Pivot Correlational Neural Network for Multimodal Video Categorization

Sunghun Kang[1][0000−0003−2632−7522], Junyeong Kim[1][0000−0002−7871−9627],
Hyunsoo Choi[2], Sungjin Kim[2], and Chang D. Yoo[1]

[1] KAIST, Daejeon, South Korea
{sunghun.kang, junyeong.kim, cd_yoo}@kaist.ac.kr,
[2] SAMSUNG ELECTRONICS CO.,LTD, Seoul, South Korea
{hsu.choi, sj9373.kim}@samsung.com

Abstract. This paper considers an architecture for multimodal video categorization referred to as Pivot Correlational Neural Network (Pivot CorrNN). The architecture consists of modal-specific streams dedicated exclusively to one specific modal input as well as modal-agnostic pivot stream that considers all modal inputs without distinction, and the architecture tries to refine the pivot prediction based on modal-specific predictions. The Pivot CorrNN consists of three modules: (1) maximizing pivot-correlation module that maximizes the correlation between the hidden states as well as the predictions of the modal-agnostic pivot stream and modal-specific streams in the network, (2) contextual Gated Recurrent Unit (cGRU) module which extends the capability of a generic GRU to take multimodal inputs in updating the pivot hidden-state, and (3) adaptive aggregation module that aggregates all modal-specific predictions as well as the modal-agnostic pivot predictions into one final prediction. We evaluate the Pivot CorrNN on two publicly available large-scale multimodal video categorization datasets, FCVID and YouTube-8M. From the experimental results, Pivot CorrNN achieves the best performance on the FCVID database and performance comparable to the state-of-the-art on YouTube-8M database.

Keywords: Video categorization · Multimodal representation · Sequential modeling · Deep learning

1 Introduction

Multimodal video categorization is a task for predicting the categories of a given video based on different modal inputs which may have been captured using diverse mixture of sensors and softwares in securing different modalities of the video. Fig. 1 shows four video examples from the FCVID dataset with groundtruth and top 3 scores obtained from the proposed algorithm referred to as Pivot CorrNN. Fortifying and supplementing among different modalities for more accurate overall prediction is a key technology that can drive future innovation in better understanding and recognizing the contents in a video. Emerging applications includes video surveillance, video recommendation, autonomous

driving and sports video analysis system. The use of deep Convolutional Neural Networks (CNNs) has lead to many dramatic progress across different tasks but generally confined to a single modality- often in the form of an image, speech or text- with an optional association with an auxiliary modality such as a text query. Indeed, studies leveraging on synergistic relationship across multiple modalities have been scarce so far.

Considerable studies have been dedicated to the topic of video categorization, but these have mainly been visual. Auditory modality has very often been ignored. Some notable past studies have focused on spatio-temporal visual representation. Karpathy *et al.* [19] trained a deep CNN on large video dataset while investigating the effectiveness of various temporal fusion. Tran *et al.* [29] extended conventional two dimensional convolution operation to three dimensional for considering spatio-temporal information in a video.

Other studies have focused on utilizing motion modality alongside with visual appearance modality. Donahue *et al.* [9] studied and compared the behaviors of various configurations of CNN-LSTM combination. Here, the outputs of two CNN-LSTM combination- one taking RGB image as input while the other taking flow image- are merged in making the final prediction. In the two stream networks [10,11,25], two separate CNN streams- one taking static image as input while the other taking optical flow- are considered, and intermediate features of the two streams leading up to the final prediction are fused either by the summation [10] or multiplicative operations [11].

Auditory modality has also been considered in a minor way. Jiang *et al.* [18] proposed regularized DNN (rDNN) which jointly exploits the feature (including audio features) and class relationship to model video semantics. Miech *et al* [23] considered an architecture with two learnable pooling layers- one taking visual input while the other taking audio input- that are merged by a fully connected layer and gated for final prediction.

Although considerable advances have been made in video categorization, there are still many unresolved issues to be investigated. First, it is often difficult to determine the relationship among heterogeneous modalities especially when the modalities involve different entities such it is difficult to determine the relationship between the modalities. For example, static image and its optical flow which involve a common entity- in this case, the pixels- can be easily be fused in the same spatial domain, while it is non-trvial to learn the relationship between static images and audio signals of the video. Second, multimodal sequential modeling should consider the complementary relationship between modalities with their contextual information. Information relevant for categorization vary across time due to various reasons such as occlusion and noise. It maybe more appropriate to emphasize one modality over the other. Third, depending on the category, one modality will provide far more significant information about the category than the other, and this needs to be taken into account. Most categories are defined well in the visual domain while there are categories better defined in the auditory domain. As depicted by Wang et al. [31], in most of the misclassification cases, there exists one modality that is failing while the other is correct. In
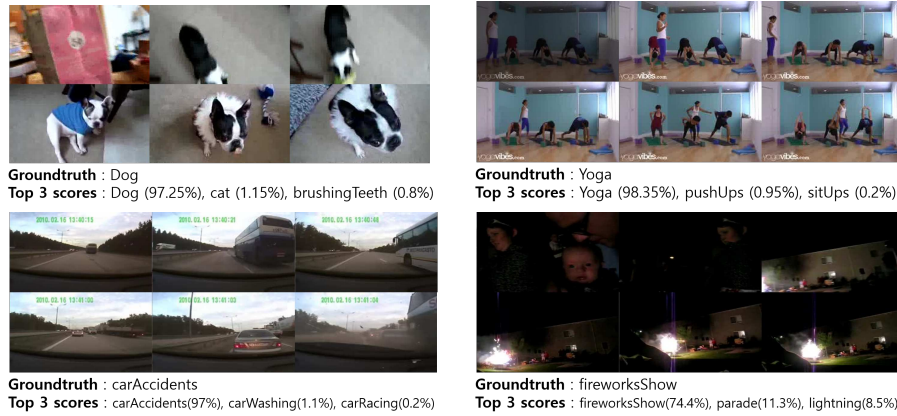
Fig. 1: Four video examples from the FCVID dataset with groundtruth and top3 scores obtained from the proposed algorithm referred to as Pivot CorrNN.

this case, it is necessary to develop a model considering the level of confidence for each modality prediction.

To overcome the above issues, this paper considers an architecture for multimodal video categorization referred to as Pivot Correlational Neural Network (Pivot CorrNN). It is trained to maximize the correlation between the hidden states as well as the predictions of the modal-agnostic pivot stream and modal-specific streams in the network, and to refine the pivot prediction based on modal-specific predictions. Here, the modal-agnostic pivot hidden state considers all modal inputs without distinction while the modal-specific hidden state is dedicated exclusively to one specific modal input. The Pivot CorrNN consists of three modules: (1) maximizing pivot-correlation module that attempts to maximally correlate the hidden states as well as the predictions of the modal-agnostic pivot stream and modal-specific streams in the network, (2) contextual Gated Recurrent Unit (cGRU) module which extends the capability of a generic GRU to take multimodal inputs in updating the pivot hidden-state, and (3) adaptive aggregation module that aggregates all modal-specific predictions as well as the modal-agnostic pivot predictions into one final prediction. The maximizing pivot correlation module that provides guidance for co-occurrence between modal-agnostic pivot and modal-specific hidden states as well as their predictions. The contextual Gated Recurrent Unit (cGRU) module which models time-varying contextual information among modalities. When making the final prediction, the adaptive aggregation module considers the confidence of each modality.

The rest of the paper is organized as follows. Section 2 reviews previous studies on video categorization and multimodal learning. Section 3 discusses proposed architecture in detail. Section 4 presents experimental results, and finally, Section 5 concludes the paper.

2   Multimodal Learning

In this section, multimodal learning is briefly reviewed. Some related works on multimodal representation learning are introduced.

Deep learning has been shown to have the capability to model multiple modalities for useful representations [3, 24, 27]. Generally speaking, the mainstream of multimodal representation learning falls into two methods: joint representation learning and coordinated representation learning. Joint representation learning. It learns a joint representation of input modalities can be obtained by concatenation, element-wise summation, and element-wise multiplication, as output of a neural network. The second method is based on similarity of separate representations. It learns a separate representation of each input modalities while each representation is constrained to be similiar with other modalities.

Research focus on the first method aims to make joint representation using various first and second order interactions between features. Ngiam *et al.* [24] propose a deep autoencoder based architecture for joint representation learning of video and audio modality. Self-reconstruction and cross-reconstruction are utilized to learn joint representation for audio-visual speech recognition. Srivastava *et al.* [27] propose a Deep Boltzmann Machine (DBM) based architecture to learn a joint density model over the space of multimodal inputs. Joint representation can be obtained even though there exist some missing modalities through Gibbs sampling. Antol *et al.* [4] propose deep neural network based architecture for VQA. The element-wise multiplication is performed to fuse image features and text features and obtain joint representation. Outer product is also used to fuse input modalities [6, 13, 20]. Since the fully parameterized bilinear model (using the outer product) becomes intractable due to the number of parameters, simplification or approximation of model complexity is needed. Fukui *et al.* [13] project outer product to lower dimensional space using count-sketch projection, Kim *et al.* [20] constrain the rank of resulting tensor and Ben-Younes *et al.* [6] utilize tucker decomposition to reduce the number of parameters while preserving the model complexity.

Research focus on the second method aims to make separate representation, and a loss function is incorporated to reduce the distance between the representations. Similarity measure such as inner product or cosine similarity can be used for coordinated representation. Weston *et al.* [32] propose WSABIE which uses inner product to measure similarity. The inner product between image feature and textual feature is calculated and maximized so that corresponding image and annotation would have a high similarity between them. Frome *et al.* [12] propose DeViSE for visual-semantic embedding. DeViSE uses a hinge ranking loss function and an inner product similar to WSABIE but utilizes deep architecture to extract the image and textual feature. Huang *et al.* [16] utilize cosine similarity to measure the similarity between query and document. The similarity is directly used to predict posterior probability among documents. Research focus on coordinated representation is based on canonical correlation analysis (CCA) [15]. The CCA is the methods that aim to learn separate representation for each modality while the correlation between them is maximized simultaneously. An-

drew *et al.* [3] propose Deep CCA (DCCA) which is a DNN extension of CCA. The DCCA learns a nonlinear projection using deep networks such that the resulting representations are highly linearly correlated with different view images. Wang *et al.* [30] propose deep canonically correlated autoencoders (DCCAE) which is a DNN-based model combining CCA and autoencoder-based terms. The DCCAE jointly optimizes autoencoder (AE) objective (reconstruction error) and canonical correlation objective. Chandar *et al.* [7] propose correlational neural networks (CorrNet) which is similar to the DCCAE in terms of jointly using reconstruction objective and correlation maximization objective. However, CorrNet only maximizes the empirical correlation within a mini-batch instead of CCA constraints maximizing canonical correlation.

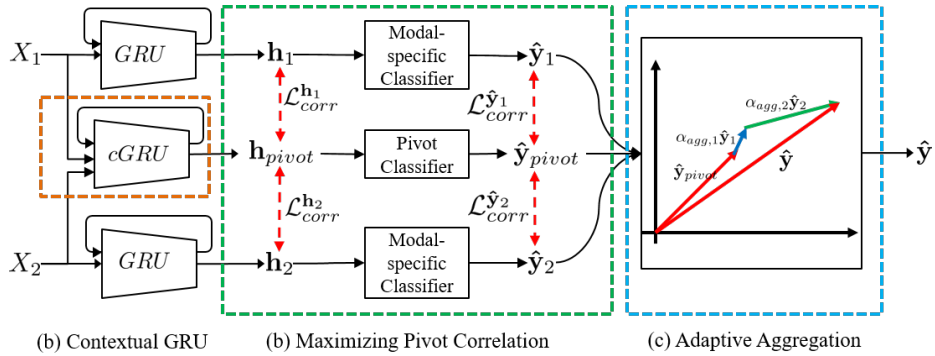## 3   Pivot Correlational Neural Network

Fig. 2: Block diagram of the proposed Pivot CorrNN in a bi-modal scenario. The Pivot CorrNN is composed of three modules: (a) Contextual Gated Recurrent Unit, (b) Maximizing Pivot Correlations, and (c) Adaptive Aggregation

In this section, the Pivot CorrNN and its modules are described. The proposed Pivot CorrNN is composed of three modules: contextual GRU (cGRU) module, maximizing pivot correlation module and adaptive aggregation module. The proposed Pivot CorrNN can be generalized for M modalities using M modal-specific GRUs and one modal-agnostic cGRU with its classifiers.

Fig. 2 shows the overall block diagram of the Pivot CorrNN illustrating the connections between modules for sequential bi-modal scenario. In the sequential bi-modal case which involves two sequential modal inputs $X_1 = \{\mathbf{x_1^t}\}_{t=1}^T$ and $X_2 = \{\mathbf{x_2^t}\}_{t=1}^T$, the Pivot CorrNN fuses the two inputs and then predicts a label $\hat{\mathbf{y}}$ corresponding to the two inputs. Two GRUs and one cGRU are utilized for obtaining two separate modal-specific hidden states ($\mathbf{h}_1$ and $\mathbf{h}_2$) and one pivot hidden state $\mathbf{h}_{pivot}$. Each hidden state is fed to its classifier for predicting corresponding labels ($\hat{\mathbf{y}}_1$, $\hat{\mathbf{y}}_2$, and $\hat{\mathbf{y}}_{pivot}$). During training proposed Pivot CorrNN,

maximizing pivot correlation module measures the correlations on both hidden state and label prediction between modal-specific and modal-agnostic pivot, and maximizes them. To produce final prediction $\hat{\mathbf{y}}$, adaptive aggregation module is involved.

The details of proposed the cGRU, maximizing pivot correlation, and adaptive aggregation modules are introduced in Section. 3.1, 3.2, and 3.3, respectively.
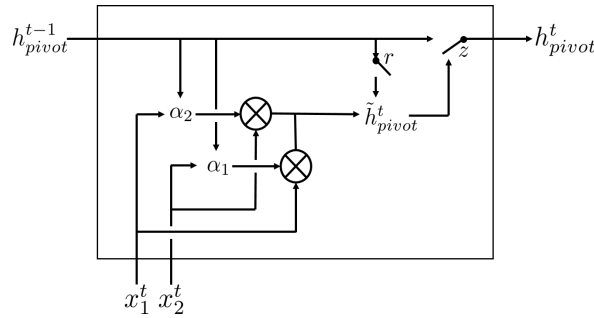
3.1   Contextual Gated Recurrent Units (cGRU)

Fig. 3: Illustration of the cGRU. Gating masks α_1, and α_2 are introduced to control contextual flow of each modality input based on previous hidden pivot state and other modality input.

The proposed contextual GRU (cGRU) is an extension of the GRU [8] that combines many modal inputs into one by concatenating the weighted inputs before the usual process of GRU takes over. The weight place on a particular modal input is determined by considering the hidden state of the cGRU and other modal inputs excluding itself.

Fig. 3 illustrates a particular cGRU taking two modal inputs $\mathbf{x}_1^t$ and $\mathbf{x}_2^t$ at time step t and updating its hidden state $\mathbf{h}_{pivot}^{t-1}$ to $\mathbf{h}_{pivot}^t$. After going through all the input sequence from $t = 1$ through $t = T$, the final modal-agnostic pivot hidden-state $\mathbf{h}_{pivot}$ is presented to the pivot classifier.

To model time-varying contextual information of each modality, two learnable sub-neural networks within cGRU are introduced. Each input modality is gated by considering the input of the other modality in the context of previous hidden pivot state $\mathbf{h}_{pivot}^{t-1}$. The gated inputs are concatenated in constructing the update gate masks as well as reset gate and the hidden pivot state. The hidden pivot state are updated in the usual GRU manner.

$$\alpha_1 = \sigma(W_{\alpha_1 h}\mathbf{h}^{t-1}_{pivot} + W_{\alpha_1 x}\mathbf{x}^t_2 + b_{\alpha_1}),$$
$$\alpha_2 = \sigma(W_{\alpha_2 h}\mathbf{h}^{t-1}_{pivot} + W_{\alpha_2 x}\mathbf{x}^t_1 + b_{\alpha_2}),$$
$$\mathbf{x}^t = [\alpha_1 \odot \mathbf{x}^t_1; \alpha_2 \odot \mathbf{x}^t_2],$$
$$\mathbf{z}^t = \sigma(W_{zh}\mathbf{h}^{t-1}_{pivot} + W_{zx}\mathbf{x}^t + b_z),$$
$$\mathbf{r}^t = \sigma(W_{rh}\mathbf{h}^{t-1}_{pivot} + W_{rx}\mathbf{x}^t + b_r),$$
$$\tilde{\mathbf{h}}^t_{pivot} = \varphi(W_{hx}\mathbf{x}^t + W_{hh}(\mathbf{r}^t \odot \mathbf{h}^{t-1}_{pivot}) + b_h),$$
$$\mathbf{h}^t_{pivot} = (1 - \mathbf{z}^t) \odot \mathbf{h}^{t-1}_{pivot} + \mathbf{z}^t \odot \tilde{\mathbf{h}}^t_{pivot},$$

where σ, φ are logistic sigmoid and hyperbolic tangent function respectively. Here, $\odot$ denotes the Hadamard product. $\mathbf{x}^t$ is the modulated input using gating masks. $\mathbf{z}^t, \mathbf{r}^t$ are the update and reset gates at time t, which are the same as original GRU. $\mathbf{h}_{pivot}$ and $\tilde{\mathbf{h}}_{pivot}$ are modal-agnostic pivot hidden state and its internal candidate hidden pivot state.

3.2   Maximizing Pivot Correlation Module

The maximizing pivot correlation module is proposed for capturing co-occurrence among modalities in both hidden states and label predictions during training. The co-occurrence expresses co-activation of neurons among modal-specific hidden states. The maximizing pivot-correlation module that attempts to maximally correlate between the hidden states as well as the predictions of the modal-agnostic pivot stream and modal-specific streams in the network The details of maximizing pivot correlation module is followed as below.

The maximizing pivot correlation in hidden states utilizes modal-specific states $\mathbf{h}_1$, and $\mathbf{h}_2$ and modal-agnostic pivot hidden state $\mathbf{h}^T_{pivot}$. The pivot correlation objective on the m-th modality hidden state $\mathcal{L}^{\mathbf{h}_m}_{corr}$ is defined as follows:

$$\mathcal{L}^{\mathbf{h}_m}_{corr} = \frac{\sum_{i=1}^N (\mathbf{h}_{m,i} - \bar{\mathbf{h}}_m)(\mathbf{h}_{pivot,i} - \bar{\mathbf{h}}_{pivot})}{\sqrt{\sum_{i=1}^N (\mathbf{h}_{m,i} - \bar{\mathbf{h}}_m)^2 \sum_{i=1}^N (\mathbf{h}_{pivot,i} - \bar{\mathbf{h}}_{pivot})^2}},$$

where the subscript i denotes the sample index. Here, $\bar{\mathbf{h}}_m = \frac{1}{N}\sum_{i=1}^N \mathbf{h}_{m,i}$ and $\bar{\mathbf{h}}_{pivot} = \frac{1}{N}\sum_{i=1}^N \mathbf{h}_{pivot,i}$ are the averages of the modal-specific and modal-agnostic hidden states, respectively. Here, $\mathbf{h}_{m,i}$ denotes the hidden state of the m-th modality of the i-th samples.

For maximizing pivot correlation objective in label predictions $\mathcal{L}^{\hat{\mathbf{y}}_m}_{corr}$ is defined as follows:

$$\mathcal{L}^{\hat{\mathbf{y}}_m}_{corr} = \frac{\sum_{i=1}^N (\hat{\mathbf{y}}_{m,i} - \bar{\mathbf{y}}_m)(\hat{\mathbf{y}}_{pivot,i} - \bar{\mathbf{y}}_{pivot})}{\sqrt{\sum_{i=1}^N (\hat{\mathbf{y}}_{m,i} - \bar{\mathbf{y}}_m)^2 \sum_{i=1}^N (\hat{\mathbf{y}}_{pivot,m} - \bar{\mathbf{y}}_{pivot})^2}},$$

where $\bar{\mathbf{y}}_m = \frac{1}{N}\sum_{i=1}^N \hat{\mathbf{y}}_{m,i}$ and $\bar{\mathbf{y}}_{pivot} = \frac{1}{N}\sum_{i=1}^N \hat{\mathbf{y}}_{pivot,i}$ denote respectively the average of the modal-specific and modal-agnostic prediction.

3.3   Adaptive Aggregation

We propose a soft-attention based late fusion algorithm referred as adaptive aggregation. The adaptive aggregation is an extension of the attention mechanism in the late fusion framework based on the confidence between modal-specific predictions and modal-agnostic pivot prediction. For M multimodal case, all the modal-specific predictions $\{\hat{\mathbf{y}}_m\}_{m=1}^{M}$ and the modal-agnostic pivot prediction $\hat{\mathbf{y}}_{pivot}$ are considered in making the final prediction $\hat{\mathbf{y}}_{agg}$ as follows:

$$\hat{\mathbf{y}}_{agg} = \sigma \left(\hat{\mathbf{y}}_{pivot} + \sum_{m=1}^{M} \alpha_{agg,m} \cdot \hat{\mathbf{y}}_m \right),$$

where $\alpha_{agg,m}$ is the scalar multimodal attention weight corresponding to the m-th modality. The multimodal attention weights are obtained using a neural network analogous to the soft-attention mechanism:

$$\alpha_{agg,m} = \frac{\exp(s_m)}{\sum_{i=1}^{M} \exp(s_i)}, \quad m = 1, \cdots, M,$$

where

$$s_m = W_s \left[\mathbf{h}_m ; \mathbf{h}_{pivot} \right] + b_s, \quad m = 1, \cdots, M.$$

Unlike widely used late fusion algorithm such as mean aggregation, the adaptive aggregation can regulate the ratio of each modality on final prediction. The learned multimodal attention weights can be viewed as the reliability of each modality. Consider a video with "surfing" label. Surfing board can be visually observed but insteads of hearing the waves we hear some music. In this case, the attention weight corresponds to visual modality label should be higher than that corresponding to audio such that final prediction is made based on visual modality rather than auditory modality.

3.4   Training

The objective loss function to train the proposed Pivot CorrNN is composed of three terms. First, $(M+2)$ cross-entropy losses are included where M denotes the number of input modalities. Additional two cross-entropy is dedicated to the pivot prediction and the prediction after the adaptive aggregation module which is responsible for the supervision in learning the confidence level of each modality prediction.

Second, M number of correlations between the hidden states as well as the predictions of each of the modal-specific and modal-agnostic subnetwork. Third, for achieving better generalization performance, ℓ^2-regularization is additionally applied. Minimizing the overall objective loss function leads to minimizing the $M+2$ classification errors, and at the same time, maximizes the pivot correlation objectives. To handle this opposite direction, the final loss function $\mathcal{L}$ is designed

to minimize cross-entropy, regularization and negative of correlations losses as below:

$$\mathcal{L} = \sum_{m=1}^{M} \left(\sum_{c=1}^{C} \mathbf{y}_c \log(\hat{\mathbf{y}}_{m,c}) + (1 - \mathbf{y}_c) \log(1 - \hat{\mathbf{y}}_{m,c}) \right)$$
$$+ \sum_{c=1}^{C} \left(\mathbf{y}_c \log(\hat{\mathbf{y}}_{pivot,c}) + (1 - \mathbf{y}_c) \log(1 - \hat{\mathbf{y}}_{pivot,c}) \right)$$
$$+ \sum_{c=1}^{C} \left(\mathbf{y}_c \log(\hat{\mathbf{y}}_{agg,c}) + (1 - \mathbf{y}_c) \log(1 - \hat{\mathbf{y}}_{agg,c}) \right)$$
$$- \lambda_1 \left(\sum_{m=1}^{M} \mathcal{L}_{corr}^{\mathbf{h}_m} + \mathcal{L}_{corr}^{\mathbf{y}_m} \right) + \lambda_2 \ell^2,$$

where, c and C indicate c-th category and the total number of categories, respectively. $\mathbf{y}_c$ is the groundtruth label for c-th category. λ_1 and λ_2 is the balancing term for controlling effectiveness of Pivot correlation and ℓ^2 regularization term.

To evaluate the pivot correlations, the entire N samples at the same time, but in practice, the empirical correlation is calculated within a single mini-batch as the same as Deep CCA [3]. Thus, the proposed maximizing pivot correlation module can be optimized using any types of gradient descent based methods including Adam [21]

4   Experiments

This section provides the experimental details of Pivot CorrNN. Initially, we describe the datasets used to train and evaluate the proposed architecture in Section. 4.1. The experimental details are described in Section. 4.2 and investigations of each proposed module are shown in Section. 4.3 as ablation study. Finally, Section. 4.4, and 4.5 show the experimental results of Pivot CorrNN for two datasets: FCVID, and YouTube-8M.

4.1   Datasets

FCVID [18] is a multi-label video categorization dataset containing 91,223 web videos manually annotated with 239 categories. The dataset represents over 4,232 hours of video with an average video duration of 167 seconds. The categories in FCVID cover a wide range of topics including objects (e.g., "car"), scenes (e.g., "beach"), social events (e.g., "tailgate party") and procedural events ("making cake"). There exist some broken videos which cannot be played, we filtered out broken videos that cannot be used for extracting features. After filtering, the remaining number of videos are 44,544 for training and 44,511 for testing. The partition of the training and testing are the same of previous paper [18]. FCVID distributes raw video and 8 different precomputed video level

features: SpectrogramSIFT, SIFT, IDT-Traj, CNN, IDT-HOG, IDT-HOF, IDT-MBH and MFCC. In this paper, 7 types of pre-extracted features (except SpectrogramSIFT) are used for evaluating proposed Pivot CorrNN. For evaluation, mean Average Precision (mAP) metric is used.

YouTube-8M [2] is the largest video categorization dataset composed of about 7 million YouTube videos. Each videos are annotated one or multiple positive labels. The number of categories are 4,716, and the averaged positive labels per videos is 3.4. The training, validation and testing split are pre-defined with 70%, 20%, and 10%, respectively. Also the dataset is released to hold competition purpose, the groundtruth labels for test split is not provided. Due to its huge size, YouTube-8M provides two types of pre-extracted feature which cover visual and auditory modalities. The visual and auditory features are extracted using pre-trained Inception-V3 [28] and VGGish [14], respectively. For measuring the quality of predictions, Global Average Precision (GAP) at top 20 is used in Kaggle competition thus the performance of test split is measured in GAP solely.

4.2   Experimental Details

The entire proposed model is implemented using Tensorflow [1] framework. All the results reported in this paper were performed with Adam optimizer [21] with a mini-batch size of 128. The hyper parameters that we used are as follows. The learning rate is set to 0.001, and exponential decay rate for the 1st and 2nd moments are set to 0.9 and 0.999, respectively. For stable gradient descent procedure in cGRU and GRU, gradient clipping is adopted with clipping norm of 1.0. For the loss functions, balancing term λ_1 for maximizing pivot correlation objective, and λ_2 for ℓ^2 regularization are set to 0.001 and 3×10^{-7}. All the experiments performed under CUDA acceleration with single NVIDIA Titan Xp (12GB of memory) GPU.

4.3   Ablation Study on FCVID

To verify the effectiveness of each module of Pivot CorrNN, we conducted ablation study on FCVID. Table. 1 presents the ablation study on FCVID. In this ablation study, two modality inputs are used: C3D [29] visual and VGGish [14] auditory features.

   The performance of baseline model (without proposed module) is shown on the first row of Table. 1. For the baseline model, C3D and VGGish features are concatenated and fed into a standard GRU instead of cGRU to produce modal-agnostic pivot hidden state. The baseline model shows 66.86% in mAP measure. Then we applied proposed modules one by one. Replacing original GRU to cGRU for modal-agnostic pivot hidden state boosts the performance about 0.7%, and achieves 67.57% in mAP measure. With maximizing pivot correlations on hidden state and prediction, the model achieves the performance of 66.68% and 68.02%, respectively. Synergistic effect is observed when maximizing correlation on both

cGRU	Max. Pivot Correlation		Adaptive Aggregation	mAP(%)
	Pivot Hidden State	Pivot Prediction		
				66.86
✓				67.57
✓	✓			67.68
✓		✓		68.02
✓	✓	✓		68.45
✓	✓	✓	✓	69.54

Table 1: Ablation study for Pivot CorrNN on FCVID. As can be seen, each module of Pivot CorrNN gracefully increases the performance with activating each module. In this study, C3D visual and VGGish auditory features are used.

pivot hidden state and prediction. Finally, with all of the proposed modules, the Pivot CorrNN shows the performance of 69.54%. The entire gain of proposed modules is about 2.7% and each of the proposed modules gracefully increases the performance.

4.4   Experimental Results on FCVID

The performances of Pivot CorrNN are shown in Table 2 for FCVID test partition. In Table 2a the performances of proposed Pivot CorrNN with previous state-of-the-art algorithms are listed. The performances of previous algorithms on FCVID were not reported their original papers except for rDNN, we referred the performance from [18]. The proposed Pivot CorrNN achieved 77.6% in mAP metric on test partition of FCVID and shows absolute mAP gain of 1.6% compared to the previous state-of-the-art results.

Model	mAP (%)
DMF [26]	72.5
DASD [17]	72.8
M-DBM [27]	74.4
SVM-MKL [22]	75.2
rDNN-F [18]	75.4
rDNN [18]	76.0
Pivot CorrNN	**77.6**

(a) Performance comparison

Feature Names	Feature Type	mAP(%)
C3D, VGGish	Frame level features	69.54
+CNN, SIFT	Appearance feature	75.33
+IDT-HOF, IDT-HOG	Motion feature	76.58
+IDT-MBH, IDT-Traj	Motion feature	77.23
+MFCC	Audio features	**77.60**

(b) Feature ablation experiments on Pivot CorrNN

Table 2: Experimental Results on test partition of FCVID. (a) shows performance comparison on Pivot CorrNN and previous algorithms, and (b) shows feature ablation results

For details of performance gains, ablation experiments on the number of modalities are conducted and shown in Table 2b. With frame level features

only, the Pivot CorrNN recorded 69.54 % mAP, and adding different types of features the performance is gracefully increased. Adding appearance, motion, and audio, 6%, 1.2%, 0.7% and 0.3% mAP gains are observed, respectively. The gains explain that there is complementary information in each feature, but there is also some redundant information.

In Table 3, the comparison for multimodal attention weights in the adaptive aggregation module is shown. In the tables, thirteen categories which are selected by descending order for visual attention weight $\alpha_{agg,1}$, and auditory attention weight $\alpha_{agg,2}$. In Table 3a, all the categories are related to actions or objects. In videos belong to those categories, there is limited information in auditory modalities to describe its context from auditory information that most of the predictions are based on the visual modalities. On the other hands, all the categories listed in Table 3b are related musical activities. Visual modality does not provide much information related to its categories, but auditory modality does.

Category	$\alpha_{agg,1}$	$\alpha_{agg,2}$
taekwondo	0.981	0.019
rafting	0.958	0.042
surfing	0.94	0.06
kiteSurfing	0.937	0.063
swimmingProfessional	0.915	0.085
egyptianPyramids	0.901	0.099
horseRiding	0.895	0.105
bikeTricks	0.88	0.12
rhythmicGymnastics	0.867	0.133
mountain	0.863	0.137
VolcanoEruption	0.858	0.142
walkingWithDog	0.852	0.148
playingFrisbeeWithDog	0.846	0.154

(a) Ordered by visual modality

Category	$\alpha_{agg,1}$	$\alpha_{agg,2}$
flutePerformance	0.091	0.909
pianoPerformance	0.126	0.874
trumpetPerformance	0.179	0.821
harmonicaPerformance	0.186	0.814
singingInKtv	0.205	0.795
celloPerformance	0.216	0.784
accordionPerformance	0.239	0.761
chorus	0.309	0.691
saxophonePerformance	0.315	0.685
beatbox	0.377	0.623
publicSpeech	0.413	0.587
violinPerformance	0.415	0.585
guitarPerformance	0.42	0.58

(b) Ordered by auditory modality

Table 3: Averaged attention weights of top thirteen categories in descending order for each modality

Fig. 4 shows the qualitative results of Pivot CorrNN. For each video sample, four still frames are extracted. The corresponding groundtruth category and the top three predictions of both pivot stream and adaptive aggregation are presented. The first two videos are sampled from the categories from Table 3a, and the remaining two videos are sampled from the categories from Table 3b. The correct predictions are colored red with its probabilistic scores. The rightmost bar graphs denote the multimodal attention weights of adaptive aggregation module. In this experiments $\alpha_{agg,1}$, and $\alpha_{agg,2}$ are dedicated to visual and auditory feature, respectively.

Groundtruth : surfing
Pivot prediction: **surfing (99.10%)**, beach (3.36%), sunset (2.85%)
Adaptive Agg. : **surfing (99.27%)**, beach (2.36%), sunset (0.85%)

Groundtruth : horseRiding
Pivot prediction : **horseRiding (96.24%)**, camel (3.51%), elephant (0.86%)
Adaptive Agg. : **horseRiding (98.23%)**, camel (4.16%), cow (1.86%)

Groundtruth : celloPerformance
Pivot prediction: symphonyOrchestra (55.0%), **celloPerformance (37.8 %)**, chamberMusic (0.82%)
Adaptive Agg. : **celloPerformance (95.21%)**, trumpetPerformance (3.79%), violinPerformance (1.52%)

Groundtruth : violinPerformance
Pivot prediction : **violinPerformance (96.8%)**, chamberMusic (2.50%), celloPerformance (2.34%)
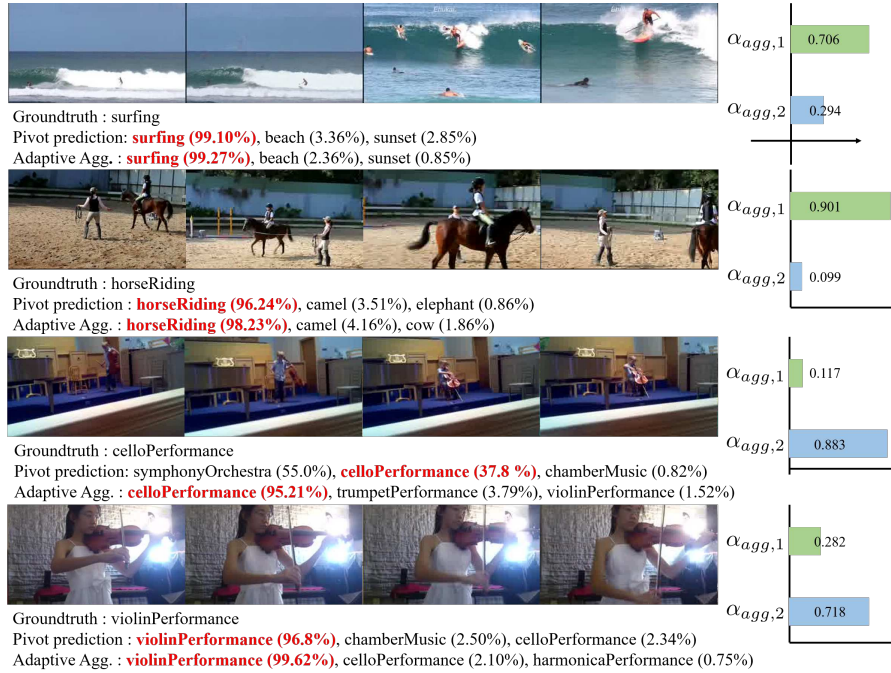Adaptive Agg. : **violinPerformance (99.62%)**, celloPerformance (2.10%), harmonicaPerformance (0.75%)

Fig. 4: Qualitative results of Pivot CorrNN. We show the groundtruth category of each video sample with top three pivot and final predictions of proposed Pivot CorrNN. The multimodal attention weights in adaptive aggregation are illustrated on the rightmost side.

Experimental results in Fig. 4 shows that the module reduces false positive errors effectively for above examples. The predictions of sampled videos are finetuned by increasing the probability of the correct predictions, and decreasing false positive predictions. Visual modality is considered more informative than auditory modality in "surfing" and "horseRiding" categories relatively two and ten times, while auditory modality is considered more informative in "celloPerformance" and "violinPerformance" categories. For sampled video which groundtruth category is "celloPerformance", the pivot prediction was 37.8% on "celloPerformance", on the other hands "symphonyOrchestraFrom" has more confidence. However, adaptive aggregation module finetuned the probability of correct category "celloPerformance" to 95.21%. From these results, adaptive aggregation module measures which modality prediction is more reliable, then it refines the final prediction with both pivot and modal-specific predictions.

4.5 Experimental Results on YouTube-8M

For evaluating proposed Pivot CorrNN on YouTube-8M dataset, two types of experiments are conducted from both video and frame level features. For the

video level features, all the frame level features from each video are averaged into a single feature vector. There is no sequential information in the video level features that cGRU is not applied for experiments of video level features. For the frame level features, all the three modules are applied for Pivot CorrNN.

Feature Level	Model	GAP (%)
Video	Logistic Regression (Concat)	76.79
Video	Pivot CorrNN (without cGRU)	77.40
Frame	Two-layer LSTM (Concat)	80.11
Frame	Pivot CorrNN (with cGRU)	81.61

Table 4: Multimodal video categorization performance of two baseline models and Pivot CorrNNs on YouTube-8M dataset

The performance comparision of Pivot CorrNN with baseline models are presented in Table 4. Logistic regressions are used for all the classifiers within the models. The performance gains are observed for the proposed Pivot CorrNN 0.7% and 1.5% in GAP metric, respectively. In these experiments, pre-extracted Inception-V3 and VGGish features are used without any additional feature encoding algorithms, such as learnable pooling methods [23], NetVLAD [5], etc. With advanced feature encoding algorithms as an additional feature, we believe proposed Pivot CorrNN will achieve better performance on YouTube-8M.

5   Conclusion

This paper considers a Pivot Correlational Neural Network (Pivot CorrNN) for multimodal video categorization by maximizing the correlation between the hidden states as well as the predictions of the modal-agnostic pivot stream and modal-specific streams in the network. The Pivot CorrNN consists of three modules: (1) maximizing pivot-correlation module that maximizes the correlation between the hidden states as well as the predictions of the modal-agnostic pivot stream and modal-specific streams in the network, (2) contextual Gated Recurrent Unit (cGRU) module that models time-varying contextual information among modalities, and (3) adaptive aggregation module that considers the confidence of each modality before making one final prediction. We evaluate the Pivot CorrNN on two publicly available large-scale multimodal video categorization dataset: FCVID, and YouTube-8M. From the experimental results, Pivot CorrNN achieves best performance on the FCVID database and the performance comparable to the state-of-the-art on YouTube-8M database.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI. vol. 16, pp. 265–283 (2016)
2. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
3. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning. pp. 1247–1255 (2013)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
5. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
6. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proc. IEEE Int. Conf. Comp. Vis. vol. 3 (2017)
7. Chandar, S., Khapra, M.M., Larochelle, H., Ravindran, B.: Correlational neural networks. Neural computation 28(2), 257–285 (2016)
8. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
9. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
10. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: Advances in neural information processing systems. pp. 3468–3476 (2016)
11. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7445–7454. IEEE (2017)
12. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems. pp. 2121–2129 (2013)
13. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
14. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. pp. 131–135. IEEE (2017)
15. Hotelling, H.: Relations between two sets of variates. Biometrika 28(3/4), 321–377 (1936)
16. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. pp. 2333–2338. ACM (2013)

17. Jiang, Y.G., Dai, Q., Wang, J., Ngo, C.W., Xue, X., Chang, S.F.: Fast semantic diffusion for large-scale context-based image and video annotation. IEEE Transactions on Image Processing **21**(6), 3080–3091 (2012)
18. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE transactions on pattern analysis and machine intelligence **40**(2), 352–364 (2018)
19. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
20. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 (2016)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: Lp-norm multiple kernel learning. Journal of Machine Learning Research **12**(Mar), 953–997 (2011)
23. Miech, A., Laptev, I., Sivic, J.: Learnable pooling with context gating for video classification. arXiv preprint arXiv:1706.06905 (2017)
24. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11). pp. 689–696 (2011)
25. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
26. Smith, J.R., Naphade, M., Natsev, A.: Multimedia semantic indexing using model vectors. In: Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on. vol. 2, pp. II–445. IEEE (2003)
27. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Advances in neural information processing systems. pp. 2222–2230 (2012)
28. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
29. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
30. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: International Conference on Machine Learning. pp. 1083–1092 (2015)
31. Wang, Y., Long, M., Wang, J., Philip, S.Y.: Spatiotemporal pyramid network for video action recognition. In: CVPR. vol. 6, p. 7 (2017)
32. Weston, J., Bengio, S., Usunier, N.: Wsabie: Scaling up to large vocabulary image annotation. In: IJCAI. vol. 11, pp. 2764–2770 (2011)