From Multimodal to Unimodal Attention in Transformers using Knowledge Distillation

Dhruv Agarwal[*†1,2], Tanay Agrawal[*‡1], Laura M. Ferrari[‡1,3], and François Bremond[‡1,3]

[1]INRIA Sophia Antipolis - Méditerranée, France
[2]Indian Institute of Information Technology, Allahabad, India
[3]Université Côte d'Azur, France
[†]drv.agwl@gmail.com [‡]name.surname@inria.fr

Abstract

Multimodal Deep Learning has garnered much interest, and transformers have triggered novel approaches, thanks to the cross-attention mechanism. Here we propose an approach to deal with two key existing challenges: the high computational resource demanded and the issue of missing modalities. We introduce for the first time the concept of knowledge distillation in transformers to use only one modality at inference time. We report a full study analyzing multiple student-teacher configurations, levels at which distillation is applied, and different methodologies. With the best configuration, we improved the state-of-the-art accuracy by 3%, we reduced the number of parameters by 2.5 times and the inference time by 22%. Such performance-computation tradeoff can be exploited in many applications and we aim at opening a new research area where the deployment of complex models with limited resources is demanded

1. Introduction

New deep learning models are introduced at a rapid rate for applications in various areas. Among others, transformers are one of those that have sparked great interest in the computer vision community for vision and multimodal learning [9]. Numerous model variants of the paper "Attention is All You Need" [23] have been proposed and many works have centered around the attention mechanism. In multimodal learning the, cross-attention concept has been proven to be an efficient way to incorporate attention in one modality based on others. A key existing challenge is the high computational resource demanded, even at inference time, and therefore the need for GPUs with large memory.

In this paper, we study how to minimize the time and resources required using a transformer based model trained on multimodal data. We utilize the concept of knowledge distillation (KD) to use only one modality at inference time. Basically, KD transfers knowledge from one deep learning model (the teacher) to another (the student), and originally, it was introduced as a technique to reduce the distance between the probability distributions of the output classes of both networks [8]. When it comes to applying KD to tranformers the first issue is related to the level on which to apply it. In a cross-attention transformer the modalities other than the primary one, the query, are taken as key and value. The primary modality has skip connections allowing it to preserve more information during back-propagation. Thus, when trying to distill information from the teacher (e.g. a cross-attention transformer), at the output level, the student does not have enough knowledge about the other modalities. Thereafter, distillation need to take place at lower feature levels. In order to do so, we propose here the use of Contrastive Representation Distillation (CRD) [21], a recent method for KD with state-of-the-art performances which enables transfer of knowledge about representations. As KD has been effectively applied in deep learning but little has been explored with transformers, a study is here presented comparing multiple student-teacher configurations, KD applied on multiple levels, and comparison of two methods for distillation.

We chose emotion recognition as our application since it is a complex problem where multimodality showed significant improvement [17]. Human emotions possess not only the behavioral component, expressed as visual attributes; they have as well the cognitive and the physiological aspects. This rich information allows us to emphasize and interact with others. For this reason, the use of multimodal

[*]These authors contributed equally to this work

inputs, as verbal (e.g. spoken words) and acoustic features (e.g. tone of voice), permitted to combine with the video data complementary salient information increasing the results accuracy. Nevertheless, the limits of multimodal analysis are various. Besides the high computational cost, one of the main issues is related to missing modalities. In the real world, we may not have all the modalities available and some parts of the data can go missing due to hardware, software or human faults. Moreover, the use of heterogeneous data causes synchronization issues, linked to the frequency at which the modalities are recorded. To address these aspects we propose a novel approach and our contributions are summarized below.

1. At the best of our knowledge, we introduce for the first time KD in a transformer architecture for multimodal machine learning. We reduce the computational cost at inference time and able to use just one modality (e.g. video), improving the state-of-the-art accuracy by 3%.

2. We develop a framework to study multiple configurations of student-teacher networks and distillation. We especially compare two methods to distill knowledge, the CRD and the use of cross-entropy loss in the attention module of transformers. We name this second method as Entropy Distillation in Attention Maps (EDAM).

The remaining of this paper is organized as follows. Section 2 reports the main findings in the state-of-the-art with respect to multimodal machine learning, transformers and KD. Section 3 describes the dataset exploited for the experiments. In Section 4, we present the developed detailed framework of the student-teacher network architectures and the KD methods applied. Section 5 discusses the results and section-6 concludes with a summary of our contributions and future perspectives.

2. Related Work

Multimodal machine learning aims at developing models that can process information from multiple input types. One of the earliest work was in audio-visual speech recognition [26], while more recently other fields have started exploiting multimodality as multimedia content indexing [20] and emotions [3, 28] or personality [12] recognition. First works in the field proposed the use of early fusion [11, 16] while late fusion has started to be explored more recently [24]. In the last years more complex architectures have moreover been proposed, as the Tensor Fusion Network for intra- and inter- modality dynamics [27] or the use of an attention gate to shift the words through visual and acoustic features [25]. From the famous work by Vaswani et al. [23], in 2019, Tsai et al. [22] proposed the Multimodal Transformer (MulT) to learn representations directly from unaligned multimodal data. In this work cross-modal attention is used to latently adapt elements across modalities, in an end-to-end manner, without explicitly aligning the data and inferring long term dependencies across modalities. In our proposed framework we exploited the MulT architecture for developing the student and teacher network architectures.

The research field of Multimodal Machine Learning brings some unique challenges, given the heterogeneity of the data, that have been deeply explored [14]. One of the main core challenges is the so called co-learning, that is related to the capability of transferring knowledge between modalities. Another core challenge is the representation of heterogeneous data, meaning learning how to summarize multimodal data exploiting their complementarity [1]. This aspect is linked to the model robustness, that is a relevant theme in real world applications, which in turn is correlated to missing data/modalities. Finally, a common general issue of these models is their high computational cost due to their complex architecture, especially when dealing with transformers [10]. This is an emerging area of interest and just lately a multimodal end-to-end model have been proposed with a sparse cross-modal attention mechanism to reduce the computational overhead [6]. An unexplored method to reduce the inference time and dealing with missing modalities is the use of KD. The idea is to leverage on multimodality to build rich representational knowledge that can be transferred to a lighter network, able to infer through just one modality. The concept of KD was firstly introduced by Bucilua et al. [2] and Hinton et al.[8], where a large teacher network is distilled into a small and faster student network to improve the performance of the student at test time. From here on a wide variety of works have been proposed [18, 5] and recently many self-supervised methods leveraging on contrastive loss achieved good performance [13, 4]. Among those works, the CRD method proposed by Tian et al. [21] reported state-of-the-art results proposing to use a contrastive learning strategy for transferring the knowledge in multiple scenarios, comprehending cross-modal transfer and ensemble distillation.

In this work we analyzed the use of CRD in multiple levels of the proposed framework and in a deeper one, in the attention maps of the transformers, we compared the CRD with what we called the EDAM method. We opted for the EDAM methodology as it is a simpler and runs with less computational cost. This comparison has been performed uniquely at the attention layer as this is the only level that outputs probability distribution, essential for the EDAM loss (i.e. cross-entropy loss).

3. Dataset

The dataset used to do experiments in the proposed framework is the CMU-MOSEI dataset [28], which is one

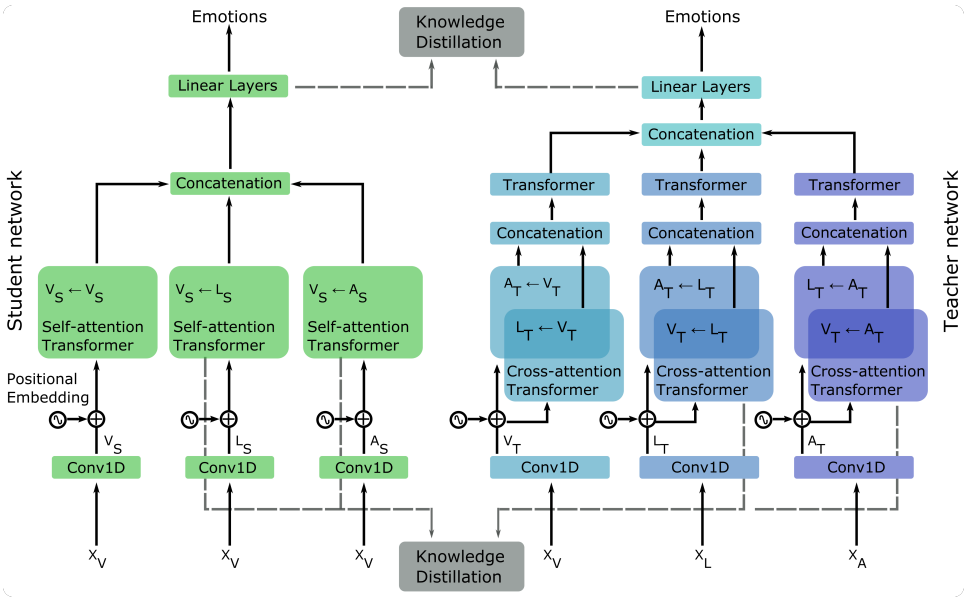

Figure 1: The proposed framework, with the student and teacher networks. The knowledge distillation is performed at high level, in the linear layers, and at lower level, inside the transformers. The student network (left) here is the simplified 3-transformers based architecture, while the teacher (right) is the Complete Teacher Network with 9 transformers

of the largest multimodal datasets for emotion recognition. It is made up of 23,454 movie review video clips taken from YouTube. The movie review clips have been selected from more than 1000 speakers and they are gender balanced. The dataset contains three data modalities, video, audio and text. The video and the audio data are directly extracted from the clips while the language data are transcribed and properly punctuated. Each sample from the dataset is labeled by human annotators with continuous values from -3 (strongly negative emotion) to 3 (strongly positive emotion), which we discretized in 7 emotion classes (-3, -2, -1, 0, 1, 2, 3) to model our emotion classification task. Due to the variable sampling rates of different modalities, there is an inherent non-alignment in the dataset. The dataset is moreover available in both the aligned and non-aligned mode and we used non-aligned one in our task. We used the extracted features by [22] for each modality from the dataset. For video, Facenet [19] was adopted to give 35 facial action units showing muscle movement for each from the video. For audio, COVAREP [7] is used to get low level acoustic features. As language is the transcript of the audio, the chosen audio features represent the non-semantic features like pitch tracking, loudness and maxima dispersion quotients. For language, Glove word embeddings (glove.840B.300d) [15] is used to encode the transcript words in 300 dimensional vectors.

4. The Proposed Framework

The framework is a student-teacher network based on transformers. The student network exploits self-attention (uni-modal attention) while the teacher exploits cross-attention (cross-modal attention) on the input sequences, as detailed in section 4.1. The outputs from the transformers are concatenated and passed through linear layers to make the final predictions on emotion classes (Fig 1). In section 4.2 we present the multiple student-teacher configurations experimented here. The capability of transferring knowledge from multiple modalities to one is performed through distillation, at four levels of the framework. As described in section 4.3 we employ KD in the linear layers and in the transformer layers (Fig 1). For each of these cases, we calculate the final loss ($Loss$) as the sum of the distillation loss(L_{KD}) and classification loss (cross-entropy loss, L_c), as stated in equation (1), which is used to train the student network. While training, we keep the trained weights of the teacher network fixed and only backpropagate gradients through the student network.

$$Loss = \alpha L_c + \beta L_{KD} \quad (1)$$

where α, β are hyperparameters in the range [0, 1]

4.1. Teacher and student Transformers

Teacher and student networks take multimodal sequences (Video, Audio, Language) as input and outputs the

likelihood of the 7 emotion classes. The architecture of the networks are derived from [22]. and are detailed below.

Teacher network: The teacher network takes in input all the modalities and pass them through the respective Conv1D layers, along the temporal dimension, with kernel size of 3. This passage ensures each element in the sequence is aware of its neighboring elements. The outputs from the Conv1D layers are added with sinusodial positional embeddings before they enter into the transformers, to carry the temporal information [23]. Since we have 3 modalities in the teacher network, we use 6 transformers applying cross-attention in every combination of Query (Q), Key (K) and Value (V) pairs, coming from 2 different modalities. For example, the transformer $V_T \leftarrow A_T$ (Fig 1) represents a transformer where Q is from modality A (Audio) and K and V are from modality V (Video). The transformers $A_T \leftarrow V_T$ and $L_T \leftarrow V_T$ form the **Video Branch** of the teacher network, as they both use video modality as Q and have the same output dimension at the end of all transformer layers. Similarly, $V_T \leftarrow A_T$ and $L_T \leftarrow A_T$ form the **Audio Branch**, and $V_T \leftarrow L_T$ and $A_T \leftarrow L_T$ form the **Language branch**. All three branches together form the **Complete Teacher Network** represented in Fig 1. We experimented on four teacher configurations. One is the Complete Teacher Network and the other three are the individual branches: the Video, the Audio and the Language branches. The output sequences from each of the 2 transformers in a branch are concatenated along the last dimension to be further input respectively to another set of 3 transformers. This set of transformers applies temporal attention on the fused output of previous transformers. The last sequence elements from the output sequences of the 3 transformers are, individually passed (for individual branches), concatenated and passed (for Complete Teacher Network), through linear layers.

Student Network: In the student network the Video is the only input modality and it is passed to the model in parallel three times as shown in Fig 1. Each of the inputs goes through a Conv1D layer with different sized kernels to downsample them. Outputs of the 2^{nd} and 3^{rd} Conv1D layers (from left in Fig 1) are used as proxy for missing audio and language modality sequences and are hence named A_S and L_S respectively. The naming scheme of the transformers in the student network is identical to that of teacher network with subscript "S" in place of "T" to distinguish between "Student" and "Teacher".

We experimented on four student configurations. One is developed same as the Complete Teacher Network, with 9 transformers. The other three configurations are a simplified version of the previous, with 3 transformers. These 3 transformers vary for each student configuration as detailed in the next section, one of those configuration (the 5th configuration in section 4.2) is reported in Fig 1. The last sequence elements of output sequences from all 3 transformers are concatenated and passed through linear layers.

The following section elaborate the different configurations of student-teacher networks.

4.2. Student-Teacher Configurations

We study multiple pairs of student and teacher networks to see the effect of distillation on the different architectures.

1. **KD from the Complete Teacher Network**: The teacher is the **Complete Teacher Network** and the student is constructed exactly the same, except for the inputs. All the 9 cross-attention transformers from the teacher are used for distillation to the corresponding self-attention transformers in the student network.

2. **KD from Video Branch**: The teacher is the **Video Branch** and the student network consists of 3 transformers: $V_S \leftarrow V_S$, $A_S \leftarrow V_S$, and $L_S \leftarrow V_S$. The transformers $A_T \leftarrow V_T$ and $L_T \leftarrow V_T$ are used for distillation to corresponding transformers in the student.

3. **KD from Language Branch**: The teacher is the **Language Branch** and the student network consists of 3 transformers, $V_S \leftarrow V_S$, $V_S \leftarrow L_S$, and $A_S \leftarrow L_S$. The transformers $V_T \leftarrow L_T$ and $A_T \leftarrow L_T$ are used for distillation to corresponding transformers in the student.

4. **KD from Audio Branch**: The teacher is the **Audio Branch** and the student network consists of 3 transformers, $V_S \leftarrow V_S$, $V_S \leftarrow A_S$, and $L_S \leftarrow A_S$. The transformers $V_T \leftarrow A_T$ and $L_T \leftarrow A_T$ are used for distillation to corresponding transformers in the student.

5. **KD from Language and Audio Branches of Complete Teacher Network**: The teacher is the **Complete Teacher Network** and the student network consists of 3 transformers, $V_S \leftarrow V_S$, $V_S \leftarrow L_S$, and $V_S \leftarrow A_S$ as depicted in the Fig 1. The transformers $V_T \leftarrow A_T$ and $V_T \leftarrow L_T$ are used for distillation to corresponding transformers in the student.

4.3. Knowledge Distillation

Since different layers in a deep network carry different information, we explore applying distillation to various stages of the network and we compare two methods for loss calculation, CRD and EDAM. We apply CRD at four stages, two at high-level features, in the final and penultimate linear layers, and the other two at the transformer level, where we compare the use of contrastive with cross-entropy loss.

Overview of Contrastive Representation Distillation (CRD): The CRD method [21] provides a general framework to bring closer the representation of "positive pairs" from student and teacher networks while pushing apart the representations of "negative pairs". A pair is called positive when the same sample from a dataset is provided to both the networks, while it is negative when different samples from the dataset are input to the networks. In the following, we use the expression, **CRDLoss(X, Y)** to imply CRD being applied on 1D feature vectors X and Y.

1. **CRD on final and penultimate layers:** CRD is applied on the final and penultimate layers of the student and teacher networks. Here the L_{KD} is calculated by applying the CRD loss at the output of the respective layers from the student and the teacher network, as stated in the following equation

$$L_{KD} = CRDLoss(Student^{(l)}, Teacher^{(l)}) \quad (2)$$

where l indicates the l^{th} layer of the networks.

2. **CRD on post-attention linear layers:** CRD is applied on the outputs of the attention module of the student-teacher transformer pairs (see Fig-1 in Supplementary Materials). Here the L_{KD} is calculated as stated in the equations (3) - (7) and we take the average of computed L_{KD} for all the pairs when we have multiple pairs in an configuration.

$$Attention(Q_X, K_Y, V_Y) = softmax(\frac{Q_X K_Y^T}{\sqrt{d_k}})V_Y \quad (3)$$

$$DisLvl(Y \leftarrow X) = Attention(Q_X, K_Y, V_Y) \quad (4)$$

where X and Y denote output sequences from Conv1D layers (i.e. V_S, A_S, L_S, V_T A_T, and L_T), and Y $\leftarrow$ X indicates that the sequence X serves as Q, while K, V of the transformer come from sequence Y.

Since our transformers have multiple attention layers, we store the $DisLvl(Y \leftarrow X)$ computed from multiple layers in the variable, $DislList$ as follows

$$DisList^{\alpha_S \leftarrow \beta_S} = [DisLvl(\alpha_S \leftarrow \beta_S)]_{i=1}^{i=l} \quad (5)$$

$$DisList^{\alpha_T \leftarrow \beta_T} = [DisLvl(\alpha_T \leftarrow \beta_T)]_{i=1}^{i=l} \quad (6)$$

where α and β are input modalities (Video, Audio, or Language), subscripts "S" and "T" refer to Student and Teacher respectively, and l is the number of layers in the student and teacher transformers. Finally, L_{KD} is calculated using equation (7)

$$L_{KD}^{\alpha \leftarrow \beta} = \frac{\sum_{i=1}^{l} CRDLoss(DisList_i^{\alpha_S \leftarrow \beta_S}, DisList_i^{\alpha_T \leftarrow \beta_T})}{l} \quad (7)$$

3. **CRD on Attention Maps:**. CRD loss is applied on the attention map of each transformer pairs (see Fig-1 in supplementary Materials). The attention map of a transformer layer is described in the following equation (8).

$$DisLvl(Y \leftarrow X) = softmax(\frac{Q_X K_Y^T}{\sqrt{d_k}}) \quad (8)$$

In our cross-attention transformers (the teacher network), Q comes from different modality than K and V while all are from the same modality in case of self-attention transformers (in student network). The teacher network benefits from multiple modalities as it has the opportunity of extracting features from all combinations of modalities put as Q and (K, V), which however is not the case with student networks. Distilling different attention maps from multiple transformers can account for the missing modalities in the student networks to an extent and can be conducive in learning richer representations. In an attention map, the sum of values in every row is 1 and every cell has value in the range $0 - 1$. We therefore see every row as a probability distribution over n classes where n is the number of columns in the attention map of dimension $m \times n$. Through distillation we aim to bring closer the probability distribution of student network to the teacher network and at the same time push apart the distributions of positive samples away from negative samples. (L_{KD}) for one transformer pair is thus calculated using equations, (5), (6) (with $DisLvl$ defined in equation (8)), and (9). In the configurations with multiple transformer pairs, average over L_{KD} obtained from all pairs serves as the final L_{KD}

$$L_{KD}^{\alpha \leftarrow \beta} = \frac{\sum_{i=1}^{l} \sum_{j=1}^{m} CRDLoss(DisList_{ij}^{\alpha_S \leftarrow \beta_S}, DisList_{ij}^{\alpha_T \leftarrow \beta_T})}{ml} \quad (9)$$

where $DisList_{ij}^{Y \leftarrow X}$ denotes the j^{th} row of the i^{th} layer attention map from transformer $Y \leftarrow X$. m = number of rows in attention map and l = number of layers in the transformer

4. **EDAM on Attention Maps:** The cross-entropy loss is applied on the attention maps of each transformer pairs. For an attention map of m × n dimension, we view the rows as probability distribution over n classes and define the distillation loss as a modification to classical cross-entropy loss between student and teacher attention map rows. Similar to the case of CRD on attention maps, we obtain attention maps from all transformer layers using equations (5) and (6) (with Attention map / DisLvl defined using equation (10) and (11))

$$DisLvl(Y \leftarrow X, t) = temp\_softmax(\frac{Q_X K_Y^T}{\sqrt{d_k}}, t) \tag{10}$$

$$temp\_softmax(X, t)^{(i)} = \frac{exp(X^{(i)}/t)}{\sum_{j=1}^{k} exp(X^{(j)}/t)} \tag{11}$$

where t is the temperature parameter which adjusts the sharpening of distribution.

and calculate the distillation loss for a pair using the equations (12) and (13).

$$L_{KD}^{\alpha \leftarrow \beta} = \frac{\sum_{i=1}^{l} \sum_{j=1}^{m} F(DisList_{ij}^{\alpha_S \leftarrow \beta_S}, DisList_{ij}^{\alpha_T \leftarrow \beta_T})}{ml} \tag{12}$$

$$F(a, b) = -alog(b) \tag{13}$$

For the configuration with distillation on multiple transformer pairs, final L_{KD} is calculated by averaging over L_{KD} obtained from all pairs.

We employ the EDAM in two ways, named **EDAM-S**↓ and **EDAM-T**↑ to deal with diverse dimensions, due to the different inputs of the teacher and the student. The teacher has multiple modalities as input while the student has only one modality, and each modality may have a different sequence length. Thereafter the attention map in the student and teacher networks can be of different dimensions. To handle this, in **EDAM-S**↓ we downsample the Video inputs to the student, A_S and L_S using Conv1D to match them with Audio and Language input in teacher network, A_T and L_T respectively.

Instead of downsampling inputs, in **EDAM-T**↑ we upsample the attention maps in the teacher transformer layers to match with the dimensions of attention maps in the student transformer layers using linear layers.

5. Results and Discussion

Table-1 reports the results obtained with the proposed framework by varying the teacher network design and the KD levels. The table is arranged with alternating student and teacher rows. The first row shows the result for just the student network, without KD. Following that, every teacher-student row pair has the teacher network performance and the performance for student, with different distillation methods (Section-4.3).

We achieved the best result with the simplified Student Network and the Complete Teacher Network (the 5th configuration in section 4.2), through the EDAM-S↓ method on attention maps. The accuracy in this case is 44.231% which improves the one-modality case (row-1 of the Table-1: 41.296 %), performed with state-of-the-art architecture [22], by approximately 3% in accuracy and 4 points in F1-score. We further compared the number of parameters of our best performing student network (0.675 Million) to the complete teacher network (1.802 Million). The parameters are reduced of around 2.5 times, which leads to a minimized inference time of 29ms (22% less than the Complete Teacher Network). Note that, the results obtained for KD from Complete Teacher Network (row-2 in Table-1) were not optimized rigorously as the behavior of the smaller student (row-5 in Table-1) was similar to it.

Notably, the improvements, in accuracy and F1-score, occurred in the lower layers of the architecture, in the attention map. This proves our initial hypothesis, suggesting that higher layers do not provide enough information in cross-attention transformers to the student network. Moreover, by comparing CRD with EDAM methods we found out that EDAM, the lighter one, gives the best results. Therefore we were able to obtain the highest accuracy and F1-score with the less computational requirement. A challenge with applying distillation to attention maps is that the dimensions of the teacher and student are different as the student capitalizes on self attention, while the teacher on cross-attention. We explored two methodologies to overcome this issue, downsampling (EDAM-S↓) and upsampling (EDAM-T↑), showing that the former works better. This result is intuitive as the features are designed to be as orthogonal as possible and reducing the map size using a linear layer will lead to loss of information.

In this work we used unaligned modalities as input. Nevertheless, if aligned modalities are adopted instead, the downsampling method for which loss of information is linked to the reduced temporal resolution, would have shown even better performance.

Note that in Table-1 we did not optimize the hyperparameters for the CRD on attention map, as the outputs were closed to the EDAM method that do not required any additional parameters for training.

Looking at the diverse teacher configurations, the use of

Network	*Description*	*Accuracy (%)*	*F1-score*
Student	Without KD	41.296	32.142
KD from the Complete Teacher Network			
Teacher	Complete Teacher Network	49.779	48.817
Student	EDAM-S↓ on Attention Maps	44.137	36.286
KD from Video Branch			
Teacher	Video Branch	49.621	48.542
Student	CRD on final linear layer	43.711	33.281
	CRD on penultimate linear layer	42.576	35.565
	CRD on attention maps	43.837	35.912
KD from Language Branch			
Teacher	Language Branch	49.217	47.250
Student	CRD on final linear layer	43.154	32.863
	CRD on penultimate linear layer	42.901	34.467
KD from Audio Branch			
Teacher	Audio Branch	49.217	47.232
Student	CRD on final linear layer	43.584	33.260
	CRD on penultimate linear layer	43.232	35.115
KD from Language and Audio Branches			
Teacher	Complete teacher network	49.779	48.817
Student	CRD on final linear layer	42.731	32.513
	CRD on penultimate linear layer	43.732	35.632
	CRD on post-attention linear layers	43.837	35.711
	CRD on attention maps	44.031	36.110
	EDAM-S↓ on attention maps	**44.231**	**36.332**
	EDAM-T↑ on attention maps	43.993	36.189

Table 1: Results on CMU-MOSEI Dataset - Unaligned

the complete teacher network or of the individual teacher branches, does not have a significant impact on the results (see results of teacher networks in the Table-1). This can be due to the high temporal and semantic correlation among the modalities with respect to the output. On the other hand, individual branches teacher are smaller in size, therefore quicker to train.

Finally, looking at the results, we noted that in some cases (e.g. in the case of Teacher with Language Branch) accuracy decreases while the F1-score increases when applying KD in deeper layers. This can be due to the fact the data is skewed and the better accuracy reached in the higher layers is obtained by giving the majority class as output most of the times. When applying KD in the deeper layers the network better learns to take into account the other classes, so the accuracy decreases. Controversially the F1 score, which is a good metric for imbalanced data, is aligned with the expected trend, where the results improve applying KD in deeper layers.

6. Conclusions

In this work we explore KD at various levels and with multiple student-teachers configurations. We establish that going deeper in the transformer network is conducive to KD by getting best results with KD on attention maps.

In this first work we used as single modality, in the student network, the video; nevertheless, the same methodology can be adopted to other modalities. Studying the multiple teacher configurations we demonstrated that using different modalities as query, key and value does not have much impact on the results. This will help future studies, to reduce the effort in the evaluation of diverse permutations.

Since this is an unexplored field, multiple directions are envisioned as future work. First, others techniques of distillation can be explored. The EDAM-S↓ method, based on simple cross-entropy loss achieved the best accuracy and F1-score, and required fewer learning parameters compared to KD with CRD loss. This serves as a proof of concept and opens the possibility for experimenting different loss functions for KD on attention maps. The accuracy of one-modality based models can be moreover improved by exploiting other modalities, such as the physiological data. We believe that, while strong emotions can be better recognized by using one modality with distilled knowledge, subtle emotions are more complicated to be addressed and physiological cues could add relevant dependencies in the final one-modality. We believe our approach is robust and extending it to other datasets and fields of work is a pertinent next step.

We aim that this paper will pave the way for a novel research area in multimodal machine learning focusing on resources reduction. This will permit the use of complex algorithms when the inference time has to be minimized, exploiting a richer embedding space and increased accuracy. An example of application is in the field of emotion recognition, where high computational cost can not be sustained and some modalities might be partly missing.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2

[2] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 2

[3] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008. 2

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020. 2

[5] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019. 2

[6] W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung. Multimodal end-to-end sparse model for emotion recognition. *arXiv preprint arXiv:2103.09666*, 2021. 2

[7] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. COVAREP - A collaborative voice analysis repository for speech technologies. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, pages 960–964, Florence, Italy, May 2014. IEEE. 3

[8] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015. 1, 2

[9] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 1

[10] N. Kitaev, Łukasz Kaiser, and A. Levskaya. Reformer: The efficient transformer, 2020. 2

[11] L.-P. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176, 2011. 2

[12] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288, 2016. 2

[13] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[14] S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger. *The handbook of multimodal-multisensor interfaces, Volume 2: Signal processing, architectures, and detection of emotion and cognition*. Morgan & Claypool, 2018. 2

[15] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 3

[16] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, 2013. 2

[17] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*, 2020. 1

[18] F. B. Rui Dai, Srijan Das. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. accepted, available at hal-03314575,. *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2

[19] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3

[20] C. G. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications*, 25(1):5–35, 2005. 2

[21] Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation, 2020. 1, 2, 5

[22] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 2, 3, 4, 6

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. 1, 2, 4

[24] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 949–954. IEEE, 2017. 2

[25] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223, 2019. 2

[26] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71, 1989. 2

[27] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, 2017. 2

[28] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018. 2