Dean&Francis

ISSN 2959-6157

The Difference between BiLSTM and Transformer - The Discussion about Suitable Scenarios for BiLSTM and Transformer

Hongzhou Lei

School of Computer Science, Central China Normal University, Wuhan, China
Corresponding author: lei23516417zhou@gmail.com

Abstract:

Action recognition is becoming more and more important with the development of autopilot cars and smart homes. Many people hope that autopilot cars can make transportation safer. Furthermore, action recognition can make their homes smart and make their daily lives more convenient. The common demand of these technologies is action recognition. This paper researches the differences between Bi-direction Long Short-Term Memory (BiLSTM) and Transformer to find their suitable scenes. This paper main research topic is motion recognition using BiLSTM and Transformer. The results show that in one train, the train epoch of BiLSTM is 178.41 seconds, Transformer is 1657.75 seconds; the train time in a similar fitter degree of BiLSTM is 19 times, but Transformer need 50 times to achieve a similar fitter degree. Therefore, BiLSTM is suitable for low-power-consumption devices and scenes that require high real-time recognition. The Transformer is suitable for unlimited-power-consumption devices and requires high recognition accuracy in scenes. In the future, this research will add more motions and add transitions between every motion. Furthermore, the research will improve codes to make results more accurate and make training more efficient.

Keywords: Skeleton-based motion recognize; Human pose estimation; BiLSTM; Transformer encoder; Real-time inference

1. Introduction

In recent years, many smart products like smart cameras, smart homes, autopilot cars, and others need motion recognition and forecasting in real time to ensure safety and utility. Autopilot cars need to handle fatigue driving and crossing at pedestrian crossings. Smart homes need to recognize many complicated

actions from the human body to make the electrical devices in the home more convenient. These products always use cameras to recognize humans and other actions.

There are some traditional methods to resolve action recognition, such as Recurrent Neural Network (RNN), Histogram of Oriented Gradients (HOG), and Improved Dense Trajectories (iDT). RNN has serious problems with gradients and cannot recognize in real time; meanwhile, this model is difficult to run parallel calculations and has a low speed of inference, so RNN is not useful in high-real-time scenes. HOG can only recognize a single frame at the same time and has poor accuracy for action recognition. Therefore, HOG has poor robustness to a variety of actions and light directions. iDT has heavy feature-stacking weight and cannot recognize actions in real time either. Therefore, these three models all have the same problem: they have high performance requirements and cannot have both real-time and accuracy, so they cannot, or are difficult to, run on limited-performance devices and in high-real-time scenes.

In modern city, a great mass of devices need high real-time calculation and limit their performance to compatible popular devices, but traditional models like RNN, HOG and iDT have some defects in this scenario. So this paper only inputs simple data about the human body key points sequence to directly recognize the RGB image [1] [2]. The key points data set just inputs action information in one dimension, so this method has a low performance requirement and has higher real-time performance than the traditional method. Every real-time human action belongs to a timing sequence, but one-frame recognition can not catch dependencies over a long time. So the action recognition needs a model that can recognize the long-term actions.

In recent years, some papers have proposed more balanced methods like Long Short-Term Memory (LSTM) and Transformer Encoder (Transformer). These methods provide a better balance of both precision and real-time recognition. Google Scholar shows a better method to improve LSTM: Bidirectional Long Short-Term Memory (BiLSTM), which achieves better precision than LSTM. BiLSTM and Transformer can recognize at the same time, so this paper contrasts these methods in recognition precision. The paper will contrast factors to research what suits scenes of two models in these factories: train epochs in similar fitting degrees, their accuracy and how time are they spent.

2. Method

The LSTM is an improved edition of the Recurrent Neural Network (RNN). It is a specialism for resolving the gradient disappearance and explosion. This model controls the information flow through the input gate, forget gate, and output gate [3]. This control method can “remember” the long dependency information. The BiLSTM is an improved model by adding a reverse propagation layer to the basic LSTM. The BiLSTM can think over the past information and future information, so this model can do better in the semantic analysis, action recognition, and sequence labelling [4]. Therefore, this paper uses the BiLSTM to conduct research.

The Transformer is fundamentally different from the traditional model RNN. This model uses the Self-Attention system to analyze the connection of the different elements [5]. In Transformer, every location vector can look over the other location vectors by Self-Attention [6]. These features can implement the common calculation [7]. Therefore, the encoder and decoder have a great expression of the machine translation, visual identity, and speech processing [8].

This paper gives two models, three motions that stand, sit and walk. Every motion has two videos from a realistic game, and uses MediaPipe to pick up 33 key bones from videos to build a dataset. This research uses PyTorch to initialize two models and uses OpenCV to read videos [9,10]. When training these models, this paper adds some arguments to improve training precision [11,12]. These arguments include Automatic Mixed Precision (AMP), gradient clipping, early-stopping and learning rate scheduling.

This paper contrasts two models in 3 aspects: their accuracy in a similar fitting degree, their training time in one train and how they spent time in one train.

3. Results

Firstly, about the accuracy(acc) rate, with the same data set to train two models, the precision value with a similar degree of fitting as figure 1 and figure 2:

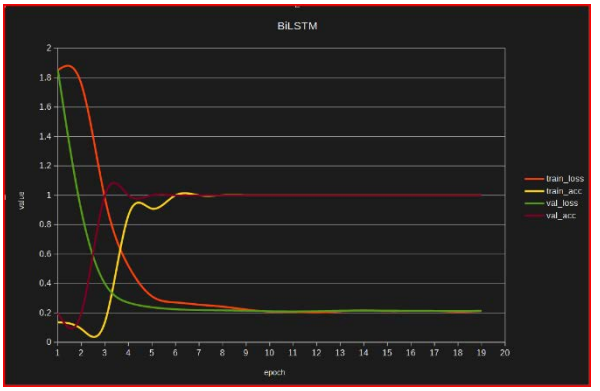

Fig. 1 BiLSTM accuracy in similar fitting degree

ISSN 2959-6157

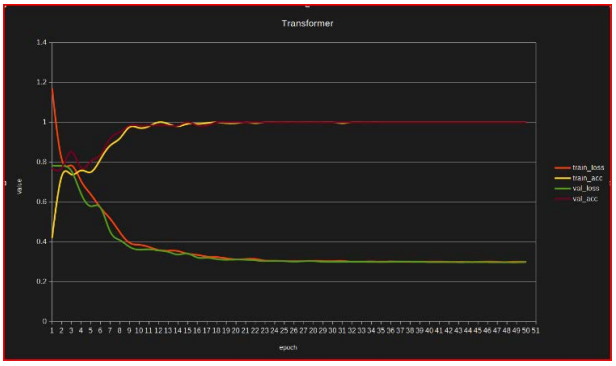

Fig. 2 Transformer accuracy in similar fitting degree

Table 1. Accuracy value

	BiLSTM		Transformer	
epoch	train_acc	val_acc	0.42	0.77
1	0.14	0.2	0.73	0.77
2	0.09	0.2	0.74	0.85
3	0.14	1	0.76	0.77
4	0.86	1	0.75	0.8
5	0.91	1	0.81	0.84
6	1	1	0.88	0.92
7	1	1	0.92	0.95
8	1	1	0.98	0.98
9	1	1	0.97	0.98
10	1	1	0.98	0.98
11	1	1	1	0.98
12	1	1	0.99	0.98
13	1	1	0.98	0.98
14	1	1	0.99	1
15	1	1	0.99	0.98
16	1	1	1	0.98
17	1	1	1	1
18	1	1	1	1
19	1	1	0.42	0.77

Table 1 shows that the BiLSTM accuracy value fluctuates between about 0.2 and 1.00, and the Transformer accuracy value fluctuates between about 0.77 and about 1.00. The line graph(1)(2) shows BiLSTM have a better fitting degree, but table 1 show Transformer have a higher average accuracy value. This table illustrates that BiLSTM train speed is faster than Transformer, and Transformer have more accuracy in the same train times.

Next, these value about performance in table 2:

Table 2. Performance value

	BiLSTM	Transformer	
FPS	19.9	12.9	Frame/second
FLOPs	3.78	7.26	Mmac
Spent train time in once train	178.41	1657.75	seconds

Table 2 show that in real-time recognition, for a frame one second, the BiLSTM has 19.9 FPS, the Transformer has 12.9 FPS; for flops, the transformer value is 7.26, BiLSTM is 3.78; for spent time in one train, BiLSTM use 178.41 seconds, Transformer use 1657.75 seconds. This table illustrates that BiLSTM has better performance than Transformer on the same device. This table illustrates that BiLSTM is more friendly for low-performance devices than Transformer, and BiLSTM can save more time to train than Transformer.

4. Discussion

This research just includes 3 actions, all simple fixed actions and does not have a transition between actions. In the future, this research will incorporate additional action labels and include transition labels between every action. The transition label can forecast action variety faster and with higher accuracy. Next, the research will add more angle of view. This research just has 1 or 2 angles of view, so the recognize may have not enough accuracy. In addition, these model in this research have little lopsided. In real-time recognize, two models all mix up three actions: the action is "sit" in real, but the label usually set to stand; if the action not include full body, the label is set to "walk" in some time. Furthermore, the research will improve codes to make training more efficient and more effective. More efficiency can save a lot of time for training the models, so improving the code is good to make the program more useful. In future, this paper will add more comparer of different models such like Temporal Conv (TCN), Spatial Temporal Graph Convolutional Networks (ST-GCN) and more.

5. Conclusion

The research result of this paper illustrates that the BiLSTM is suitable for high requirements scenes and low dissipation devices such as personal computers, home safety systems and autopilot car self-calculation core. These devices need lower requirements of recognition accuracy, and these devices can not have high dissipation because they can not make high power consumption and can not make high temperature to ensure the user's safety and keep their cost economic. The transformer is suitable for unlimited power consumption and has enough dissipation devices. Those devices are common in big centre calculation servers. That server is often used for urban safety monitoring and the cloud autopilot calculation centre. These devices need high accuracy of recognition to keep people safe and make people more convenient, and most scenes have a low requirement for real-time recognition. Besides the server, the Transformer can be used for action data acquisition about films and games. Many papers illustrates that high accuracy can make the process more convenient.

References

[1] Lugaresi F, et al. MediaPipe: A Framework for Building Perception Pipelines. CVPR Workshops, 2019.

[2] Bazarevsky V, et al. BlazePose: On-device Real-time Body Pose Tracking. ArXiv:2006.10204, 2020.

[3] Grishchenko I, et al. BlazePose GHUM Holistic: Real-time 3D Human Landmarks and Pose Estimation. ArXiv:2206.11678, 2022.

[4] Plizzari C, Cannici M, Matteucci M. Skeleton-based Action Recognition with Spatio-Temporal Transformer. ICPR, 2021 (arXiv preprint).

[5] Yan S, Xiong Y, Lin D Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. AAAI, 2018.

[6] Si C, et al. An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-based Action Recognition (AGC-LSTM). CVPR, 2019.

[7] Wang L, et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. ECCV, 2016.

[8] Bewley A, et al. Simple Online and Realtime Tracking (SORT). ICIP, 2016 (arXiv preprint).

[9] Micikevicius P, et al. Mixed Precision Training. ICLR, 2018 (arXiv:1710.03740).

[10] Loshchilov I, Hutter F Decoupled Weight Decay Regularization (AdamW). ICLR, 2019.

[11] Müller R, Kornblith S, Hinton G When Does Label Smoothing Help? NeurIPS, 2019.

[12] Chaudhry A, et al. On Tiny Episodic Memories in Continual Learning. arXiv:1902.10486, 2019.