Masked Feature Modelling for the unsupervised pre-training of a Graph Attention Network block for bottom-up video event recognition

Dimitrios Daskalakis, Nikolaos Gkalelis, Vasileios Mezaris
CERTH-ITI
Thessaloniki, Greece, 57001
{dimidask, gkalelis, bmezaris}@iti.gr

***Abstract*—In this paper, we introduce Masked Feature Modelling (MFM), a novel approach for the unsupervised pre-training of a Graph Attention Network (GAT) block. MFM utilizes a pretrained Visual Tokenizer to reconstruct masked features of objects within a video, leveraging the MiniKinetics dataset. We then incorporate the pre-trained GAT block into a state-of-the-art bottom-up supervised video-event recognition architecture, ViGAT, to improve the model's starting point and overall accuracy. Experimental evaluations on the YLI-MED dataset demonstrate the effectiveness of MFM in improving event recognition performance.**

***Index Terms*—masked image modelling, masked feature modelling, graph, attention, event recognition**

I. Introduction

In recent years, Vision Transformers (ViTs) have emerged as a dominant approach in video and image analysis, gradually surpassing Convolutional Neural Networks in various applications. However, one key challenge faced by ViTs is their requirement for abundant data and extensive annotations to achieve optimal training results. To tackle this annotation-hungry nature of ViTs, several techniques were studied, such as Transfer Learning and, more recently, Masked Image Modelling (MIM) and other self-supervised pre-training techniques [2], [3]. The primary objective of MIM is to learn to reconstruct the masked patches of an image in order to capture comprehensive contextual information using a representation model, as e.g. in the case of BEiT [4], where the pretraining process involves utilizing two different views of each image: image patches and visual tokens. Although studied in the image domain, masking has not yet been thoroughly explored in video event recognition or video-related tasks in general.

We present a new approach called Masked Feature Modelling (MFM), particularly tailored to videos (Fig.1). In summary, our major contributions are:

- We are the first, to the best of our knowledge, to apply vector-quantized visual tokenizer MFM techniques [5] to extracted object features within videos for unsupervised pretraining of Graph Attention Networks.
- We show that the Graph Attention Networks pretrained through the proposed MFM technique can provide improvements in event recognition accuracy in video.

This work was supported by the EU Horizon 2020 programme under grant agreement 101021866 (CRiTERIA). The work of N. Gkalelis was done while he was with CERTH-ITI. Source code is available at https://github.com/bmezaris/masked-ViGAT.

II. Related Work

This brief survey of related work touches upon the three domains that are most closely related to this work: a) Transfer Learning and Unsupervised Pre-Training, b) Masked Image Modelling and c) Video Event Recognition.

Transfer Learning (TL) [6] leverages learned feature maps from models trained on large datasets, offering benefits in downstream tasks (reduced training time, improved performance). One possible TL approach is Weight Initialization; this has shown promising results in various applications, e.g. [7]. Unsupervised pretraining, as in e.g. [8], allows learning these initial weights of the deep network by training on large volumes of unlabeled data.

Under the unsupervised pretraining paradigm, Masked Image Modeling (MIM) is a technique that leverages the reconstruction of masked image content to learn new representations. Recently, this approach has been applied in training ViTs, as in BEiT [4], which introduces the prediction of discrete visual tokens as a key element. In the video domain, MIM has been used in [9], [10], [11]. However, none of these previous studies explored *feature masking* in images or videos.

In Video Event Recognition, two dominant directions have emerged: i) training a Transformer or similar network from scratch using images or video frames, as in e.g. [12], and ii) in order to reduce computational cost, utilize pretrained models to extract feature representations, which are then fed into the respective classifier models, as in e.g. [13]. Two notable approaches of the latter category are ViGAT [1], depicted in Fig.2, and its faster approximation Gated-ViGAT [14], which incorporate an object detector to extract bottom-up (object) information from video frames. The extracted objects and frames are processed by a pretrained ViT backbone [15] to derive feature representations; and, these are fed to a trainable attention-based head network made of three Graph Attention Network (GAT) blocks, ω_1, ω_2 and ω_3, to recognize and explain events within the video.

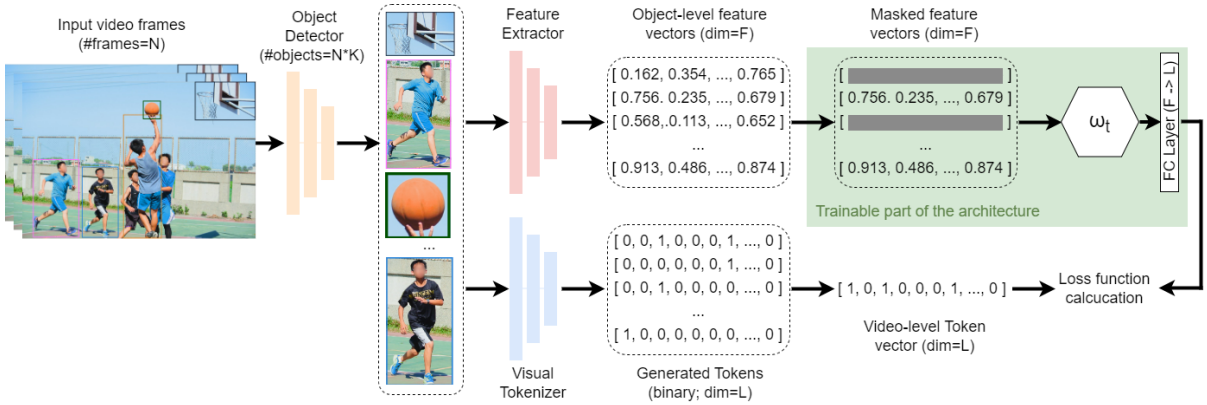

Fig. 1. An illustration of the main contribution of our paper, training a Graph Attention (GAT) block, ω_t [1] through Masked Feature Modelling and Tokens. Initially, we detect the objects within a video frame, pass them through the feature extractor to acquire object-level features. In parallel, a Tokenizer creates object-level token vectors of the objects based on their cosine similarity to a predefined Visual Vocabulary, to subsequently generate a video-level token vector, which is used as supervision in our unsupervised architecture. In each training iteration, a portion (e.g. 40%) of the extracted object-level features are masked and the modified set of features is fed through ω_t and a fully connected (FC) layer. The binary cross-entropy loss is used for the unsupervised training, with the help of the generated token, of the trainable components of this architecture: ω_t and FC.

In this paper, our goal is to (pre-)train the GAT blocks of ViGAT (Fig.2) without supervision. Building upon the weight-sharing technique described in [1], we accomplish this by training a new GAT block ω_t using local (i.e. object) information and without supervision; and, subsequently, utilizing the pretrained ω_t to initialize one or more GAT blocks of ViGAT. For unsupervised training on a large source-dataset, we apply masked feature modelling using the Vision Tokenizer introduced in [4]. Once this unsupervised training is completed, we initialize the selected GAT blocks of ViGAT with our ω_t and conduct supervised training and evaluation in a video-event recognition task on a small target-dataset. We should note that the object-centric transformer presented in [16] also utilizes object features and a masked unsupervised pretraining step. However, in contrast to [16] that utilizes a self-supervised contrastive loss (InfoNCE [17]) at scene level, here we use a pretrained vector-quantized visual tokenizer with a cross-entropy loss [5]. The latter approach puts more emphasis on local- (in our case object-) level reconstruction. This design choice is motivated by recent results in the image classification domain, which showed that BEiT v2 (utilizing a vector-quantized visual tokenizer) clearly outperforms MoCo v3 that is based on the InfoNCE loss [18] (see Table 2 in [5]).

III. Proposed Method

Our method comprises an object detector, a feature extractor, a Visual Tokenizer, and a GAT block referred to as ω_t. ω_t adopts the structure defined in [1], i.e. it is made of an attention mechanism, a GAT head of two layers, and a graph pooling stage; the interested reader is referred to [1] for details on the GAT block structure. In the core of our method is the training of ω_t using tokens generated from a Visual Tokenizer [5], in conjunction with a novel unsupervised approach based on MIM that utilizes features instead of images, called Masked Feature Modelling (MFM).

To acquire object-level features, we utilize the method employed in [1]. That is, a video is represented with a sequence of N frames, and an object detector along with a feature extractor is used to obtain matrix $\mathbf{X}^{(n)} \in \mathbb{R}^{K \times F}$ representing frame n,

$$\mathbf{X}^{(n)} = [\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_K^{(n)}]^T, \tag{1}$$

where K is the number of objects extracted from each frame, $\mathbf{x}_k^{(n)} \in \mathbb{R}^F$ is the feature embedding for object k in frame n.

In order to achieve unsupervised learning, we utilize the pretrained Visual Tokenizer [5]. As seen in Fig.1, by feeding an object image into the Tokenizer, we obtain new representations, called tokens, that serve as valuable supervision for our procedure. The Tokenizer consists of a vision Transformer encoder, a quantizer and a visual vocabulary (Codebook) containing L distinct embeddings. Inside the Tokenizer, the object images are partitioned to Q patches, and each patch is transformed to an embedding $\mathbf{h}_{j,k}^{(n)}$, which corresponds to the jth patch of the kth object of frame n. Then the quantizer looks up the nearest neighbor in the visual vocabulary for each representation $\mathbf{h}_{j,k}^{(n)}$, according to cosine similarity, and thereby produces the respective visual token vector $\mathbf{z}_{j,k}^{(n)} = [z_{1,j,k}^{(n)}, \ldots, z_{L,j,k}^{(n)}]^T$, where, $z_{i,j,k}^{(n)} \in \{0, 1\}$, and $z_{i,j,k}^{(n)}$ equals 1 if the jth patch of the kth object in the nth frame belongs to the ith codebook embedding, and 0 otherwise. A visual token vector $\mathbf{v}$ for the overall video is then obtained using

$$\mathbf{u} = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{Q} \mathbf{z}_{j,k}^{(n)}, \tag{2}$$

and the function top_r that returns 1 for the r largest elements of $\mathbf{u}$ and 0 for the rest,

$$\mathbf{v} = \mathsf{top}_r(\mathbf{u}). \tag{3}$$

To proceed with the unsupervised training, given the feature matrix $\mathbf{X}^{(n)}$, we mask $\Gamma\%$ of the objects in each frame

$$\mathbf{x}_k^{(n)} = \delta(k \in \mathcal{M})\mathbf{p} + (1 - \delta(k \in \mathcal{M}))\mathbf{x}_k^{(n)}, \tag{4}$$

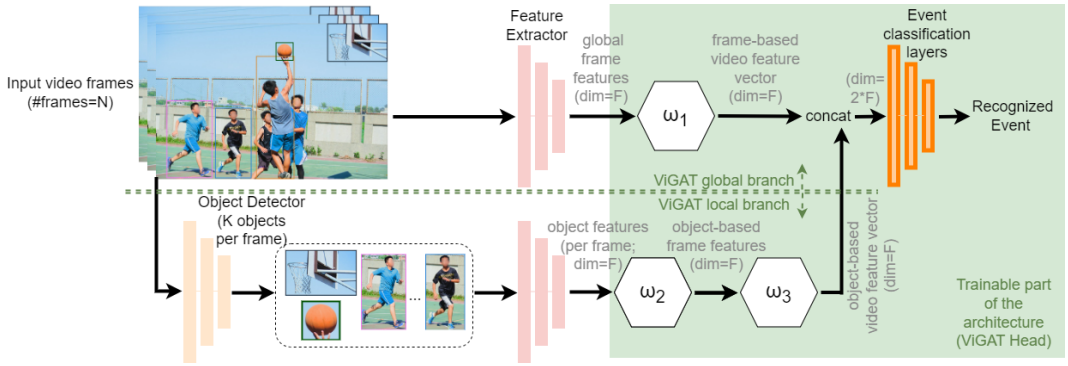

Fig. 2. The block diagram of ViGAT [1]. We adopt this method for the supervised video event recognition task. It encompasses an object detector, a feature extractor, and the ViGAT head. The ViGAT head consists of three Graph Attention Network (GAT) blocks (ω_1, ω_2, ω_3) responsible for processing global (frame-level) and local (object-level) features. Finally, a pair of event classification layers utilize the video-level feature vectors coming out of GAT blocks ω_1 and ω_3 to recognize the event occurring in the video.

where, $\delta()$ is the indicator function, $\mathbf{p} \in \mathbb{R}^F$ is a shared learnable object feature embedding and $\mathcal{M}$ is the set of object indices k, randomly selected to be masked. The resulting, so-called masked feature matrix $\tilde{\mathbf{X}}^{(n)}$, is processed by ω_t to produce a latent representation in its output; and then, a FC layer of F inputs and L outputs, equipped with an appropriate nonlinearity (e.g. sigmoid), is utilized to transform the output of ω_t to a score vector $\mathbf{g} \in \mathbb{R}^L$, containing L score values with respect to the codebook vocabulary for the overall video. The standard cross-entropy loss is then used to compute the dissimilarity between $\mathbf{g}$ and the visual token vector $\mathbf{v}$.

In this way, using a large unlabelled video dataset, the GAT block ω_t can be trained effectively in an entirely unsupervised manner. Subsequently, it can be used for initializing and re-training a supervised event-recognition architecture (ViGAT) on a smaller target-dataset, to effect knowledge transfer.

In the supervised event-recognition task, we utilize the pretrained ω_t to initialize one or both of ω_2 and ω_3 depicted in Fig.2. The entire ViGAT architecture is then trained on the YLI-MED dataset.

IV. Experiments

A. Datasets and experimental setup

We use two established, publicly available video datasets:

i) MiniKinetics [19] is a subset of Kinetics [20]. It comprises 200 action classes, 121215 training and 9867 testing video clips. Each clip, sampled from a distinct YouTube video, has a duration of 10 seconds and is annotated with a single event/action class label. We utilize this as the source dataset, i.e. for training ω_t in an unsupervised way without using the existing class-label annotations, as illustrated in Fig.1.

ii) YLI-MED [21] is a video corpus based on YFCC100M, containing 1823 videos and 10 event categories. The dataset is divided into standard training and testing partitions of 1000 and 823 videos, respectively. We employ this much smaller dataset as the target one, i.e. for supervised learning, taking advantage of the previously trained ω_t as illustrated in Fig.2.

In order to accurately represent each video within the two datasets, we initially employ uniform sampling, resulting in a sequence of $N = 9$ or $N = 25$ frames (depending on the experiment) for YLI-MED and $N = 30$ frames for MiniKinetics. Our approach consists of the following:

i) An object detector named Detic [22], which is pretrained on ImageNet21K and fine-tuned on the CoCo dataset.

ii) A ViT-L/14-Clip backbone, utilizing the OpenAI CLIP model [23] for the extraction of object- and frame-level features. This backbone utilizes a 14×14 grid to patchify the input image, i.e., $Q = 14^2$ patches are produced per frame; the derived feature embeddings' dimension is $F = 1024$.

iii) The pretrained Visual Tokenizer provided in [5] with Codebook size $L = 8192$.

For object detection, we set the number of objects K to be extracted as 50. For the masking procedure we set $\Gamma = 40\%$. Our unsupervised architecture for pretraining ω_t (Fig.1) was trained on MiniKinetics for 200 epochs (with learning rate initially set to 10^{-3} and subsequently multiplied by 0.1 at epochs 50 and 100).

Then, our supervised event recognition architecture (Fig.2) was trained on the YLI-MED dataset for 200 epochs (with learning rate initially set to 10^{-4} and subsequently multiplied by 0.1 at epochs 60 and 110). The same training was performed on either the local branch alone or the entire architecture, depending on the experiment.

In alignment with the relevant literature, top-1 accuracy is used as evaluation metric on YLI-MED.

B. Event recognition results

We evaluate the performance of our unsupervised pretrained GAT block ω_t, by using it for the initialization of GAT blocks in our supervised event-recognition architecture ViGAT, on the YLI-MED dataset. The upper part of Table I shows our experiments using just the local branch of the entire ViGAT architecture, as depicted in Fig.2. To study the behavior of ω_t and minimize the influence of frame-level features, we used a small number of frames in this experiment, specifically $N = 9$. The results demonstrate that our pretrained ω_t outperforms the randomly initialized GAT blocks by 0.58%.

TABLE I
EVALUATION OF THE USE OF THE UNSUPERVISED PRETRAINED GAT BLOCK ω_t FOR WEIGHT INITIALIZATION IN VIGAT (FIG.2), ON YLI-MED, EXCLUDING OR INCLUDING VIGAT'S GLOBAL BRANCH.

N	ω_1	ω_2	ω_3	top-1(%)
9	-	Rand Init	Rand Init	87.12
	-	Pretr. ω_t	Pretr. ω_t	**88.70**
25	Rand Init	Rand Init	Rand Init	90.77
	Rand Init	Pretr. ω_t	Pretr. ω_t	**91.62**

TABLE II
ABLATION STUDY DEPICTING THE TOP-1 ACCURACY OF THE LOCAL BRANCH OF VIGAT ON YLI-MED THROUGHOUT DIFFERENT SCENARIOS OF USING NO TRAINABLE COMPONENTS, A PRETRAINED ω_t OR A RANDOMLY INITIALIZED GAT BLOCK.

ω_2	ω_3	weight sharing	top-1(%)
Mean Pooling	Mean Pooling	no	80.92
Pretrained ω_t	Rand Init	no	85.18
Rand Init	Mean Pooling	no	86.51
Rand Init	Rand Init	yes	87.12
Pretrained ω_t	Mean Pooling	no	88.34
Pretrained ω_t	Pretrained ω_t	yes	**88.70**

We also experimented with the complete ViGAT supervised event-recognition architecture of Fig 2, using a total of $N = 25$ frames for both the local and global branches (to allow also the global branch to effectively learn). As presented in the lower part of Table I, by utilizing the unsupervised pretrained ω_t for the initialization of the local-branch GAT blocks, ViGAT yields a significant improvement of 0.85% compared to using randomly initialized GAT blocks. This highlights the effectiveness and potential of unsupervised pretraining in capturing meaningful representations.

C. Ablation Study

In Table II we compare different variants of our model by substituting the GAT blocks initialized with ω_t with alternative training approaches or a simple non-trainable mean pooling. These comparisons specifically focus on the local branch of the ViGAT architecture of Fig.2 (setting $N = 9$), in order to evaluate the effectiveness of knowledge transfer. We observe that utilizing ω_t for initializing both the objects- and the frames-GAT blocks, ω_2 and ω_3, where these two GAT blocks also share weights during the subsequent training on YLI-MED, is the best-performing strategy. It yields significantly improved results compared to either using a single randomly initialized GAT block or relying on randomly initialized GAT blocks for both objects and frames.

V. CONCLUSION

In this paper, we introduced Masked Feature Modelling (MFM), and demonstrated the use of MFM for the unsupervised pre-training of a key component of the state-of-the-art ViGAT event recognition method. Our results showed that using the outcome of unsupervised pre-training for the initialization of certain blocks of ViGAT enables the latter to reach higher accuracy in a downstream task, i.e. when further trained in a supervised way on a small target dataset for event recognition. We believe that this work contributes to advancing our understanding of the benefits and applications of masking techniques in video analysis; and, that such Masked Feature Modelling can be widely applicable in various video classification problems, when the employed learning architecture leverages "features" (typically, visual information embeddings generated by pre-trained deep networks).

REFERENCES

[1] N. Gkalelis, D. Daskalakis, and V. Mezaris, "ViGAT: Bottom-up event recognition and explanation in video using factorized graph attention network," *IEEE Access*, vol. 10, pp. 108 797–108 816, 2022.

[2] S. Atito, M. Awais, and J. Kittler, "Sit: Self-supervised vision transformer," *arXiv preprint arXiv:2104.03602*, 2021.

[3] M. Goulão and A. L. Oliveira, "Pretraining the vision transformer using self-supervised methods for vision based deep reinforcement learning," *arXiv preprint arXiv:2209.10901*, 2022.

[4] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *ICLR*, 2022.

[5] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "BEiT v2: Masked image modeling with vector-quantized visual tokenizers," *arXiv preprint arXiv:2208.06366*, 2022.

[6] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[7] Z. Mnasri, S. Rovetta, and F. Masulli, "Audio surveillance of roads using deep learning and autoencoder-based sample weight initialization," in *IEEE MELECON*, 2020, pp. 99–103.

[8] Z. Yang, X. Jin, K. Zheng, and F. Zhao, "Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification," in *CVPR*, 2022, pp. 14 278–14 287.

[9] R. Girdhar, A. El-Nouby, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Omnimae: Single model masked pretraining on images and videos," in *CVPR*, 2023, pp. 10 406–10 417.

[10] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *CVPR*, 2022, pp. 14 668–14 678.

[11] L. Yu, Y. Cheng, K. Sohn, J. Lezama, H. Zhang *et al.*, "Magvit: Masked generative video transformer," in *CVPR*, 2023, pp. 10 459–10 469.

[12] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *ICCV*, 2021, pp. 6836–6846.

[13] Y. Wang, Y. Yue, X. Xu, A. Hassani, V. Kulikov, N. Orlov, S. Song, H. Shi, and G. Huang, "Adafocusv3: On unified spatial-temporal dynamic video recognition," in *ECCV*, 2022, pp. 226–243.

[14] N. Gkalelis, D. Daskalakis, and V. Mezaris, "Gated-ViGAT: Efficient bottom-up event recognition and explanation using a new frame selection policy and gating mechanism," in *IEEE ISM*, 2022, pp. 113–120.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[16] C.-Y. Wu and P. Krähenbühl, "Towards Long-Form Video Understanding," in *CVPR*, 2021.

[17] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[18] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *ICCV*, 2021, pp. 9620–9629.

[19] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, vol. 11219, 2018, pp. 318–335.

[20] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[21] J. Bernd, D. Borth, B. Elizalde, G. Friedland, H. Gallagher, L. Gottlieb, A. Janin, S. Karabashlieva, J. Takahashi, and J. Won, "The YLI-MED corpus: Characteristics, procedures, and plans," *arXiv preprint arXiv:1503.04250*, 2015.

[22] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *ECCV*, 2022, pp. 350–368.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.