SkateboardAI: The Coolest Video Action Recognition for Skateboarding (Student Abstract)

Hanxiao Chen
Department of Automation
Harbin Institute of Technology
Harbin, China 150001
hanxiaochen@hit.edu.cn

Abstract

Impressed by the coolest skateboarding sports program from 2021 Tokyo Olympic Games, we are the first to curate the original real-world video datasets "SkateboardAI" in the wild, even self-design and implement diverse uni-modal and multi-modal video action recognition approaches to recognize different tricks accurately. For uni-modal methods, we separately apply (1) CNN and LSTM; (2) CNN and BiLSTM; (3) CNN and BiLSTM with effective attention mechanisms; (4) Transformer-based action recognition pipeline. Transferred to the multi-modal conditions, we investigated the two-stream Inflated-3D architecture on "SkateboardAI" datasets to compare its performance with uni-modal cases. In sum, our objective is developing an excellent AI sport referee for the coolest skateboarding competitions.

SkateboardAI Datasets

We develop the new skateboarding video datasets "SkateboardAI" from multiple platforms including YouTube, Twitter, and Instagram. In special, we choose 15 different fundamental tricks: *Ollie, Kickflip, Shuvit, Manual, Hardflip, 50-50 grind, 5-0 grind, Backside 180, BacksideAir, Boardslide, Boneless180, Smithgrind, Benihana, Impossible, Treflip.* Such tricks (Fig. 1) are usually performed in matches, thus it's valuable to train an excellent AI skateboard referee within diverse competitions. For each category, we collect 50 custom videos so that the total number is 750. In this case, we apply 45 videos for training and the other 5 items for validation with each "SkateboardAI" video class.

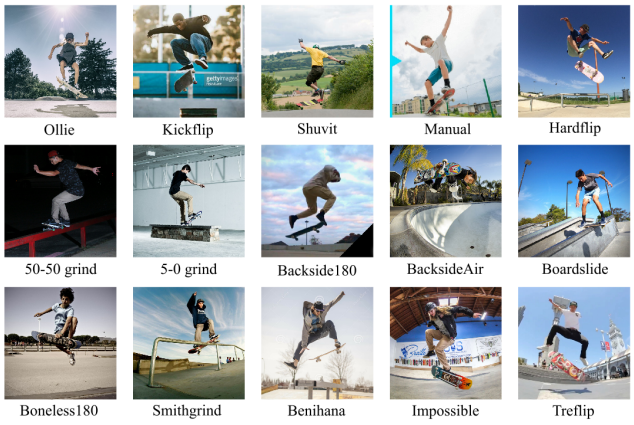

Figure 1: 15 different classes in "SkateboardAI".

Diverse Approaches

CNN-LSTM & CNN-BiLSTM

Following the fundamental CNN-LSTM training pipeline, we apply the cv2.CAP_PROP_FRAME_COUNT to compute frame number on "SkateboardAI" videos and use reasonable sampling techniques to input specific number of frame sequences (e.g., 45, 60) for feature extraction, then pass them together to the following LSTM model along the temporal dimension into one index of video category. Moreover, Bidirectional LSTM consisting of two forward and backwards LSTMs can achieve better performance since adjacent video frames contain strong correlations so that BiLSTM can preserve all information from past and future to improve reasonable context for the CNN-BiLSTM method as in Fig. 2.

CNN-BiLSTM with Attention

It is scientifically researched that "attention" is a very important mechanism in human perception system, so we integrate a simple but efficient attention block as (You and Korhonen 2020) into the CNN-BiLSTM architecture to appropriately utilize "attention" mechanisms on the most useful and considerable extracted temporal video features for better classification. In theory, it is possible to add attention on the output of CNNs before LSTM layers or after LSTM layers that produce a sequence of outputs, which can derive two different training pipelines: (i) CNN-BiLSTM-Attention; (ii) CNN-Attention-BiLSTM. Fig. 3 concisely introduces the CNN-BiLSTM-Attention framework that combines the attention block after BiLSTM with two permutation sections. Actually, the attention block can be developed by adding a fully connected layer with an activation function (Softmax) serving as probability distribution to be multiplied with outputs.

Transformer-based Method

Consider that Transformer (Vaswani et al. 2017) is a ubiquitous deep learning approach to learn context and track relationships in the sequential data (e.g., sentence, videos), we

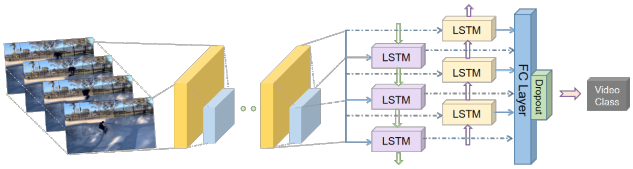

Figure 2: CNN-BiLSTM action recognition pipeline.

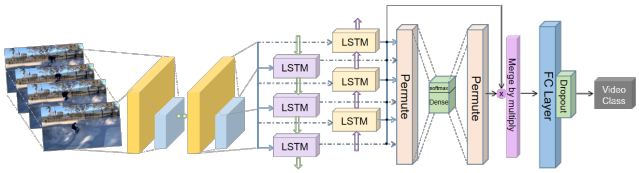

Figure 3: CNN-BiLSTM-Attention framework.

also explore and implement the transformer-based method for "SkateboardAI" action recognition. Firstly, we apply an ImageNet pre-trained CNN backbone (e.g., DenseNet121, ResNet50) for feature extraction and pad each video shorter to certain sequence length. Then, we embed the positions of frames present inside videos with an Embedding Layer and add these positional embeddings to the pre-computed CNN video feature maps to make self-attention layers within the Transformer consider order-agnostic information. Next, we employ the Global Max Pooling Layer and Dropout to the outputs of Transformer model for further video recognition.

I3D Multi-modal Method

Transferred to the multi-modal case, we investigate the popular Inflated 3D ConvNet (I3D) (Carreira and Zisserman 2017) method to recognize skateboarding tricks. Similar to the official implementation on Kinetics Datasets, we directly feed the whole video into the pipeline rather than the derived video frame RGB images. In fact, the I3D model is a version of Inception-V1 with batch normalization which has been pre-trained on ImageNet and then "inflated" from 2 dimensions into 3 dimensions. I3D utilizes a two-stream architecture with two good modalites of videos: RGB and optical flow. Each stream is separate and the output of models will be combined only at the logit-level for class prediction.

Experiments

We conduct extensive experiments with diverse models to recognize "SkateboardAI" datasets. For CNN-LSTM based approaches consisting of CNN-BiLSTM, CNN-BiLSTM-Attention, and CNN-Attention-BiLSTM, we investigate 4 different CNN backbones including ResNet50, ResNet152, DenseNet and VGG16, even set the input sequence length as 45 and resize the video frame to be (299,299) for 100 training epochs, then we collect three important metrics for each condition: *Training time, Training accuracy, and Validation accuracy*. As for Transformer solutions, we implement experiments with two CNN backbones including DenseNet121 and ResNet50, even investigate 45 or 100 input sequences for 1000 epochs with two different frame resize (224,224)

Method	Train time/s	Train_acc	Val_acc
ResNet50+LSTM (2048)	3676.15	0.9956	0.8000
ResNet50+BiLSTM	**3593.21**	1.0000	0.8133
ResNet50+"A"+BiLSTM	3967.73	1.0000	**0.8400**
ResNet50+BiLSTM+"A"	3916.80	0.9926	0.8133
ResNet152+LSTM (2048)	6145.99	0.9926	0.7867
ResNet152+BiLSTM	6443.21	0.9985	0.8000
ResNet152+"A"+BiLSTM	6422.44	1.0000	0.8133
ResNet152+BiLSTM+"A"	6372.50	0.9985	0.7733
DenseNet121+LSTM (1024)	2881.21	0.9822	0.7467
DenseNet121+BiLSTM	3536.98	1.0000	0.7067
DenseNet121+"A"+BiLSTM	3067.68	0.9970	0.8000
DenseNet121+BiLSTM+"A"	2817.63	1.0000	0.6933
Vgg16+LSTM (512)	2405.80	0.9867	0.7067
Vgg16+BiLSTM	2347.70	1.0000	0.6933
Vgg16+"A"+BiLSTM	3247.76	1.0000	0.6533
Vgg16+BiLSTM+"A"	3324.39	0.9970	0.7867
T_DenseNet121 (crop_center, 45)	17936.21	1.0000	0.3529
T_ResNet50 (crop_center, 45)	36845.02	1.0000	0.2647
T_DenseNet121 (tf.resize, 100)	34133.97	1.0000	0.3333
T_ResNet50 (tf.resize, 100)	68310.90	1.0000	0
I3D_10 epochs	432166.05	0.072	0.081
I3D_50 epochs	1721886.04	0.074	0.078

Table 1: Experiments ("A": Attention, "T": Transformer).

techniques containing "crop_center" and "tf.image.resize". Unlike the uni-modal methods, I3D co-processes both RGB and optical flow modalities for each video with longer training time so that we train 10 & 50 epochs concisely. Table 1 presents the whole numerical results for all implemented methods. Obviously, we discover that ResNet50-Attention-BiLSTM achieves the best performance on our "SkateboardAI" with 84% validation accuracy in just 3967.73s training time. Compared with ResNet50-LSTM & ResNet50-BiLSTM, our integrated attention mechanism can exactly improve the classification accuracy and the attention block can perform much better before the BiLSTM instead of after it. Moreover, Transformer-based methods and I3D can not address "SkateboardAI" action recognition very well as CNN-LSTM cases because of different model architectures, much longer training time and lower evaluation accuracy. More experimental results and detailed analysis can refer to https://github.com/2000222/Skateboard-AI.

References

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 4724–4733.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, 5998–6008.

You, J.; and Korhonen, J. 2020. Attention Boosted Deep Networks For Video Classification. In *IEEE International Conference on Image Processing, ICIP 2020*, 1761–1765.