Quoc Viet

ABSTRACT

Capturing images has become a simple task due to the popularity and technological advancements in cameras and cellphones. As the quantity of pictures being taken increases, the task of organizing them while preserving their significance is becoming more challenging. Solving this problem requires creating a system that can identify the type of album, select the important photos to store, and automatically delete the rest. Such a system could also significantly reduce the storage requirements and create attractive story videos. In this study, we design a multitask network architecture that can simultaneously learn event recognition and image importance, thereby preventing the need for event-type information. This approach combines the strengths of convolution neural networks for image description with an attention and transformer mechanism for album description to perform both event recognition and image significance determination, providing a viable and effective approach with faster prediction times for both image importance and event identification. Our approach surpasses state-of-the-art methods by improving 3% on image importance tasks and achieving 67.21% accuracy on event recognition tasks in the ML-CUFED dataset. The results are evaluated on multiple backbones and parameters to demonstrate the generalization of the proposed methodology.

Key words: event recognition, image importance, transformer, multitask learning, attention network, photo album.

Computer Vision Department, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam

Correspondence

Vo Hoai Viet, Computer Vision Department, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam

Email: vhviet@fit.hcmus.edu.vn

History

- Received: 20-09-2024
- Revised: 07-08-2025
- Accepted: 27-09-2025
- Published Online: 06-01-2026

DOI :

https://doi.org/10.32508/stdj.v29i1.4361

INTRODUCTION

The prevalence of cameras today, specifically in mobile phones, has made taking pictures commonplace and practically effortless. People attending an important event, such as the Thanksgiving holiday with family and friends, may take numerous photos to capture the interesting moments that take place during the event. As a result, they may have a large number of photos from the holiday on their phone. If they want to share their meaningful moments with friends and family, they will have a laborious and time-consuming task of selecting significant photographs from their large collection of images. They will also have to face the same task of selecting important and meaningful photos before saving them to their computer or phone or uploading them to the cloud. For many people, manual album organizing is no longer practical; as such, there is a growing need for this task to be carried out automatically. The work of categorizing a particular set of images in a classical photo album into a predetermined list of noteworthy names is known as event recognition (e.g., birthdays, Christmas, vacations, sporting events, etc.). There are typically three main challenges associated with recognizing events in photo albums: (1) The analysis involves a collection of several disorganized photos, requiring seconds to hours or even days; (2) Common photo galleries typically contain ineffective images; (3) Image encoders must contain both low-level and high-level features, but currently available datasets are small and potentially insufficient to support representation learning. An additional challenge involves selecting relevant images in photo albums because the image importance of albums depends on the accuracy of event recognition. For example, if a person needs to choose the most significant photo from their vacation in Dubai, an image of the beach at Jumeirah Beach Residence is clearly important to keep. However, if the album is specific for a birthday, then the most significant photo would include the person having the birthday and the birthday cake, because the beach scenery is only in the background. To address this challenge, this research poses the following question, as visualized in Figure 1: "***Can we build a model that recognizes the event label and scores the importance of images?***".

Cite this article : Hoai Viet V, Quoc Viet L. **Multitask learning based on attention and transformer mechanism for event recognition and importance image prediction in photo albums**. *Sci. Tech. Dev. J.* 2026; 29(1):3911-3918.

Existing categorization methods previously proposed address different aspects of these challenges in different ways. Using basic or weighted averaging, multi-image analysis was proposed [1,2]. Recurrent neural networks (RNNs) were also employed in previous studies [3–6]. To recognize events in photo albums, RNNs must be able to capture global relationships between distant sequence components. Image importance ground-truth usage [3,5,7] was employed to determine how to use the most pertinent album photographs, although this necessitates thorough annotation of each image's relevance. By using an attention layer to compare photos to an event, Guo et al. [8] ignored this annotation in their proposed method. They learned the image significantly in an implicit manner when training the model. Nevertheless, the implicit expected visual significance was not assessed. They also suggested using a combination of different networks and hierarchical feature extraction to handle the image representation problem, which could require a more complicated solution. Savchenko [9] similarly utilized an attention to combining picture-embedding descriptors by a learnable model. However, their image descriptor also contained image-captioning information, which decreases performance and yields lower classification accuracy. In this study, we reveal significant findings on event classification and image importance on benchmark datasets and offer a feasible and effective multi-task learning solution based on a transformer and attention mechanism.

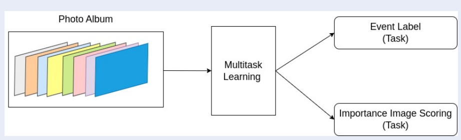

Figure 1: Overview of multitask learning in a photo album

Transformers [10] are increasingly being used as sequential data classification models. Transformers enable multi-image processing and use an attention technique to concentrate on the most crucial components. Information propagation is restricted, even if LSTMs [11] ideally attempt to store short and long information in memory, because the impact of the strongest gradients may fade. Transformers have enabled advancements in the processing of serial data due to the limitations on the robust processing window. As a result of parallel computing, transformers can successfully apply global attention over the entire collection of images while considering the relationships between remote sequence parts. We apply knowledge from the deep CNN model learned on a generic classification task for images to handle high-level picture representation while training with small datasets. We address how our methodology may be applied, such as in a plugin-in module with pre-existing image-recognition frameworks, with the settings remaining constant and only the transformer component being trained. Automatic photograph organization in typical real-world applications requires both single photograph categorization and abnormal event classification. In practice, it may be beneficial to use pre-calculated picture representations or share image representations across these two jobs. Our proposed method permits sharing of parameters and computations as well as knowledge transfer from large datasets, which shares similar advantages with multi-task learning. Understanding which photographs contain important information and separating those from less important photographs is a crucial component of automatic album arrangement. This semantic content analysis work can improve event recognition, and it may also be helpful for other ranking applications for photo collections, such as album stories or summarization, or the removal of clutter from photos. The subjective nature of image importance prediction presents a significant challenge; hence, only annotated datasets are currently available. Previous methods trained models to predict image importance explicitly from labeled data [3,7]. Human-annotated datasets are limited because the relevance of an image is not clearly defined, and they may be difficult to scale due to the laborious process of annotating the ground truth for both event name and importance score. In our approach, we propose using the transformer's attentional learnings to forecast the significance of album images. We show that our model can estimate image importance even when it has not been specifically trained on this task-specific annotation. This research makes several innovative contributions.

1) We propose a multitask model that trains both album-level representation for event classification and image-level feature representation for image importance tasks.

2) We investigate a novel fusion of transformers and attention mechanisms for photo album understanding to recognize events and score important images. This methodology significantly outperforms state-of-the-art (SOTA) approaches on image importance tasks and achieves high performance on event recognition tasks.

3) Evaluation and testing on multiple backbones and several samples on each album, leveraging a pre-trained feature extractor for image representation. The results on the ML-CUFED dataset outperform previous approaches by achieving 3% higher accuracy on image importance tasks.

RELATED WORK

In general, the term "event recognition" can apply to a variety of media types, including single photos (such as those shared on social media), individual photo albums, videos, and audio. The task of per-image event recognition has been the subject of numerous publications [12–17]. The issue of picture representation is addressed via per-image event recognition; however, multi-photo analysis or handling of irrelevant photos have not been addressed. Public datasets [12,16,18], as well as more current datasets [19], are created by collecting unrelated photos from various sources. Each image in the datasets is named by the event name individually. Three datasets for personal albums that are publicly accessible, Holidays [20], PEC [21], and ML-CUFED [3,7], were all collected from individual photo albums. They offer a more accurate representation of personal user albums by including both relevant and irrelevant photographs in each album. While some earlier studies evaluated image and album event detection tasks, they are typically viewed as separate. As personal photograph applications are becoming more widespread, many pertinent insights are provided by event categorization at the image-level based on image features, although the focus of this study is event recognition of personal photo collections. The difficulty of representing an album and a single image in a personal photo album was previously combined to examine different methods for event detection. Guo [22] proposes the use of an ensemble of networks to handle image representation and minimal dataset size. A coarse network is pre-trained on the massive places-365 scene classification dataset [23]; subsequently, a fine network is pre-trained on the large object-classification dataset ImageNet [24]. Each network is pre-trained using a different large dataset. The author shows that, to identify an event type, both high-level and low-level features are needed. These ideas are expanded upon in a subsequent study [8], which increases the ensemble's size by including a new network and attention layers. For each of the three convolution layers, the features of the album photographs are aggregated using a weighted average. By comparing the event label representation and the picture feature representation using word2vec [25], an auxiliary loss is used to determine the weights. In order to demonstrate the improvement, kids explicitly learn the value of images. When one produces an image caption using image embeddings, Savchenko [9] also utilizes an ensemble of two networks. In the past, aggregation of album photos was performed by utilizing averaging [22], attention-weighted averaging [9], RNNs, LSTMs [3–5], and GRUs [26]. In reality, these RNNs have a small temporal window. Transformer architecture offered a breakthrough mechanism in NLP [10] by giving the entire input sequence global attention. Applying transformers to photo collections is appropriate given that they have recently been employed for videos [27]. Other related studies address the task of predicting the relevance of images in addition to the event recognition task. A Siamese network [3] explores the prediction of importance scores for images given the sort of event type in general. A more recent study [4] extended that work, offering an iterative approach that alternated predicting the importance of the picture and the event in the album; another study [5] further addressed the prediction of the importance score for all event types. The contribution to event recognition using picture significance prediction was also shown [3,5]. These methods all employ the CUFED dataset's annotated image importance to supervise the training of the prediction network. The substantial level of effort required to annotate such data for this specific activity is well described in another study [4]. A major contribution of that work is automatic expansion of the annotations. Estimating image relevance by using an unsupervised approach may be beneficial and even applicable to other subjective content-based tasks. We demonstrate how an image's relevance can be predicted without specifically training on its annotation.

PROPOSED METHOD

If we were given a photo album with N_A images and a time of 1, we would forecast the event type of the album as being $1 \ldots N_{cls}$. The three basic obstacles to album event recognition are as follows. First, albums are disorganized N_A groups of photographs of varying lengths. Second, albums frequently contain photographs that are only somewhat relevant. Third, while event detection datasets for albums are very modest, high-level image representation is needed. Every real-world system is constrained by computation efficiency and parameter requirements, which presents another obstacle that should be considered. Within the context of these obstacles, our method strives to achieve high recognition rates. Specifically, we seek to expand the application of transformer architecture to albums in response to the recent reported success of transformers in various domains, including video[27,28], NLP[10], and vision[12] (image collections). Several changes must be implemented before transformers can be applied to an image collection. Since the length of an album can vary, we only select one image from each album while sampling it. The next step is to apply a feature encoder that has been pre-trained on a sizable image dataset in a classification task to construct a feature descriptor for each sampling image. Transformers are then used to aggregate these embeddings, producing the prediction layer for the event classification. Additionally, a forecast of visual relevance is taken from the attention block. Transformation training implicitly teaches image importance prediction. The subsequent section describes the many steps in our process. The general architecture of the suggested solution is illustrated in Figure 2.

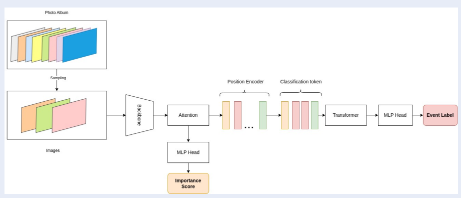

Figure 2: Proposed multitask architecture for even trecognition and importance image scoring in a photo album

Album Sampling

As previously stated, our goal is a viable, effective solution. Since it is well known that transformers rely quadratically on the length of serial input, we simplify this problem by sampling just a few photos from each album. Fewer images are needed to achieve high identification rates since the attention mechanism enhances the more pertinent ones. The sampling of images improves efficiency and addresses the issue of varying album length. We define the fixed length of the input sequence for both training and inference phases as the sampled images per album S_A. Then, we randomly select S_A photographs from an album of N_A images $I_{t_{t=1}}^{N_a}$. Each batch B is sampled by selecting S_A pictures from B_A albums: $B = B_A S_A \quad (1)$

In the case of $N_A < S_A$ in album photos, we duplicate $S_A - N_A$ photos and add them to S_A photos for matching with required input data.

Album Representation

We use an image event recognition dataset that may not be sufficiently large to train an adequate image encoder backbone for the sampled images with s=1. As a result, we extract the sampled album picture embeddings using an image classification pre-trained backbone. This method enables the backbone network to be used for both knowledge transfer and parameter exchange for image classification into album representation based on visual features at the image level. Additionally, this approach permits offline image embedding aggregation for pre-computed image embeddings using just the transformer network. The MTResNet is trained on Open Images V6[29] as an image feature encoder and is shown to transfer information in experimental results. This is a very large and diverse classification dataset that removes the need to fuse models trained independently for both low-level and high-level content. MTResNet[29] is an effective, powerful backbone. The following section demonstrates the effects of pre-trained dataset selection.

Attention Module

In our proposed method, we adapted the attention module, which aids in the goal of increasing attention to the key contexts and participants in the target event to better extract attributes for the issue. After extensive testing, we discovered that ParNet Attention[30] outperformed the non-attention model in terms of output. Non-deep network architecture is utilized by the ParNet block. The VGG block is frequently used as the building block for this network architecture, but it has the drawback of being more challenging to train than ResNet. According to a recent study, the "Structural reparameterization" technique can be trained easily. Researchers employ Rep-initialization VGG's block[31], after which change is required to meet the shallow architecture. One issue with deep networks is that there are relatively few options for feature extraction in large feature fields when utilizing simply conv3x3 (Receptive field). The Skip-Squeeze-Excitation (SSE) class builder is based on Squeeze-Excitation (SE) to address this issue. The ParNet Attention's architecture is visualized in Figure 3. We apply the multitasking learning approach to simultaneously address the challenges of event recognition and critical photo prediction in event photo albums after completing the attention module.

Event Recognition using Transformers

The architecture of the transformers utilized for aggregating the embedding vectors is modeled after the STAM for action classification in video[27]. This approach treats the sampled images as tokens and incorporates them into the transformer. A learnable encoding for positional information is used for each picture embedding in x_s:

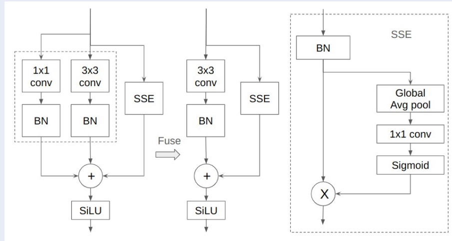

Figure 3: Architecture of a ParNet Block[30] that includes 1×1 convolution, 3×3 convolution, and SSE. These blocks are computed in parallel for faster inference and SSE for more receptive fields that do not require depth.

$z_s^{(0)} = x_s + z_{pos}^s,\ s \in 1,\ 2, \ldots,\ S_A$ (2)

The picture embedding dimension (in this example, $D = 2048$) determines the size of the positional encoding. This is performed for every photo, creating an album's $S_A \times D$ matrix of image embeddings. The album-level descriptor is expanded to $(S_A + 1) \times D$ by connecting a classification token CLS to the album representation to apply classification using transformers. The dimension serves as the input picture token, and the CLS is a trainable weight that relies on pertinent information[32]. The stacked L transformer encoder layers are placed in the $(S_A + 1) \times D$ album descriptor. The first encoder layer input is a layer with sufficient heads to directly simulate $\left(CLS,\ z_s^{(0)}{}_{s=1}^{S_A}\right)$. Each encoder is made up of an MSA[33] with H heads since it may be used for both short- and long-distance interactions. The learnable $W_K^{(l,h)}$, $W_Q^{(l,h)}$, $W_V^{(l,h)}$ matrices, into Query (Q), Key (K), and Value (V) for each input, and formalized by LN[34] and linearly projected by each head $h \in 1, 2, \ldots, H$ embeddings that are not the learnable $W_K^{(l,h)}$, $W_Q^{(l,h)}$, $W_V^{(l,h)}$.

$Q_s^{(l,h)} = W_Q^{(l,h)} LN\left(z_s^{(l1)}\right)$ (3)

$K_s^{(l,h)} = W_K^{(l,h)} LN\left(z_s^{(l1)}\right)$ (4)

$V_s^{(l,h)} = W_V^{(l,h)} LN\left(z_s^{(l1)}\right)$ (5)

The weighted values are calculated using a dot-product operation and scaling by head dimensions $D_H = D/H$ according to the following formula:

$z(Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{D_H}}\right) \times V$ (6)

Each layer's output from the attention heads is combined, then sent through two multi-layer perceptron (MLP)[35] layers activated by GeLU[36]. Due to the new skip connections, LayerNorm[34] is applied prior to the MLP. With the designed architecture, the MLP and MSA layers function like residual layers. A multi-label prediction and an attention vector connecting each picture token to the CLS are the network's outputs; this result is examined in more detail in Section 4. An attention-based aggregation for albums, which exploits the more relevant photographs and decreases the contribution of the images that are less relevant, is the primary advantage of the transformer architecture. The different picture weights can be considered the weighted value extracted from the multi-head attention. With small sampled photos, effective recognition is possible when pertinent images are highlighted by the attention mechanism. In our implementation, we choose $H = 8$ heads for the MSA module and $H = 6$ transformer encoder layers.

Image Importance Prediction

To predict the image importance score, we use an MLP[35] consisting of two FC hidden layers and one dropout layer. The output, which is a score between 0 and 1, predicts the significance score of the image per album.

EXPERIMENT

Our approach consists of a number of modules with the goal of combining the most effective modules for each component of the architecture. We contrast various options and analyze their effects separately to support our module choices. The ML-CUFED dataset[3] was used for the majority of our ablation study trials. Additionally, it is the sole multi-label event in the dataset, which presents a difficult and pragmatic issue representative of real applications.

ML-CUFED Dataset

ML-CUFED[3] is the expanded version of the CUration of Flickr Events Dataset (CUFED)[7] for event recognition dataset to multi-label events. The dataset consists of a total of 94,798 photographs in 1,883 albums, with between 30 and 100 images per album. The dataset contains 23 different event classifications, including sports, holidays, and personal occasions such as birthdays. The dataset has an "image significance" term with score and event name in the ground truth for each image collection. This annotation was completed by asking a number of commentators about the pertinence of each photo to a specific event and then averaging their scores. The average Spearman correlation among the annotators is 0.4, calculated using multiple 2-group splits, demonstrating the subjective nature of this work. Following earlier studies[3,7], we split the dataset into 4:1 for training and validation for both training and testing in the experiment methodology.

Figure 4: Sample images of architecture events from the ML-CUFED dataset

Table 1: Recognition rates of events in different samples in the album

Number of samples	mAP (%)
8	59.88
16	65.39
24	65.62
32	**67.21**

Table 2: The recognition rates of event recognition methods in the ML-CUFED dataset

Method	Accuracy (%)
Multitask_TResnet	**67.21**
Multitask_DenseNet 121	60.06
Multitask_ResNet50	65.87

Event Recognition Results

We evaluated our proposed approach for an event recognition task on the ML-CUFED public dataset. We show how the quantity of photographs per album (SA) affects the results of recognition. The transformer's strength is due in part to its capacity to adapt to weigh the various images and draw greater attention to the more important ones. As a result, the model can classify events accurately even with fewer photos. Table 1 demonstrates that the results surpass those of earlier methods that utilized all album photographs, even when utilizing either SA = 8 or SA = 16. The comparison in this aspect is not ideal because our designed network employs fewer photographs even for SA = 32 in practice, as we consider this number to be approximately 30–50% of the number of images in many albums. Although the design of the transformer requires a specific number of photographs, it can still identify the event type using only a small subset of the album's images due to its strong aggregation capabilities. For this dataset, we achieved high performance, as shown in $; these results are evaluated with the mAP metric. Our result for ML-CUFED achieved a mAP of 67.21% for the TResnet backbone, which is the best performance for our architecture.

Importance Image Results

We adopted two evaluation metrics to assess the different proposed methods. For a given album, we only obtain the top t% results as relevant images and evaluate these metrics at different values of t. First, we evaluate our models using mean average precision (MAP). Information retrieval assessment methods are frequently used. The precision-recall curve is the averaged area under the overall curve of albums. The MAP@t% can be calculated using the album collection and the top t% of the photographs as the relevant images:

$$AP(S)\,@t\% = \int_0^1 p(r)\,d(r) \simeq \frac{1}{N}\sum_{k=1}^{n} \frac{p(k)\times rel(k)}{n\times t\%} \quad (7)$$

$$MAP(U)\,@t\% = \frac{1}{N}\sum_{k=1}^{N} AP(S_i)\,@t\% \quad (8)$$

where S_i is the ith album and U is the album collection. The factors *n, p(k),* and *rel(k)* determine how big the album S is and whether the kth-ranked photo from our system is a relevant picture or one of the top *t%* results in the ground truth. Second, we determine the precision (P), which is the proportion of relevant images in the photos that were recovered to all relevant images at each *t*. The difference from *MAP, P* focuses on how many significant photos are obtained at a cutoff value of *t*, and is not concerned with the locations of the other essential images in the ranking system or in the retrieval list. *P* is also an intuitive approach to validate the efficacy of our projected image ranking result, although it is less informative than *MAP*. We only show findings for $t \leq 30$ because we are addressing importance scoring and image selection, and are primarily interested in both *MAP* and *P* metrics when *t%* is small.

As can be seen in Table 3 and Table 4 , the TResnet backbone-based model outperforms the Resnet backbone when used with the dataset in terms of MAP@t and P@t results. This demonstrates that TResnet retrieves features more effectively than Resnet. It is based on an extension from ResNet50 that is refined on an architecture with SpaceToDepth Stem, Anti-Alias Downsampling, In-Place Activated BatchNorm,

Table 3: Comparison of predictions with MAP@t metrics on the ML-CUFED dataset

Method	MAP@t(%)					
	5	**10**	**15**	**20**	**25**	**30**
SiameseNet [3]	0.305	0.364	0.417	0.471	0.519	0.563
CNN-LSTM-Iterative [4]	0.302	0.371	0.419	0.47	0.52	0.568
Resnet50_L2_loss	0.315	0.479	0.533	0.577	0.622	0.653
TResnet_m L2_loss	0.338	0.5	0.555	0.606	0.641	0.671
Multitask_ResNet50	0.304	0.4775	0.542	0.625	0.584	0.657
Tresnet_m_Multask	0.33	0.501	0.568	0.606	0.643	0.67

Table 4: Comparison of predictions with P@t metrics on the ML-CUFED dataset

Method	P@t(%)					
	5	**10**	**15**	**20**	**25**	**30**
SiameseNet [3]	0.216	0.301	0.36	0.411	0.459	0.504
CNN-LSTM-Iterative [4]	0.205	0.3	0.36	0.413	0.459	0.507
Resnet50_L2_loss	0.193	0.275	0.335	0.397	0.448	0.497
TResnet_m L2_loss	**0.207**	**0.289**	**0.36**	**0.416**	**0.466**	**0.51**
Multitask_ResNet50	0.198	0.2799	0.339	0.397	0.451	0.49
Tresnet_m_Multask	**0.209**	**0.293**	**0.352**	**0.415**	**0.461**	**0.51**

Novel Block-type Selection, and Optimized SE Layers to optimize for GPU resources and feature extraction in small kernels. We then selected TResnet as the main algorithm for the crucial picture prediction challenge in the event photo album. The scoring method and the learning method, which integrate album score prediction and album event recognition into one, will be compared next as two approaches to the significant photo album prediction problem. The results of the single model show that the combined model performs better and is stronger.

CONCLUSIONS

Recognizing events in photo albums is a semantic challenge that requires both low-level image analysis and high-level content interpretation, as well as the selection and aggregation of vital images relevant to specific events within the album when recognizing events. In this work, we propose a multitasking architecture based on CNN backbone with an attention and transformer mechanism for understanding photo albums. We tested numerous samples in the album and many different backbones to demonstrate the generalization in our architecture. In addition to significantly outperforming SOTA identification results, our method can recognize event labels and detect important images in an album. On the ML-CUFED dataset, we demonstrated how a transformer's design may handle a variable-length image collection that contains irrelevant images. The results show that our proposed multitask method with the TResnet backbone achieves the highest performance for both event recognition and importance scoring tasks. Additionally, we believe this strategy can be applied to real-world personal photo applications and leveraged subjectively. A limitation of our proposed method is that it does not consider the role of objects and the relationship between them in event classification, as well as image scoring. In a future study, we will investigate object detection and image structure parsing at the semantic level in recognizing events and scoring the importance of images in photo albums. More specifically, we will consider a graph-based approach to build object relationships and semantics in image contexts in a specific event album.

ABBREVIATIONS

CNN: Convolution Neural Network
CLS: classify token
FC: Full Connected
GeLU: Gaussian Error Linear Unit

LSTM: Long Short Term Memory
ML-CUFED: Cuation of Flickr Events Dataset
MLP: Multi-Layer Perceptron
MSA: Multi-head Self-Attention
mAP: Mean Average Precision
NLP: Natural Language Processing
P: Precision
RNN: Recurrent neural networks
SOTA: state-of-the-art
SE: Squeeze-Excitation
SSE: Skip-Squeeze-and-Excitation
STAM: SpatioTemporal Attention based Memory
VGG: Visual Geometry Group

COMPETING INTERESTS

The authors hereby declare there are no conflicts of interest associated with this work.

AUTHORS' CONTRIBUTIONS

Vo Hoai Viet: Conceptualization, Supervision, Methodology, Data Analysis, Writing – review & editing. Le Quoc Viet: Investigation, Data curation, Methodology, Software, Writing – original draft.

ACKNOWLEDGEMENTS

The authors would like to thank Zetgoo and the Computer Vision department, University of Science, VNU-HCM for supporting this work.

REFERENCES

1. Guo C, Tian X. Event recognition in personal photo collections using hierarchical model and multiple features. In: and others, editor. 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP); 2015. p. 1–6.
2. Savchenko AV. Event recognition with automatic album detection based on sequential processing, neural attention and image captioning; 2019.
3. Wang Y, Lin Z, Shen X, Mech R, Miller G, Cottrell GW. Recognizing and curating photo albums via event-specific image importance; 2017.
4. Qi F, Yang X, Zhang T, Xu C. Discriminative multimodal embedding for event classification. vol. Volume 395. Neurocomputing; 2020.
5. Choi S, Park H, Hwang S. Meta-supervision for attention using counterfactual estimation. Data Sci Eng. 2020;5(2):193–204.
6. Guo X, Polania LF, Zhu B, Boncelet C. Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases; 2019.
7. Wang Y, Lin Z, Shen X, Mech R, Miller G, Cottrell GW. Event-specific image importance. In: 2016 {IEEE} {C}onference on {C}omputer {V}ision and {P}attern {R}ecognition ({CVPR}); 2016. p. 4810–4819.
8. Guo C, Tian X. Multigranular event recognition of personal photo albums. IEEE Trans Multimed. 2018;20(7):1837–47.
9. Savchenko A. Event recognition with automatic album detection based on sequential grouping of confidence scores and neural attention. In: 2020 {I}nternational {J}oint {C}onference on {N}eural {N}etworks ({IJCNN}); 2020. p. 1–8.
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention is all you need; 2017.
11. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–1780.
12. Li LJ, Fei-Fei L. What, where and who? classifying events by scene and object recognition. IEEE 11th International Conference on Computer Vision. 2007;p. 1–8.
13. Park S, Kwak N. Cultural event recognition by subregion classification with convolutional neural network. In: 2015 {IEEE} {C}onference on {C}omputer {V}ision and {P}attern {R}ecognition {W}orkshops ({CVPRW}); 2015. p. 45–50.
14. Wang L, Wang Z, Qiao Y, Van Gool L. Transferring object-scene convolutional neural networks for event recognition in still images; 2016.
15. Ahmad K, Mekhalfi ML, Conci N, Melgani F, Natale FD. Ensemble of deep models for event recognition,"ACM Trans. Multimedia Comput. Commun. Appl., vol. 14, no. 2. ACM Trans Multimed Comput Commun Appl. 2018;14(2):1–20.
16. Ahsan U, Sun C, Hays J, Essa I. Complex event recognition from images with few training examples; 2017.
17. Laib L, Allili MS, Ait-Aoudia S. A probabilistic topic model for event-based image classification and multi-label annotation. Image Commun. 2019;76(C):283–294.
18. Xiong Y, Zhu K, Lin D, Tang X. Recognize complex events from static images by fusing deep channels. In: 2015 {IEEE} {C}onference on {C}omputer {V}ision and {P}attern {R}ecognition ({CVPR}); 2015. p. 1600–1609.
19. Muller-Budack E, Springstein M, Hakimov S, Mrutzek K. Ontology-driven event type classification in images; 2020.
20. Tsai SF, Huang TS, Tang F. Album-based object-centric event recognition. In: 2011 {IEEE} {I}nternational {C}onference on {M}ultimedia and {E}xpo; 2011. p. 1–6.
21. Bossard L, Guillaumin M, Van L. Event recognition in photo collections with a stopwatch hmm. In: 2013 {IEEE} {I}nternational {C}onference on {C}omputer {V}ision; 2013. p. 1193–1200.
22. Guo C, Tian X. Event recognition in personal photo collections using hierarchical model and multiple features. In: and others, editor. IEEE 17th International Workshop on Multimedia Signal Processing (MMSP); 2015. p. 1–6.
23. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell. 2018;40(6):1452–64.
24. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 {IEEE} {C}onference on {C}omputer {V}ision and {P}attern {R}ecognition; 2009. p. 248–255.
25. v u˚vrek RR, Sojka P. Software framework for topic modeling with large corpora; 2010. p. 45–50.
26. Chung J, Gulcehre C, Cho K. Empirical evaluation of gated recurrent neural networks on sequence modeling; 2014.
27. Sharir G, Noy A, Zelnik-Manor L. An image is worth 16x16 words, what is a video worth?; 2021.
28. Arnab A, Dehghani M, Heigold G, Sun C. Vivit: A video vision transformer; 2021.
29. Ridnik T, Lawen H, Noy A, Baruch EB, Sharir G, Friedman I. Tresnet: High performance gpu-dedicated architecture; 2020.
30. Goyal A, Bochkovskiy A, Deng J, Koltun V. Non-deep networks; 2021.
31. Ding X, Zhang X, Ma N, Han J, Ding G, Sun J. Repvgg: Making vgg-style convnets great again. In: {P}roceedings of the {IEEE}/{CVF} {C}onference on {C}omputer {V}ision and {P}attern {R}ecognition; 2021. p. 13.
32. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding; 2018.
33. Cordonnier JB, Loukas A, Jaggi M. Multi-head attention: Collaborate instead of concatenate; 2020.
34. Ba JL, Kiros JR, Hinton GE. Layer normalization; 2016.
35. Singh G, Sachan M. Multi-layer perceptron (mlp) neural network technique for offline handwritten gurmukhi character recognition. In: 2014 {IEEE} {I}nternational {C}onference on {C}omputational {I}ntelligence and {C}omputing {R}esearch; 2014. p. 1–5.
36. Hendrycks D, Gimpel K. Gaussian error linear units (gelus); 2016.