Hallucinating Optical Flow Features for Video Classification

Yongyi Tang , **Lin Ma**[*] and **Lianqiang Zhou**
Tencent AI Lab
{yongyi.tang92, forest.linma}@gmail.com, tomcatzhou@tencent.com

Abstract

Appearance and motion are two key components to depict and characterize the video content. Currently, the two-stream models have achieved state-of-the-art performances on video classification. However, extracting motion information, specifically in the form of optical flow features, is extremely computationally expensive, especially for large-scale video classification. In this paper, we propose a motion hallucination network, namely MoNet, to imagine the optical flow features from the appearance features, with no reliance on the optical flow computation. Specifically, MoNet models the temporal relationships of the appearance features and exploits the contextual relationships of the optical flow features with concurrent connections. Extensive experimental results demonstrate that the proposed MoNet can effectively and efficiently hallucinate the optical flow features, which together with the appearance features consistently improve the video classification performances. Moreover, MoNet can help cutting down almost a half of computational and data-storage burdens for the two-stream video classification. Our code is available at: https://github.com/YongyiTang92/MoNet-Features.

1 Introduction

As a fundamental problem of video analysis, video classification provides discriminative information of the video content, which can help video proposal [Liu *et al.*, 2019], captioning [Wang *et al.*, 2018], grounding [Chen *et al.*, 2018a], and so on. However, the video sequence contains rich motion information, such as the object movements and temporal correlations between different events, making the video classification much more challenging compared with image classification.

Recently, the two-stream models [Simonyan and Zisserman, 2014; Carreira and Zisserman, 2017; Gao *et al.*, 2018] simultaneously encode the appearance and motion information and achieve the state-of-the-art performances on video classification. As shown in Figure 1, the motion information is encoded by the optical flow stream, which complements the appearance stream for video classification. However, the optical flow acquisition is extremely computationally expensive, and thereby introduces high latency for the video applications. For example, even with GPUs, FlowNet 2.0 [Ilg *et al.*, 2017] takes about 3 times of the video length to estimate its corresponding optical flow. Meanwhile the optical flow occupies the similar storage space to the RGB images. Although extensive researches have been proposed for balancing the speed and accuracy [Dosovitskiy *et al.*, 2015; Ilg *et al.*, 2017], optical flow is still computation-intensive and storage-intensive, especially for the large-scale video classification [Abu-El-Haija *et al.*, 2016; Tang *et al.*, 2018].

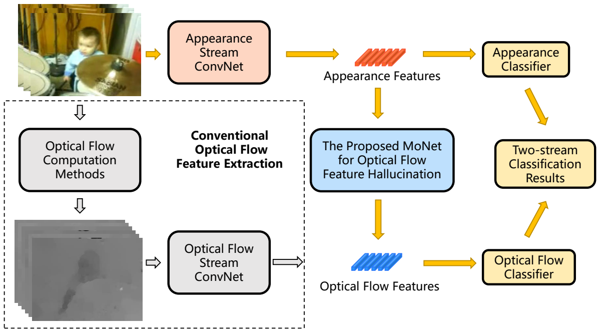

Appearance Stream ConvNet
Appearance Features
Appearance Classifier
Optical Flow Computation Methods
Conventional Optical Flow Feature Extraction
The Proposed MoNet for Optical Flow Feature Hallucination
Two-stream Classification Results
Optical Flow Stream ConvNet
Optical Flow Features
Optical Flow Classifier

Figure 1: The traditional two-stream network relies on the computationally expensive methods, such as TV-L1 methods or FlowNets, to estimate the optical flow images and the ConvNet to extract the optical flow features. On the contrary, we propose one novel MoNet to hallucinate the optical flow features from the appearance features with no reliance on the computation-intensive and storage-intensive procedures. The final classification results are yielded by considering both the appearance stream and the hallucinated optical flow stream.

In order to ease the burdens of computing optical flow, the ConvNets have been used to hallucinate optical flow images from videos [Zhu *et al.*, 2017] or images [Gao *et al.*, 2018]. An accumulated motion vectors [Wu *et al.*, 2018] are used in stead of the optical flow for the compressed video classification. However, these methods mainly focus on easing the optical flow computation, while extracting optical features

[*] Contact Author

is still of high computational budge. Instead, to further relieve the problems, we substitute both optical flow estimation and feature extraction processes with a feature hallucination network that imagines the optical flow features from the appearance features, as shown in Figure 1. Such hallucination process bypasses the resource-intensive procedures in optical flow estimation and feature extraction, and can thereby benefit the large-scale video classification.

Since optical flow features are highly related to the corresponding appearance features, we can formulate the feature hallucination as one sequence-to-sequence translation problem. Recurrent neural networks (RNNs), yielding encouraging results on sequence modeling, such as machine translation [Hochreiter and Schmidhuber, 1997; Cho *et al.*, 2014], captioning [Wang *et al.*, 2018; Jiang *et al.*, 2018; Chen *et al.*, 2018b], and video classification [Donahue *et al.*, 2015], are naturally suitable for such translation problem. However, the optical flow features concurrently relate to each other within the local context region, especially for the optical flow features encoded by the 3D-ConvNets. The traditional RNNs can only model the feature relationships of one temporal direction at each time. And even the bidirectional RNN [Schuster and Paliwal, 1997] can only capture bidirectional information asynchronously with two separated RNNs,which cannot effectively model the complicated translations between appearance and optical flow features (as illustrated in Section 4.1).

In this paper, we propose a motion hallucination network (MoNet) that imagines the optical flow features from the appearance ones. Unlike traditional RNNs, the proposed MoNet models the temporal relationships of the appearance features and exploits the contextual relationships of the optical flow features with concurrent connections. As such, MoNet helps exploiting the temporal relationships between appearance features and propagating contextual information within local context regions for optical flow feature hallucination. The hallucinated optical flow features, as the complementary information to the appearance features, brings consistent performance improvements for the two-stream video classification. Moreover, with bypassing the optical flow estimation and optical flow feature extraction with ConvNets, the computational and data-storage burdens can be significantly eased.

To summarize, the contributions of this paper are listed in the following. First, we propose to hallucinate optical flow features from the video appearance features for two-stream video classification. It gets rid of the computationally expensive optical flow estimation and feature extraction procedures. Second, we propose a motion hallucination network (MoNet) that models the temporal relationships of the appearance features and exploits the contextual relationships of the optical flow features with concurrent connections, which helps propagating contextual information within local context regions for optical flow feature hallucination. Finally, by hallucinating optical flow features, our MoNet can be deployed for efficient two-stream video classifications with consistent improved performances.

2 Related Works

2.1 Optical Flow for Video Classification

Optical flow is commonly used to describe motion pattern in videos, which can be represented as gray scale images representing motion magnitude along horizontal and vertical directions. Several well-known non-parametric optical flow estimation methods have been proposed for accurately estimating optical flow images, including the Brox method [Brox *et al.*, 2004] and the TV-L1 method [Zach *et al.*, 2007]. Recently, ConvNets based optical flow estimation methods, such as the FlowNet [Dosovitskiy *et al.*, 2015] and the FlowNet 2.0 [Ilg *et al.*, 2017] have been proposed, which take the advantages of GPUs for computational accelerations. However, dense optical flow estimation requires intensive computations at each pixel for every video frame as well as the corresponding large storage spaces. After optical flow estimation, deep ConvNets are used to encode optical flow features for two-stream video classification [Simonyan and Zisserman, 2014; Zhu *et al.*, 2017; Carreira and Zisserman, 2017; Peng and Schmid, 2016], which are also of high computational cost. In this paper, we try to hallucinate optical flow features from appearance features. Thus the optical flow estimation and the corresponding feature extraction will be bypassed, which can thereby significantly cut down the computational cost.

2.2 Sequence Modeling

The recurrent neural networks (RNNs) can successfully model those sequences such as text [Cho *et al.*, 2014], speech [Graves and Jaitly, 2014], and video sequences [Chen *et al.*, 2019; Feng *et al.*, 2018; Donahue *et al.*, 2015; Feng *et al.*, 2019]. Vanilla RNNs use the hidden states to process sequential information, which suffer from gradient vanish on long-term sequence modeling. With the gating mechanism that controls the information, the long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] and the gated recurrent unit (GRU) [Cho *et al.*, 2014] successfully ease the problem.Realizing as RNNs, conditional random fields succeed in modeling structural relationships for structured prediction [Zheng *et al.*, 2015]. However, the RNN based models are formulated as causal models that only consider the past and current inputs, missing the ability of modeling future information and backward input The bidirectional RNNs [Schuster and Paliwal, 1997] ease the problem with two independent RNNs modeling the relationships of two temporal directions asynchronously.

3 Methodology

We formulate the optical flow feature hallucination problem as one sequence-to-sequence problem. Formally, we denote $\mathbf{X} = \{x_1, ..., x_t, ..., x_T\}$ and $\mathbf{S} = \{s_1, ..., s_t, ..., s_T\}$ as the length-T sequences of the appearance and optical flow features, respectively. The goal of optical flow feature hallucination aims to learning a mapping function $f : \mathbf{X} \rightarrow \mathbf{S}$ to make the hallucinated optical flow features as close as possible to the ground-truth ones $\hat{\mathbf{S}} = \{\hat{s}_1, ..., \hat{s}_t, ..., \hat{s}_T\}$, which are extracted from the optical flow images with the ConvNet.

For hallucinating the optical flow features from the appearance ones, we propose a motion hallucination network

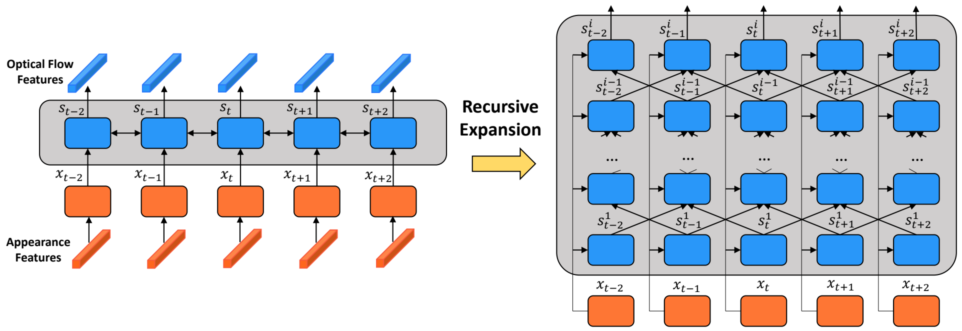

Optical Flow Features
Appearance Features
s_{t-2} s_{t-1} s_t s_{t+1} s_{t+2}
x_{t-2} x_{t-1} x_t x_{t+1} x_{t+2}
Recursive Expansion
s^i_{t-2} s^i_{t-1} s^i_t s^i_{t+1} s^i_{t+2}
s^{i-1}_{t-2} s^{i-1}_{t-1} s^{i-1}_t s^{i-1}_{t+1} s^{i-1}_{t+2}
s^1_{t-2} s^1_{t-1} s^1_t s^1_{t+1} s^1_{t+2}

Figure 2: Left: The proposed MoNet architecture. The MoNet models temporal relationships of the appearance features and exploits contextual relationships of the optical flow features at the same time. Right: Recursive expansion of the MoNet. We expand the MoNet in a recursive manner, such that the optical flow feature s^i_t at the t-th time step of the i-th layer is yielded by concurrently modeling the appearance feature x_t and adjacent optical flow features s^{i-1}_{t-1} and s^{i-1}_{t+1} of the previous layer.

(MoNet) that models the temporal relationships of the appearance features and exploits the contextual relationships of the optical flow features with concurrent connections. The hallucinated optical flow features further cooperate with the appearance features for the two-stream video classification. As such, it can reduces the computational and data-storage budgets for the optical flow estimation and feature extraction.

In this section, we first review the background of the RNN for the sequence-to-sequence translation problem. Then, we illustrate how the proposed MoNet takes the advantages of sequence-to-sequence translation for the optical flow feature hallucination. Moreover, we discuss the relations between our proposed MoNet with the existing models, specifically the GRU and ConvNet. Finally, the two-stream video classification with hallucinated optical flow features is introduced.

3.1 Background

RNN is naturally suitable for the sequence-to-sequence translation problem. For the optical flow feature hallucination problem, the vanilla RNN takes the estimated optical flow feature s_{t-1} at the previous time step and the appearance feature x_t to hallucinate the optical flow feature s_t at the t-th time step:

$$s_t = \mathrm{RNN}(x_t, s_{t-1}). \tag{1}$$

However, missing the control of dependencies with internal memories, RNN suffers from long term dependency modeling and temporal correlation modeling. The long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] and the gated recurrent unit (GRU) [Cho *et al.*, 2014] ease such problems by introducing the gating mechanism in the internal memories of RNN and have achieved great successes in the sequence-to-sequence translation tasks.

Comparing with the LSTM, the GRU, with a more concise architecture, can be formulated as follows to hallucinate the optical flow features:

$$r_t = \sigma(\mathbf{W}_r x_t + \mathbf{U}_r s_{t-1}), \tag{2}$$

$$z_t = \sigma(\mathbf{W}_z x_t + \mathbf{U}_z s_{t-1}), \tag{3}$$

$$h_t = \phi(\mathbf{W}_h x_t + \mathbf{U}_h(r_t \circ s_{t-1})), \tag{4}$$

$$s_t = z_t \circ s_{t-1} + (1 - z_t) \circ h_t, \tag{5}$$

where r_t, z_t, h_t, s_t, σ, and ϕ are the reset gate, update gate, cell hidden state, cell output, sigmoid activation, and Tanh activation, respectively. '$\circ$' denotes the element-wise multiplication. $\mathbf{W}$ and $\mathbf{U}$ are the learnable parameters, while we omit the biases for simplicity. With the gating mechanism, the GRU is able to capture the long-term dependencies of the feature sequence. However, the GRU can only model the feature relationships of one temporal direction at each time. Even with a bidirectional architecture, it only models the feature relationships of two temporal directions asynchronously, which lacks the abilities of exploiting concurrent contextual feature relationships within local regions. As such, the performances of the hallucination as well as the two-stream video classification cannot be ensured, which will be illustrated in the following experiment section.

3.2 Motion Hallucination Network (MoNet)

To better hallucinate optical flow features, temporal relationships of appearance features and contextual relationships of optical flow features should be concurrently modeled. We propose the MoNet, as shown in Figure 2, that can concurrently exploit the relationships between the appearance and optical flow features within the local contextual region. Specifically, the proposed MoNet consists of the hidden state h_t, the update gates r, and the output gates z:

$$r_{t,f} = \sigma(\mathbf{W}_r x_t + \mathbf{U}_{r,f} s_{t-1}), \tag{6}$$

$$r_{t,b} = \sigma(\mathbf{W}_r x_t + \mathbf{U}_{r,b} s_{t+1}), \tag{7}$$

$$z_{t,f} = \sigma(\mathbf{W}_z x_t + \mathbf{U}_{z,f} s_{t-1}), \tag{8}$$

$$z_{t,b} = \sigma(\mathbf{W}_z x_t + \mathbf{U}_{z,b} s_{t+1}), \tag{9}$$

$$h_t = \psi(\mathbf{W}_h x_t + \mathbf{U}_h[s_{t+1} \circ r_{t,b}, s_{t-1} \circ r_{t,f}]^\top), \tag{10}$$

where ψ is the ReLU activation function. The subscripts b, f denote backward and forward with respect to time t, respectively. Finally, the hallucinated optical flow feature of current time step s_t is inferred from the hidden state h_t and its neighboring optical flow features s_{t-1} and s_{t+1}:

$$s_t = \tilde{z}_t \circ h_t + \tilde{z}_{t,b} \circ s_{t+1} + \tilde{z}_{t,f} \circ s_{t-1}, \tag{11}$$

where $[\tilde{z}_t, \tilde{z}_{t,b}, \tilde{z}_{t,f}] = \mathrm{softmax}([1, z_{t,b}, z_{t,f}])$.

With such designed architecture, the proposed MoNet is able to hallucinate optical flow features with flexible temporal dependencies by controlling information from both directions with the designed gates. For those optical flow features lying around scene boundaries in videos, they present low correlations and high variances with respect to their neighbors. The reset gates $r_{t,b}$ and $r_{t,f}$ can thereby suppress the irrelevant information and update the hidden state with the corresponding appearance feature x_t. In addition, the output gate z is used to control the information flows from the neighbors to further refine the hallucinated optical flow features. As such, for hallucinating the optical flow feature at each time step, the proposed MoNet exploits the contextual feature relationships with respect to the corresponding temporal relationships of appearance features.

As aforementioned, the hallucination of current optical flow feature s_t relies on the contextual optical flow features s_{t-1} and s_{t+1} at the same time. The conventional RNN is intractable for this problem since they only consider information along one temporal direction at each time. The MoNet is proposed to handle such problem. To ease the implementation, we unfold the connections between adjacent units and construct layer-wise information propagation, as shown in Figure 2. The recursive expansion of the MoNet is similar to the neural message passing schedule [Gilmer *et al.*, 2017]. Moreover, such recursive expansion enables each MoNet unit to access the adjacent optical flow features s_{t-1}^{i-1} and s_{t+1}^{i-1} at the $(i-1)$-th layer and the appearance feature x_t concurrently for hallucinating s_t^i at the i-th layer:

$$s_t^i = \text{MoNet}(x_t, s_{t+1}^{i-1}, s_{t-1}^{i-1}). \tag{12}$$

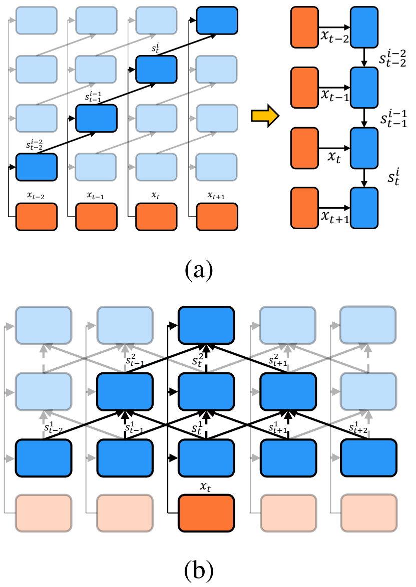
Each MoNet unit as in Eq. (12) is realized by the Eqs. (6)-(11).

On one hand, with such recursive expansion, the hallucinated optical flow features can be refined layer by layer, resulting in more consistent feature patterns. From a message-propagation point of view, the recursive expansion helps the model to capture contextual information from the adjacent optical flow features and the input appearance features. On the other hand, with the recursive expansion, the MoNet can be realized in a more efficient way by parallel matrix multiplications.

Relations with Existing Models

Here, we discuss the relations between the proposed MoNet and the existing models, specifically the RNN and ConvNet. Inheriting the gating mechanism from the GRU [Cho *et al.*, 2014], the MoNet can be regarded as a generalization of the GRU, as shown in Figure 3 (a). By only considering the connecting path where $i = t$ and dropping the backward connections, the MoNet degenerates to the GRU model resulting $s_t = GRU(x_t, s_{t-1})$. Comparing with GRU, the MoNet takes the hallucinated optical flow feature s_{t+1} into account at each time step, which makes the model concurrently exploit feature relationships within local contextual regions and thereby becomes more expressive for feature hallucination.

Compared with RNN, the ConvNet is able to simultaneously model feature relations within local regions, as shown in Figure 3 (b). However, the vanilla ConvNet cannot control the dependencies between the input appearance features and the adjacent optical flow features. Instead, the proposed MoNet adopts the gating mechanism inheriting from the GRU. With the gating mechanism and shared parameters, the proposed MoNet can more comprehensively exploit the feature relationships and control the dependencies of both the appearance features and the optical flow features.



(a)

(b)

Figure 3: Illustration of the proposed MoNet from the generalization point of view. (a): The path where $i = t$ in MoNet is equivalent to the GRU. (b): The MoNet shares one similar architecture with ConvNets by simultaneously modeling local region features.

3.3 Two-Stream Video Classification

With the optical flow features hallucinated by the MoNet, the two-stream video classification can be developed, as illustrated in the Figure 1, which only relies on the video frame information. For the appearance and optical flow classifiers, the linear classifier, the temporal relation networks (TRN) [Zhou *et al.*, 2018], the NetVLAD [Miech *et al.*, 2017], and the NeXtVLAD [Lin *et al.*, 2018] can be used, resulting the predictions of the appearance stream $p(\mathbf{X})$ and the optical flow stream $p(\mathbf{S})$. By ensembling these two predictions together, the results of the hallucinated two-stream video classification are obtained.

3.4 Training

The goal of the proposed MoNet is to hallucinate optical flow features as close as the ground-truth ones, which thereby help boosting the video classification performances. Thus, we consider not only the similarity of the features but also the corresponding classification results to constitute the objective function for training our proposed MoNet.

To hallucinate optical flow features that approximate the ground-truth ones, we minimize the mean square error at each time step. For enabling the hallucinated optical flow features for video classification, we minimize the L1 distance of the classification probabilities $p(\mathbf{S})$ and $p(\hat{\mathbf{S}})$ yielded by the sequences of hallucinated optical flow feature and ground-truth

Models	Classification Accuracy
MLP	48.21%
GRU	55.92%
LSTM	55.57%
Bi-direction GRU	55.14%
Bi-direction LSTM	56.03%
IndRNN	27.98%
1D-ConvNet	62.45%
GRU-2layer	56.16%
LSTM-2layer	60.30%
MoNet-5layers	**63.42**%

Table 1: Top-1 accuracy on Kinetics-400 validation set with optical flow features hallucinated by different sequence-to-sequence models.

Models	Classification Accuracy
I3D OF-stream	58.72% / 63.40%*
MoNet-2layers	62.58%
MoNet-3layers	62.94%
MoNet-4layers	63.26%
MoNet-5layers	**63.42**%

Table 2: Top-1 accuracy of the ground-truth I3D optical flow features and the hallucinated optical flow features. * indicates the result of the Kinetics-400 test set.

one, respectively. As such, the objective function for training the MoNet is defined as follows:

$$\begin{aligned}\mathcal{L}_{dis}(\mathbf{S},\hat{\mathbf{S}}) =& \frac{1}{N \times D \times T} \sum_{N} \sum_{T} ||s_t - \hat{s}_t||_2^2 \\ &+ \alpha \times \frac{1}{N \times C} \sum_{N} |p(\mathbf{S}) - p(\hat{\mathbf{S}})|,\end{aligned} \tag{13}$$

where C, D and N denote the number of classes, the dimension of features, and the number of examples, respectively. To balance the contributions of the terms, we empirically set α as 10.0.

4 Experimental Results and Discussions

4.1 Hallucinated Optical Flow Feature for Action Recognition

In this paper, we propose the MoNet for hallucinating optical flow features that encode motion information, which are expected to present similar classifying abilities as the ground-truth ones. Therefore, We first examine the effectiveness of the hallucinated optical flow features by different neural networks on the action recognition task.

Implementation Details

We utilize the Kinetics-400 action recognition dataset [Kay *et al.*, 2017] following the public validation split. It contains over 300 thousand of 10-second video clips with 400 different human action labels in total. We sample the video at 25 frames per second, and estimate the optical flow images by the TV-L1 method [Zach *et al.*, 2007]. After pretrained and finetuned on the ImageNet and the Kinetics-400, respectively, we took the last pooling features of the Inflated 3D ConvNet (I3D) [Carreira and Zisserman, 2017] as the features for both appearance and optical flow streams.

Classifiers	Appearance	Two-Stream	Hallucinated Two-Stream
Linear	71.32%	74.24%	72.12%
TRN	66.11%	70.49%	67.60%
NetVLAD	69.68%	74.51%	71.78%
NeXtVLAD	69.25%	74.51%	72.40%

Table 3: Top-1 accuracy of the hallucinated two-stream classification with different classifiers on the Kinetics-400 validation set.

For training the MoNet, we use the finetuned classifiers on the ground-truth features to produce the classification results of hallucinated optical flow features, $p(\mathbf{S})$. We empirically set the learning rate to $2e^{-4}$ and decrease it by $1/10$ every 15 epochs with gradient norms clipping to 1.0. We cease the training while the validation accuracy saturates at around 40 epochs. In addition, expanding more than 5 layers for the MoNet did not introduce further improvements. One reason may be attributed to that deeper architectures suffer from gradient vanish making network hard to converge.

Evaluation and Discussion

We first compare the classification performances of the hallucinated optical flow features using different sequence-to-sequence translation models as shown in Table 1. These results are evaluated on the Kinetics-400 validation set.

The hallucinated optical flow features by the MoNet achieves 63.42% accuracy with 5-layers expansions, outperforming over 3% against the RNN models, including the recent IndRNN [Li *et al.*, 2018], GRU, LSTM, and the corresponding bidirectional variants. The bidirectional variants of RNNs with doubled parameters achieve similar classification performance, which means that independent bidirectional modeling is not beneficial to the optical flow feature hallucination. The proposed MoNet models temporal relationships of the appearance features and the contextual relationships of the optical flow features concurrently, which thereby benefits the feature hallucination and video classification. Also, the proposed MoNet surpasses the 1D-ConvNet about 1% ,which illustrates the importance of the gating mechanism in our proposed MoNet, which can help suppressing irrelevant information.

We then compare the hallucinated optical flow features with the ground-truth I3D optical flow features in Table 2. The optical flow stream of the I3D model achieves 58.72% and 63.40% accuracy on the Kinetics-400 validation set and test set, respectively. The hallucinated optical flow features consistently outperform the I3D OF-features over 4.8% on the validation set. One reason is that the ground-truth optical flow stream only models motion information from optical flow images, which may not achieve the best performances. Another reason is that the proposed MoNet can effectively hallucinate optical flow features from the appearance ones, which yield even better classification performances. Moreover, with recursive layer increases, the classification performances are consistently improved, which further demonstrate the effectiveness of designed recursive expansion strategy.

	Appearance Stream				Hallucinated Two-Stream			
Classifiers	Hit@1	PERR	GAP@20	MAP@20	Hit@1	PERR	GAP@20	MAP@20
Linear	80.86%	**70.20%**	67.41%	28.11%	**81.08%**	**70.20%**	**67.64%**	**29.24%**
NeXtVLAD	87.98%	79.45%	78.55%	44.97%	**88.62%**	**80.41%**	**79.39%**	**46.70%**

Table 4: Evaluation of the hallucinated two-stream classification on the YouTube-8M validation set. Please note that the optical flow features are hallucinated from the Inception appearance features.

4.2 Two-Stream Video Classification with Hallucinated Optical Flow Features

Two-stream networks take the advantages of both the appearance and motion, thus yielding superior performances on video classification. Hallucinating optical flow features with the MoNet avoids high computational cost and latency for the optical flow stream. Thus, it is believed to be able to efficiently and effectively complement the appearance feature and benefit two-stream video classification. We further finetune the classifier of the hallucinated optical flow features, and fuse them with the appearance classifier for the two-stream video classification.

Classification Results on Kinetics-400

We first evaluate the two-stream classifications on the Kinetics-400. After training the MoNet, we finetune the classifier of both the appearance and the optical flow streams for 20 epochs. As shown in Table 3, with different classifiers, the two-stream models with the hallucinated optical flow stream consistently improve the results against the single appearance stream. Specifically, the two-stream models with the linear classifier and the NeXtVLAD classifier [Lin *et al.*, 2018] achieves 71.32% and 72.40% on the top-1 accuracy, respectively. It illustrates that the hallucinated optical flow stream effectively complements the appearance stream for action recognition. Compared with the two-stream models with ground-truth optical flow features, the performances of the hallucinated two-stream models are about 2% lower. However, the hallucinated two-stream models only rely on the appearance stream, which decreases half of the floating-point operations, from 425 GFLOPs (two-stream) to 224 GFLOPs (hallucinated two-stream), and saves the computational and storage consumptions of the optical flow images.

Classification Results on YouTube-8M

The YouTube-8M dataset is a challenging large-scale multi-labels video dataset, which consists of 6 millions of YouTube videos with 3 labels per video on average. Thus, it takes at least a hundred days to extract the optical flow images even with GPU and occupies over 100TB storage space. As such, extracting optical flow images for such large-scale video dataset are intractable. The proposed MoNet helps to improve video classification by hallucinating motion representations, with no reliance of the optical flow images.

Constrained by storage and data delivery, the YouTube-8M dataset only provides the appearance features extracted by the Inception network [Szegedy *et al.*, 2015] trained on the ImageNet [Deng *et al.*, 2009] at 1 FPS, while the optical flow features are not provided. Instead, we train the MoNet on the kinetics-400 dataset based on the same appearance features from the Inception network and the I3D optical flow features. After training, the MoNet is performed on the YouTube-8M video dataset to hallucinate optical flow feature for the two-stream classification. We follow the similar data split as in [Lin *et al.*, 2018; Tang *et al.*, 2018] that reserves 15% of video for validation, which contains about 100k video clips. We adopt four different metrics to evaluate the classification results on this dataset including Hit@1, Global Average Precision at 20 (GAP@20), Mean Average Precision at 20 (MAP@20), and Precision at Equal Recall Rate (PERR).

As shown in Table 4, the hallucinated two-stream equipped with the NeXtVLAD classifier achieves the best result, reaching 88.62%, 80.41%, 79.39% and 46.79% of Hit@1, PERR, GAP@20 and MAP@20, respectively. Compared with the single appearance stream, the two-stream models with the hallucinated optical flow features by the MoNet consistently improve the video classification performances. Thus, the proposed MoNet indeed models motion information, which complements the appearance information and helps to improve large-scale video classifications.

5 Conclusion

In this paper, we propose a novel network, namely MoNet, to hallucinate the optical flow features from the appearance ones, without the heavy optical flow computation. MoNet models temporal relationships of the appearance features and contextual relationships of the optical flow features in a concurrent way, which can effectively hallucinate the optical flow features. By incorporating the hallucinated optical flow features with the appearence features, the video classification performances can be consistently improved on the large-scale Kinetics-400 and Youtube-8M datasets.

References

[Abu-El-Haija *et al.*, 2016] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint 1609.08675*, 2016.

[Brox *et al.*, 2004] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *ECCV*, 2004.

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[Chen *et al.*, 2018a] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018.

[Chen *et al.*, 2018b] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *CVPR*, 2018.

[Chen *et al.*, 2019] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *AAAI*, 2019.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint 1406.1078*, 2014.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Donahue *et al.*, 2015] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[Dosovitskiy *et al.*, 2015] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *CVPR*, 2015.

[Feng *et al.*, 2018] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, and Jiebo Luo. Video re-localization. In *ECCV*, 2018.

[Feng *et al.*, 2019] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Spatio-temporal video re-localization by warp lstm. In *CVPR*, 2019.

[Gao *et al.*, 2018] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2flow: Motion hallucination from static images for action recognition. In *CVPR*, 2018.

[Gilmer *et al.*, 2017] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. *ICML*, 2017.

[Graves and Jaitly, 2014] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, 2014.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Ilg *et al.*, 2017] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017.

[Jiang *et al.*, 2018] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *ECCV*, 2018.

[Kay *et al.*, 2017] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint 1705.06950*, 2017.

[Li *et al.*, 2018] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *CVPR*, 2018.

[Lin *et al.*, 2018] Rongcheng Lin, Jing Xiao, and Jianping Fan. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. *arXiv:1811.05014*, 2018.

[Liu *et al.*, 2019] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. Multi-granulairty generator for temporal action proposal. In *CVPR*, 2019.

[Miech *et al.*, 2017] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint 1706.06905*, 2017.

[Peng and Schmid, 2016] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *ECCV*, 2016.

[Schuster and Paliwal, 1997] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[Tang *et al.*, 2018] Yongyi Tang, Xing Zhang, Jingwen Wang, Shaoxiang Chen, Lin Ma, and Yu-Gang Jiang. Non-local netvlad encoding for video classification. *arXiv:1810.00207*, 2018.

[Wang *et al.*, 2018] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *CVPR*, 2018.

[Wu *et al.*, 2018] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Compressed video action recognition. In *CVPR*, 2018.

[Zach *et al.*, 2007] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007.

[Zheng *et al.*, 2015] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

[Zhou *et al.*, 2018] Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. *ECCV*, 2018.

[Zhu *et al.*, 2017] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander G Hauptmann. Hidden two-stream convolutional networks for action recognition. *arXiv preprint 1704.00389*, 2017.