CM3T: Framework for Efficient Multimodal Learning for Inhomogeneous Interaction Datasets

Tanay Agrawal
INRIA
Sophia Antipolis, France
tanay.agrawal@inria.fr

Mohammed Guermal
INRIA
Sophia Antipolis, France

Michal Balazia
INRIA
Sophia Antipolis, France

Francois Bremond
INRIA
Sophia Antipolis, France

Abstract

Challenges in cross-learning involve inhomogeneous or even inadequate amount of training data and lack of resources for retraining large pretrained models. Inspired by transfer learning techniques in NLP, adapters and prefix tuning, this paper presents a new model-agnostic plugin architecture for cross-learning, called CM3T, that adapts transformer-based models to new or missing information. We introduce two adapter blocks: multi-head vision adapters for transfer learning and cross-attention adapters for multimodal learning. Training becomes substantially efficient as the backbone and other plugins do not need to be finetuned along with these additions. Comparative and ablation studies on three datasets Epic-Kitchens-100, MPI-IGroupInteraction and UDIVA v0.5 show efficacy of this framework on different recording settings and tasks. With only 12.8% trainable parameters compared to the backbone to process video input and only 22.3% trainable parameters for two additional modalities, we achieve comparable and even better results than the state-of-the-art. CM3T has no specific requirements for training or pretraining and is a step towards bridging the gap between a general model and specific practical applications of video classification.

1. Introduction

Video classification is a big field in computer vision with various sub-tasks and datasets for each of these tasks. Recently, there has been an increase in tasks, datasets, and recorded modalities. Most work is specific to a task with corresponding datasets or a subset of these modalities, and their modification for a new input protocol is tedious. Methods including late and early fusion and cross-attention are generally used for combining them, but they are not the most efficient way to treat this wide variety of data. Thus, there is a need for a method that can handle this increase in data having high variability in structure and which learns robust relations that are shareable among tasks and datasets. The field of parameter efficient transfer learning (PETL) is increasing in popularity to answer this problem. The basic idea consists in adding adapters[1] (i.e., plugin architectures of very few trainable parameters) to a backbone and only train these while keeping the backbone frozen. With increasing model and dataset sizes, PETL techniques facilitate finetuning only adapters with less resources and time compared to full-finetuning (i.e., backbones + adapters).

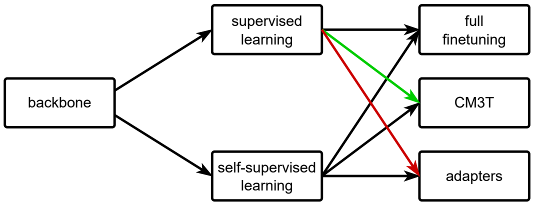

Figure 1. This is a representation of the main problem CM3T aims to solve. Backbones pretrained using self-supervised learning provide good general features, thus all methods of finetuning work well. In the case of supervised pretraining, adapters fail to perform well (in red) and CM3T is introduced to solve this (in green).

The video backbones used as a starting point for PETL can be pretrained using either i) the traditional supervised method on big datasets or using ii) more sophisticated self-supervised methods which result in better general features, such as VideoMAE [36] or contrastive learning such as CLIP [29]. Existing PETL techniques only work well after using the latter (i.e., self-supervised pretrained backbones). But, self-supervised pretrained backbones are not widely available for use off the shelf and their training is resource intensive. For example, dual-path adapters [28] and ST-adapters [27] require a backbone pretrained with CLIP. However, most works on self-supervised pretraining methods only use Vit/ViViT. Swin/Video-Swin transform-

[1] In this paper, we refer to *adapters* including a mix of multiple techniques as in the M&M adapters.

ers have not been pretrained using these self-supervised methods despite their superior performance. The main motivation behind this work is to propose new adapters to work well with traditional supervised pretrained backbones. Figure 1 summarizes this, the red arrow signifies the problem we are trying to solve and the solution is in green.

We introduce CM3T (Cross Multimodal Multi-dataset Multitask Transformer), a novel PETL technique, which can leverage these new adapters. CM3T takes a frozen backbone, for example, the Video Swin Transformer [22] pretrained (i.e., fully fine-tuned) on Kinetics-400 or Something-Something v2, and adds plugins (i.e., adapters) in parallel without changing the backbone architecture. Only these plugins need to be trained for downstream tasks and different datasets. Inspired by the Mix-and-Match (M&M) adapters [38], we combine prefix tuning with a newly introduced plugin, multi-head vision adapters. These adapters (shown in blue in Figure 2.) improve upon existing scaled parallel adapters by separating the processing for different spatial chunks into different heads of the input. This greatly increases performance as interaction datasets generally have almost fixed cameras and various objects and parts of the body always occur in particular spatial locations which generally remain the same. In addition, an approximation for prefix tuning, which has been proven to work well, is used as done by [10], but with some modifications. This is shown in red in Figure 2. The details are discussed in Section 3.

Furthermore, the above idea can be further extended to cross-modal learning where the weights of the pretrained model do not have to be changed to incorporate new modalities, just as the backbone doesn't have to be changed to adapt to new datasets. This facilitates the use of existing work to build more complex systems. For this, we introduce the third and final module in Figure 2, called cross-attention adapters (in green) for multimodal learning. Since cross-attention has been established as an effective manner for multimodal learning, we show how to incorporate it in place of linear layers in adapters, allowing their use for multimodal learning as well. It allows CM3T to learn the relationships between vision and other modalities while retaining its other advantages. This is a challenging task to execute in a resource efficient manner as increasing the number of input modalities generally increases the number of branches, hence the resources used. But, the theory of adapters allow us to overcome this. Thus, this contribution is significant as shown by the results in Section 4. Challenges in processing multimodal data include heterogeneity of the present modalities, lack of correlation between modalities (for example different pitches in the audio could correspond to the same text), and the need for many training samples for convergence. Building upon each challenge above in order, CM3T addresses these challenges with the following additions. Adding a new modality is cumbersome as it requires retraining parts of the backbone along with the new branches for the modality itself, but with this framework, it would just be a new plugin which is trainable by itself. To capture the relationship between different modalities, we add an additional module to capture the relationships between all modalities other than vision (the backbone), when available. To make training faster and convergence easier as compared to using the generic embedding from large transformer models, the downsampling layer in adapters provides a good embedding to use for cross-attention. Additionally, training cross-modal adapters across datasets improves performance and provides a good pretrained feature extractor for small datasets.

To show that CM3T is suitable for multimodal, multi-dataset and multitask learning, we experiment on three different datasets with different recording scenarios and tasks: Epic-Kitchens-100 (EK-100) with first-person human-object interaction videos, MPIIGroupInteraction (MPIIGI) and UDIVA v0.5 (UDIVA) with human-human interactions in group settings while talking or doing different tasks respectively. We choose a mix of small and large multimodal interaction datasets to show the efficacy of our work in different settings. We show that we achieve comparable accuracy to state-of-the-art for all the datasets using only 12.8% trainable parameters as compared to the backbone to process video input and only 22.3% trainable parameters to process two additional modalities. We perform additional experiments to study how CM3T works in different scenarios and explore the reasons for the results obtained.

In summary, our contributions are:

- We introduce multi-head vision adapters which perform well with traditional supervised pretraining, in contrast to existing PETL techniques.
- We introduce cross-attention adapters which are easier to modify than traditional multimodal methods. They also benefit from weight sharing, similarly to traditional adapters, by storing the relations between vision and other modalities and reusing them later.
- We provide a framework, CM3T, for combining these techniques along with an approximation of prefix tuning to achieve state-of-the-art performance.

2. Related Work

2.1. Parameter Efficient Task Adaptation

Transformer-based backbones, such as Video Swin Transformer [22] or ViVit [4], are state-of-the-art feature extractors which are carefully trained on big datasets using either supervised or the better performing self-supervised methods. But finetuning these models is resource-intensive and does not converge for small datasets. The main theory behind all PETL work for computer vision is that finetuning any general feature extractor involves learning the environment in which the new data is recorded and the intricacies of the new task. The basic spatial understanding of the video remains the same. Thus, we can use this basic understanding by these pretrained models and employ only a few additional parameters to learn the new information.

The field of NLP has seen a lot of work following the above idea, such as adapters [11], LoRA [12], and prefix tuning [21]. These methods get similar results while adding less than 10% parameters to existing models which are trained to learn the new task while the pretrained weights are frozen. These have also been extended to computer vision [6, 15, 25, 34].

There are three recent PETL methods which show good results: (1) only updating new parameters added to the model or the input [11, 17, 20, 21]; (2) updating some of the parameters of the model in a sparse manner [35,40,41]; and (3) low-rank factorization of weight matrices to reduce the number of parameters to be updated while keeping the weight matrix approximately the same [13]. Combining these approaches, [10,24] propose a unified parameter efficient training framework. Among these approaches, adapters, which belong to the first category, have been used in computer vision [30, 31] and natural language processing [11, 16, 23]. While adapters add more parameters into models, prompt-based approaches instead add trainable parameters to inputs [9, 20, 21], and experiments have shown their value in language and vision tasks. We use both techniques in [10] as an inspiration for CM3T. VL Adapters [34] compare various adapter techniques [11, 16, 17] applied to question answering tasks, but not to pure vision tasks. Their work aims to use adapters to project vision and language pretrained model embeddings into the language model's space whereas we show that it is possible to do it across vision datasets and also be used to add new modalities.

AdaptFormer [6] uses adapters with only the linear layers of a transformer and achieves better results than full finetuning. But it uses VideoMAE [37] for pretraining ViT [18] which is not feasible if resources are limited and cannot be used to make a generalized framework. Their method fails with models not carefully pretrained using self-supervised methods. Similarly, ST-adapters [27] use ViT pretrained using CLIP. They convert image models to video models using convolutions for time aggregation in addition to the upsampling and downsampling linear layers in an adapter and it works well, except for the case when traditional supervised pretraining is employed. Visual prompt tuning (VPT) [15] uses prompt tuning for images, but prompts alone do not work well for videos which is also mentioned by [6].

The paper [25] shows that adapters only work for vision if the bottleneck dimension is large. They introduce a pruning technique to reduce the size of these adapters. We introduce multi-head vision adapters as an alternative that works well even with a small bottleneck dimension and without any specific pretraining method. Dual-path adaptation from image to video transformers [28] show better results compared to others using supervised training methods, but it is still not comparable to full finetuning. They also have a specific input method that limits the maximum temporal size of input that can be provide which makes their model less scalable and not suitable for all datasets and downstream tasks.

2.2. Multimodal Learning

There is an inherent difference between videos and other modalities, such as audio or text, and thus it is challenging to combine them into one model. VATT [3] uses early fusion, where they concatenate all input modalities. Although the earlier the fusion, the better the results, there is a tradeoff with the amount of data required for training as it is harder for models with early stage fusion to converge which leads to tedious self-supervised learning.

Some works design a specialized architecture for fusion at feature level [1, 26]. These work better but there are limitations as the fusion is done after downsampling the input features which leads to loss of information and poor cross-modality relations. [8,19,33] have feature level fusion with minimal downsampling, but lack in handling specific modalities differently. So, there is a need for a model which can benefit from cross-modality learning at different levels. To answer this and so make the model flexible, we propose using cross-attention added to each block of a transformer architecture. State-of-the-art methods M&M Mix [38] and MuMu [14] are either modality specific or have a rigid architecture making it hard to add and remove modalities. This work addresses these drawbacks by having a flexible architecture that can accommodate any type of input.

3. CM3T Framework

We define an easy way to use existing multimodal data and pretrained models when approaching a video classification or video understanding task. This will assist in bridging the gap between research and practical applications. This section discusses some of the technical details of the background and then the methodology of our work.

3.1. Choosing a Pretrained Model

Our method is focused on transformer-based backbones which have produced state-of-the-art results for various vision tasks. We use the Video Swin Transformer (Video Swin-B) [22], but the following steps of the framework are model invariant and the backbone can be chosen according to the need. The reason for choosing Video Swin-B is that different blocks process the input at different spatial resolutions. Depending on the side input (other modalities),

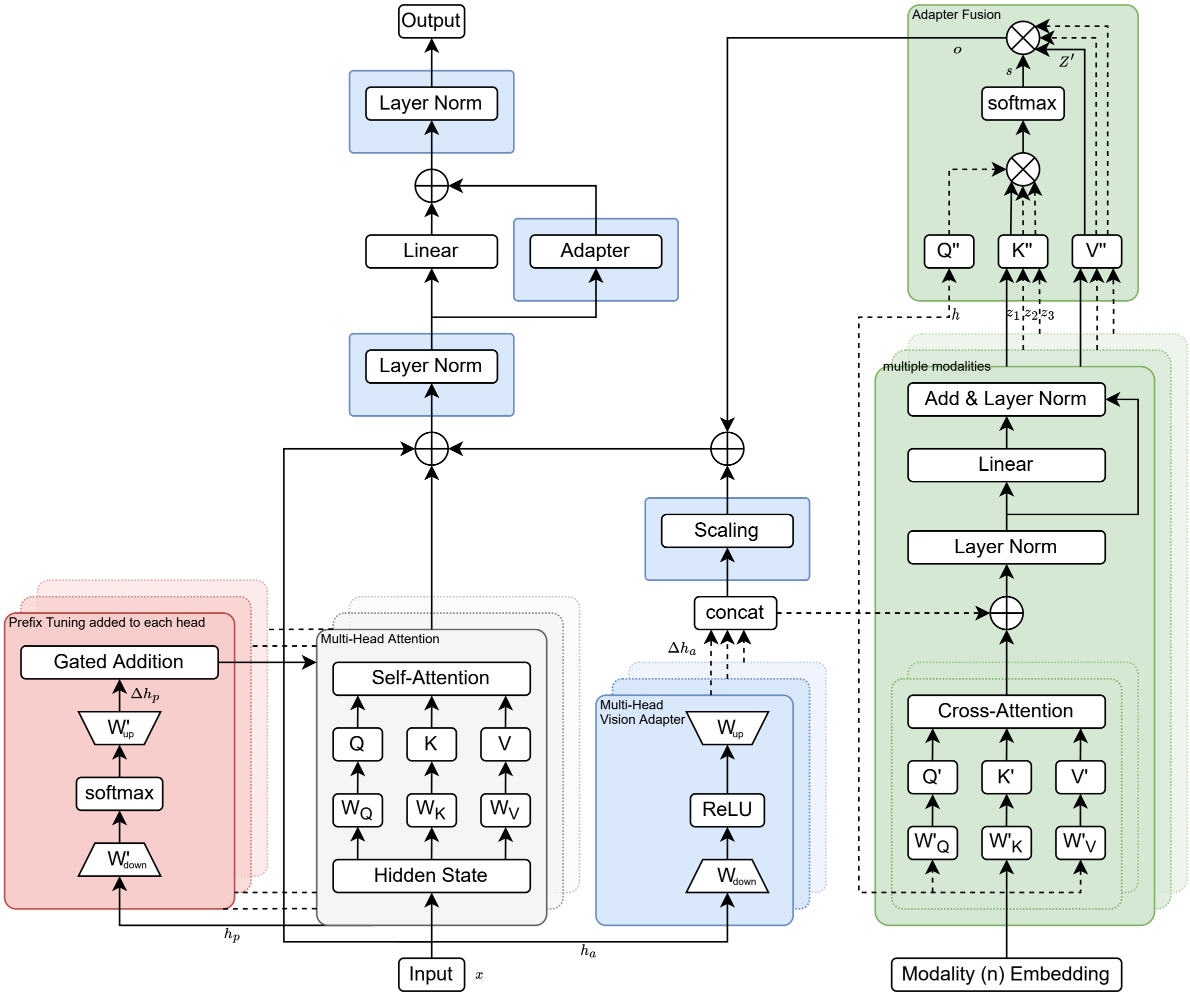

Figure 2. Detailed architecture of CM3T. Colored parts are the ones that are finetuned and the rest are frozen. It has three separate blocks added to it which are shown in three different colors. Prefix tuning is complicated to show in detail, so only a schematic is shown. The rest of the details are described in Section 3.

cross-attention performs well with different blocks, that is, different spatial resolutions.

3.2. Finetuning or Using Adapters

Once we have a pretrained vision model, the next step is to finetune and adapt it to the target dataset. If computational resources or time are a constraint, adding adapters and prefix tuning and training them in place of full finetuning produces comparable results with significantly fewer parameters to train. There is also the possibility of combining this step with the following steps (in this section and the next one) for end-to-end learning, but we perform each step separately to compare their performance with corresponding state-of-the-art. The results for end-to-end training are also shown in the next section.

3.2.1 Background

We take inspiration from scaled parallel adapters and prefix tuning (PT) as used by [10]. Figure 2 shows all the additions to the pretrained model along with our modifications. Multi-head vision adapters (MHVA) (in blue) and prefix tuning (in red) are discussed in this subsection and cross-attention adapters (CAA) (in green) are discussed in Section 3.3.

Mathematically, adapters from [11] are defined as

$$y = s \cdot \Delta h_a \tag{1}$$

$$\Delta h_a = ReLU(h_a W_{down}) \cdot W_{up} \tag{2}$$

where $h_a = x$ is the input of size d, $W_{down} \in \mathbb{R}^{d \times r}$ is the weight matrix for the down-projection layer with bottleneck dimension r, $W_{up} \in \mathbb{R}^{r \times d}$ is the up-projection layer, and s is the scaling factor. We use this in parallel instead of sequential, similar to [10]. We also use their definition for prefix tuning (for simplification, Figure 2 does not show recurrent connection for prefix tuning),

$$h_p \leftarrow (1-\lambda)\cdot h_p + \lambda \cdot \Delta h_p \tag{3}$$

$$\Delta h_p = softmax\left(h_p W_{down}^{'}\right)\cdot W_{up}^{'} \tag{4}$$

$$W_{down}^{'} = W_q P_k^T \qquad W_{up}^{'} = P_v \tag{5}$$

where W_q is the weight matrix for getting query vector from the input h_p, $P_k = C{\cdot}W_k$ and $P_v = C{\cdot}W_v$ are prefix tuning vectors which are learned using W_k and W_v (key and query weight matrices of the transformer backbone). Here, C is a learned embedding which is randomly initialized and λ is the factor used for gated addition. The red part of Figure 2 shows prefix tuning added to transformers, it is added in parallel to each head of multi-head attention.

3.2.2 Incorporating Multi-Head Vision Adapter and Prefix Tuning into CM3T

Using adapters for vision tasks is more challenging than NLP as language understanding does not change with the task or dataset, but video datasets have a wide variety of settings, such as indoor or outdoor recording scene, different views and camera angles, lighting changes, and more. Finetuning allows the networks to overcome these changes, but it is hard for adapters owing to less capability to change the original model's activations. But with a few changes, adapters can show performance comparable to fully finetuned models. Blue parts of Figure 2 mark the adapters.

AdaptFormer [6] adds scaled parallel adapters to linear layers only and achieves better results than finetuning owing to a sophisticated pretrained ViT model using VideoMAE [37]. We achieve very poor performance with the same method without this specific pretraining, even when coupled with prefix tuning. So, this leads to our first change, inspired by multi-head attention, we introduce **Multi-Head Vision Adapter (MHVA)**. This is different from multi-head attention as the input is divided along the window dimension of Video Swin transformers (or spatial patch dimension in ViViT) and not the channel dimension. Essentially, there are different linear layers for different sets of windows/patches. We saw that increasing the bottleneck dimension in adapters only increased the performance slightly (as shown by [25]), but adding the above change allowed the network to learn better even with a smaller bottleneck dimension. Overall, the parameters do not increase by a big margin as compared to traditional adapters as we use a smaller bottleneck dimension. To define the change mathematically, the input h is divided along the window dimension to get $\{h_1, h_2, h_3, \ldots\}$. Each has its own parallel adapter and the output is concatenated along the same dimension before scaling and addition. Extending Equation 3,

$$\{h_{a1}, h_{a2}, \ldots\} \leftarrow \{h_{a1}, h_{a2}, \ldots\} + s\cdot\Delta\{h_{a1}, h_{a2}, \ldots\} \tag{6}$$

where each operation is performed element-wise.

Our second change is that we make the scaling factor for adapters (s in Equation 1) added to linear layers learnable, allowing greater change to activations. Attention in pretrained models might focus on features that are not relevant to the new downstream task or dataset, but this change allows adapters to overcome this. For EK-100, when the value is fixed at 4.0, we achieve 1.1% lower performance.

Without the two changes mentioned above, adapters have very poor performance for the domain of computer vision with traditionally available pretrained models. These adapters are named multi-head vision adapters. These are specific to Video Swin transformers, but the same concept can be applied to modify adapters for any model using different linear layers in adapters for different sets of windows to which attention is applied. Section 4 shows results for ViViT-B as a backbone model too. The reason for good performance with this addition is that it gives the adapters the ability to learn different representations for different chunks of the input.

The third change is more specific as compared to the first two. We use ReLU activation in place of tanH with a lower dropout for **Prefix Tuning (PT)** and that provides a smoother training curve and easier convergence. It also allows for 0.7% gain in accuracy on EK-100 dataset.

3.3. Adding Other Modalities (CAA)

Cross-attention adapters are used for adding modalities to the model received from the previous step. Cross-attention adapters are simply obtained by replacing the two linear layers in the adapters with a cross-attention module. Each added modality has its own adapter. The query and value inputs to this adapter are taken from the concatenation of hidden states from the bottleneck hidden state in the multi-head vision adapter $Q = V = ReLU(xW_{down})$, where h is the input to the Video Swin-B block and W_{down} is the same as that in Equation 2. The key is taken as the feature embedding from the new modality.

To merge all the adapters trained for different modalities, in place of simple addition, **AdapterFusion** [32] is used which captures the interaction between different side inputs i.e, modalities other than vision. It is an attention block where each head has the same query as that of attention in cross-attention adapter for each modality, described above, let's say h. The key and value for each head are taken from the output z_n of each cross-attention adapter with n signifying the n-th modality. The module is expressed as

$$s' = softmax\left(h^T W_Q \bigotimes z_n^T W_K\right), n \in \{1, \ldots, N\} \tag{7}$$

$$z'_n = z_n W_V \quad , \quad n \in \{1, \ldots, N\} \tag{8}$$

$$Z'_n = [z'_0, \ldots, z'_N] \tag{9}$$

$$o = s'^T Z' \tag{10}$$

where o is the output and $W_Q, W_K\ and\ W_V$ are weight matrices for query, key and value respectively, and N is the number of side modalities.

To incorporate a new modality into the model, there are two additions, a new cross-attention adapter and a new concatenation to s and Z'_n vectors above. One disadvantage of this is that model size keeps increasing with more modalities. To alleviate this, the cross-attention module proposed by [2] is used in place of the traditional one and results are shown in Section 4. It makes adding new modalities hard, but it is a trade-off between flexibility and optimizing the usage of resources.

4. Experiments

4.1. Datasets

To show robustness, we experiment using three datasets with different tasks and modalities. First, an egocentric **Epic-Kitchens-100** [7] consisting of three modalities RGB, optical-flow, and audio for actions related to human-object-interactions. Second, **MPIIGroupInteraction** [5] which is a body language dataset aiming at understanding human behavior in human-to-human interactions. For this dataset, we use the following modalities, RGB and audio. Finally, we have **UDIVA v0.5** [26] which tackles the task of human personality analysis, using also different modalities such as RGB, transcript, and audio. Our approach shows effectiveness on all three tasks, proving our approach of bringing adapters mechanisms into vision problems to tackle all the challenges mentioned in the previous parts.

Epic-Kitchens-100 (EK-100) is a first-view and human-object interaction dataset. It contains 89,977 segments of fine-grained actions annotated from 700 long videos. Footage length amounts to 100 hours. It consists of a total of 97 verbs and 300 nouns, each action is a combination of a verb + noun and has a total of 3806 action classes.

MPIIGroupInteraction dataset (MPIIGI) is 26 hours of spontaneous human behavior with 15 distinct body language classes. This dataset presents a novel set of actions which are challenging in computer vision and human-behavior understanding. It consists of body language behaviors such as gesturing, grooming, or fumbling.

UDIVA v0.5 dataset (UDIVA) is 90.5 hours of dyadic interactions among 147 participants distributed in 188 sessions, recorded using multiple audiovisual and physiological sensors. But only half of the data has been released. UDIVA's main task is personality recognition. It has 5 main classes: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (OCEAN).

4.2. SOTA Comparison

In this section we compare our results to the existing SOTA methods for each dataset and related PETL methods. The aim is to achieve similar performance to the methods we compare against while having considerably less trainable parameters.

4.2.1 SOTA Comparison on EK-100 dataset

1) Multimodal methods: Table 1 shows the highest accuracy of M&M Mix [38] on EK-100 dataset [7]. M&M Mix [38] processes each of the three modalities using three branches of ViViT at different spatial resolutions using different sizes of input tubelets and different variants of ViViT. They use additional modules to share information across views and models for different modalities. One branch has more parameters than Video Swin-B, so the total number of parameters is more than three times the number of parameters of Video Swin-B. When taking Video Swin-B trained on Kinetics-400 as the backbone, we achieve performance comparable to the state-of-the-art (SOTA) (only 1.4% worse) with a minuscule number of trained parameters (more than 13 times less). Using a self-supervised trained backbone, CLIP, we achieve SOTA results (with the base variant of the backbone).

2) PETL methods: Table 1 shows the comparison against SOTA PETL techniques, Dual-path adapters [28] and ST-Adapters [27]. They do not provide these results and the results stated are from our own experiments using their code. We achieve considerably better performance when the pretraining protocol is the same. The results shows that it is hard to overcome the gap created by better pretraining as PETL techniques add minimal processing capacity. But, our plugins allow better performance when self-supervised pretrained backbones are not available.

To show the robustness of our proposed adapters design, we compare the proposed MHVA against the typical adapters from AdaptFormer [6]. We compare the results of CM3T and adapters without additional modalities. Scaled parallel adapters (used in AdaptFormer) with PT achieve 28.7% whereas MHVA achieves 39.8%. This shows that our design of adapters is more robust. The motivation for the change discussed in the methodology section is thus justified from these results.

4.2.2 SOTA Comparison on UDIVA and MPIIGI

For UDIVA and MPIIGI, we compare to FAt transformers [2], the SOTA for these datasets. FAt transformers have a lot of additions, specifically for UDIVA, which is the reason for their good performance. They have additional input branches with face crops and contextual videos and a complex method for preprocessing too. Tables 2 and 3 show a

Table 1. SOTA comparison on EK-100. *Acronyms- MHVA: Multi-head vision adapter, PT: Prefix Tuning, CAA: Cross Attention Adapters, CM3T: MHVA + PT + CAA, K400: Kinetics-400. Epochs presented are the number of epochs taken for convergence. All backbones other than CLIP-B are pretrained on Kinetics-400.*

Method	Backbone	Top-1 accuracy (%)	Epochs	GFPLOs
Multimodal methods				
M&M Mix [38]	ViViT	49.6	50	>4790
CM3T	Video Swin-B	48.2	22	616
CM3T	CLIP-B	**50.1**	18	754
PETL Methods				
Dual-Path Adapters [28]	ViT-B	35.8	21	642
ST-Adapters [27]	ViT-B	34.3	18	911
Dual-Path Adapters [28]	CLIP-B	44.8	24	642
ST-Adapters [27]	CLIP-B	44.1	18	911
Adaptformer [6] + PT	Video Swin-B	28.7	6	357
MHVA + PT	Video Swin-B	39.8	14	449
MHVA + PT	CLIP-B	**45.5**	13	589

comparison against the published results. As for MPIIGI, we achieve better results with transfer learning techniques than FAt transformers which are fully finetuned. There are two reasons for this. One is that MPIIGI is a small dataset and it is easier for these PETL techniques to converge. The second reason is that Kinetics-400 is very close to MPIIGI and the CM3T backbone networks are initialized very well. This enables adapters to work better. In summary, CM3T achieves results equivalent to the SOTA for these two datasets with around 5 times less trainable parameters as compared to the previous SOTA. Using CLIP backbone, we achieve SOTA results.

We also show that our findings are consistent in the domain of PETL methods as we outperform ST-Adapters using our plugins.

Table 2. SOTA comparison on UDIVA. Acronyms from Table 1.

Method	Backbone	Mean MSE	Epochs
Multimodal methods			
FAt transformers [2]	-	0.72	30
CM3T	Video Swin-B	0.69	27
CM3T	CLIP	**0.65**	22
PETL Methods			
ST-Adapters [27]	CLIP	0.91	14
MHVA + PT	CLIP-B	**0.8**	14

Table 3. SOTA comparison on MPIIGI. Acronyms from Table 1.

Method	Backbone	mAP	Epochs
Multimodal methods			
FAt transformers [2]	-	0.899	18
CM3T	Video Swin-B	0.901	9
CM3T	CLIP	**0.918**	11
PETL Methods			
ST-Adapters [27]	CLIP	0.886	14
MHVA + PT	CLIP-B	0.**894**	9

4.2.3 Baseline Comparison

Video Swin is one of the SOTA transformers trained on many datasets and tasks, hence it is chosen as the backbone. In Table 4 we compare to full-finetuning the backbone vs. frozen backbone and only our plugins trained. For each dataset, we compare for multimodal input and only RGB input.

For only RGB input, we achieve slightly lower results than full-finetuning. This is in tune with what we expect as traditionally pretrained backbones do not provide good generalizable features that can extend to other datasets and adapters have a limited capacity to take into account the distribution shift of the input. But as shown for EK-100, our plugins perform better than SOTA PETL methods when the same pretraining is applied.

Looking at multimodal input, for EK-100 dataset, CM3T achieves an accuracy of only 0.7% lower than the fully finetuned model. Our method achieves comparable results with only 22.3% parameters whereas Video Swin-B combining CAA goes up to 109.5% parameters (compared to Video Swin-B). Top-1 accuracy is the metric used here.

Moreover, for UDIVA [26] and MPIIGI [5], we achieve the same results with CM3T as with full-finetuning and again with only 22.3% of the total number of parameters in Video Swin-B. Mean MSE and mAP are used as metrics for them respectively.

Table 4. Baseline comparison. Acronyms from Table 1. Top-1 accuracy for EK-100, MSE for UDIVA and mAP for MPIIGI. Number of trained parameters are reported on a relative scale, 100% is equivalent to 88M.

Method	Backbone	Eval. metric	Epochs	Trained params
RGB input (EK-100)				
Full finetuning	Video Swin-B	**41.7%**	49	100.0%
MHVA + PT	Video Swin-B	39.8%	14	12.8%
Multimodal input (EK-100)				
Full finetuning + CAA	Video Swin-B	**48.9%**	56	109.5%
CM3T	Video Swin-B	48.2%	22	22.3%
RGB input (UDIVA)				
Full finetuning	Video Swin-B	**0.82**	51	100.0%
MHVA + PT	Video Swin-B	0.85	35	12.8%
Multimodal input (UDIVA)				
Full finetuning + CAA	Video Swin-B	0.69	32	116.1%
CM3T	Video Swin-B	0.69	27	28.9%
RGB input (MPIIGI)				
Full finetuning	Video Swin-B	**0.887**	17	100.0%
MHVA + PT	Video Swin-B	0.882	8	12.8%
Multimodal input (MPIIGI)				
Full finetuning + CAA	Video Swin-B	0.901	18	116.1%
CM3T	Video Swin-B	0.901	9	28.9%

4.2.4 Cross-Attention Module

An interesting thing to note is that MTV-B which is the base model for M&M Mix and uses only RGB videos as input, achieves 46.7% accuracy and there is only a 2.9% accuracy increase when optical flow and audio are added to it. We achieve a higher increase of 8.4% with CM3T when the two modalities are added. This might be because MTV-B is a better backbone as compared to Video Swin-B and captures most of the information present in optical flow already as optical flow is also a visual feature. Thus adding optical flow does not increase performance for them as much as us. This proves the efficacy of cross-attention adapters as we achieve similar performance to M&M Mix, even when we are comparatively farther as compared to MTV-B. Moreover, we compare two methods for cross-attention: MMCA [2] and our proposed CAA and we observe that with our proposed solution we can achieve 0.5% higher accuracy, showcasing robustness and efficacy of the proposed CAA. All results are in Table 5.

Table 5. Experiments for efficacy of CAA. Acronyms from Table 1. *MMCA: multimodality cross-attention [2]*

Method	Backbone	Accuracy (%)	Epochs	Trained params
RGB input				
MTV-B [39]	-	46.7	80	>100.0%
MHVA + PT	Video Swin-B	39.8	14	12.8%
Multimodal input				
M&M Mix [38]	-	49.6	50	>300.0%
CM3T: MHVA + PT + CAA	Video Swin-B	48.2	22	22.3%
MHVA + PT + (MMCA [2])	Video Swin-B	47.7%	24	22.7%

4.3. MHVA / PT

MHVA and PT work well, as shown above. But, PT alone does not work very well as it tries to find learnable fixed inputs to be added to the actual input to provide context, but since supervised pretrained models do not give good relevant features for a different dataset, these inputs are not very useful unless combined with MHVA which provides a way for the model to learn the distribution shift in the input associated with the new dataset.

4.4. Different Backbones

This experiment supports our claim of CM3T being model-agnostic. Our plugins trained along with frozen ViViT-B achieve even better performance than full-finetuning. Table 6 shows CM3T achieving better results with Video Swin as it is a better backbone, but this comparison does not say anything about out modules and is included here just for completeness.

Table 6. Results using different backbones. Experiments were done on EK-100 dataset. Acronyms from Table 1.

Method	Backbone	Accuracy (%)
Backbone with supervised pretraining using K400		
Full finetuning	ViViT-B	37.4%
MHVA + PT	ViViT-B	38.1%
CM3T	ViViT-B	44.3%

4.4.1 Computational Resources

We state that CM3T saves computational resources and we have already discussed a reduction in trainable parameters. Table 1 shows that fewer epochs are required for the convergence of models with our plugins and also low FLOPs. For just finetuning RGB models, multi-head vision adapters and prefix tuning require a third of the time as compared to full finetuning. For adding a new modality, given an embedding corresponding to features of the new modality, only 5.8M additional parameters are required (with Video Swin-B as the backbone).

5. Conclusion

In this work, we presented CM3T (Cross Multimodal Multi-dataset Multitask Transformer), a framework for using common pretrained video classification models with a transformer-based architecture. The framework consists of three modules, two introduced by us, multi-head vision adapters and cross-attention adapters, and one already existing, prefix tuning. We show that in contrast to previous related works, these work well without specific pretraining or training methods (self-supervised methods) and study different variants. This work helps bridge the gap between research and practical applications of video classification models by making it easier to adapt existing work to new datasets and tasks, and also to utilize additional available modalities. Also, the framework benefits from weight sharing across different datasets for the same modalities.

The limitation of this approach is that if the dataset used for pretraining is very dissimilar to the target one, the results will not be good. The frozen pretrained model needs to have the relevant information for the target task or dataset. Using various data augmentation, self-learning methods, or fully finetuned smaller models might give better results. For future work, combining adapters with selective finetuning of the model might resolve the above issue while keeping a low number of trainable parameters.

Ethical Statement

This research adheres to the highest ethical standards, ensuring the welfare, dignity, and rights of all individuals involved. Even though this work does not record new data, the used data was collected with an informed consent obtained from all participants. No harm or bias has been detected during our research activities. We contribute to the advancement of knowledge while prioritizing the well-being of those involved within the scope of the European GDPR regulations.

Acknowledgements

This work was supported by European Union under Horizon Europe project GAIN (GA #101078950) and by the French National Research Agency under UCA^{JEDI} Investments into the Future (ANR-15-IDEX-01).

References

[1] Tanay Agrawal, Dhruv Agarwal, Michal Balazia, Neelabh Sinha, and Francois Bremond. Multimodal personality recognition using cross-attention transformer and behaviour encoding. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*, pages 501–508. INSTICC, SciTePress, 2022. 3

[2] Tanay Agrawal, Michal Balazia, Philipp Müller, and François Brémond. Multimodal vision transformers with forced attention for behavior analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3392–3402, January 2023. 6, 7, 8

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. *CoRR*, abs/2104.11178, 2021. 3

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, October 2021. 2

[5] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and François Brémond. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 70–79, 2022. 6, 7

[6] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022. 3, 5, 6, 7

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 6

[8] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7, Seattle, USA, July 2020. Association for Computational Linguistics. 3

[9] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*, 2021. 3

[10] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 2, 3, 4, 5

[11] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019. 3, 4

[12] Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 3

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[14] Md Mofijul Islam and Tariq Iqbal. Mumu: Cooperative multitask learning-based guided multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(1), pages 1043–1051, 2022. 3

[15] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. 3

[16] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021. 3

[17] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Annual Meeting of the Association for Computational Linguistics*, 2021. 3

[18] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 3

[19] Ayush Kumar and Jithendra Vepa. Gated mechanism for attention based multimodal sentiment analysis. *CoRR*, abs/2003.01043, 2020. 3

[20] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 3

[21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, June 2022. 2, 3

[23] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021. 3

[24] Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Wen-tau Yih, and Madian Khabsa. Unipelt: A unified framework for parameter-efficient language model tuning. *arXiv preprint arXiv:2110.07577*, 2021. 3

[25] Imad Eddine Marouf, Enzo Tartaglione, and Stéphane Lathuilière. Tiny adapters for vision transformers, 2023. 3, 5

[26] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1–12, 2021. 3, 6, 7

[27] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning for action recognition. *arXiv preprint arXiv:2206.13559*, 2022. 1, 3, 6, 7

[28] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Dual-path adaptation from image to video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2203–2213, June 2023. 1, 3, 6, 7

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021. 1

[30] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017. 3

[31] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[32] Ipek Baris Schlicht, Lucie Flek, and Paolo Rosso. Multilingual detection of check-worthy claims using world languages and adapter fusion. *arXiv preprint arXiv:2301.05494*, 2023. 5

[33] Aman Shenoy and Ashish Sardana. Multilogue-net: A context-aware rnn for multi-modal emotion detection and sentiment analysis in conversation. *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 2020. 3

[34] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 3

[35] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021. 3

[36] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 1

[37] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 3, 5

[38] Xuehan Xiong, Anurag Arnab, Arsha Nagrani, and Cordelia Schmid. M&m mix: A multimodal multiview transformer ensemble. *arXiv preprint arXiv:2206.09852*, 2022. 2, 3, 6, 7, 8

[39] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022. 8

[40] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 3

[41] Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021. 3