COGS260 Image Recognition
Instructor: Prof. Zhuowen Tu

A spatiotemporal model with visual attention for video classification

Mo Shan

Department of Electrical and Computer Engineering

June 7, 2017

Outline

Motivation

Proposed model

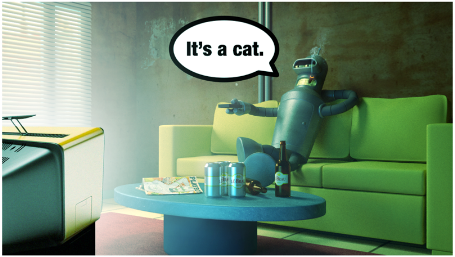
Experiment

Conclusion

Motivation

Video classification

- Semantic understanding of sequential visual input is important for robots in localization and object detection.
- Eg, search for a cat in a living room, instead of in a gym.

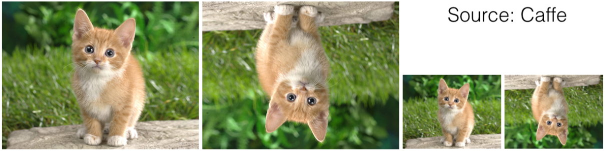


Source: Harvey M., Five video classification methods

Motivation

Rotation and scale

- Existing benchmark contains videos of daily scenes.
- Objects in real world could be rotated and scaled.



Source: Caffe

Original

Rotated

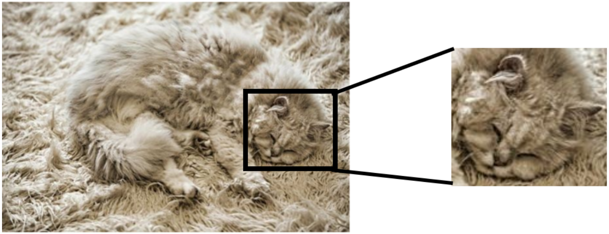
Scaled

Rotated & scaled

Motivation

Visual attention

- Attention mechanism reduces complexity and avoids cluttering. This makes it easier to deal with rotated and scaled images.

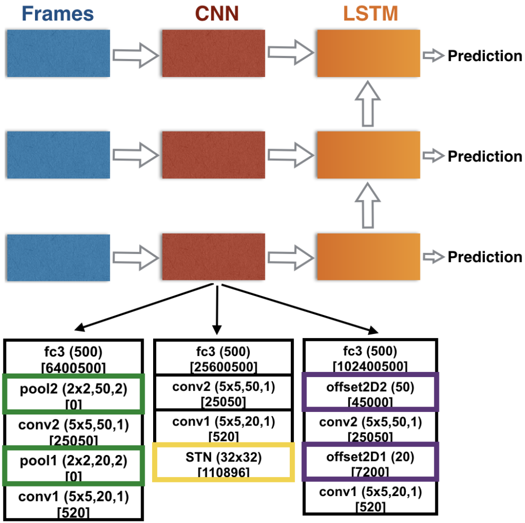


Source: cs231n, Stanford

Proposed model

Architecture

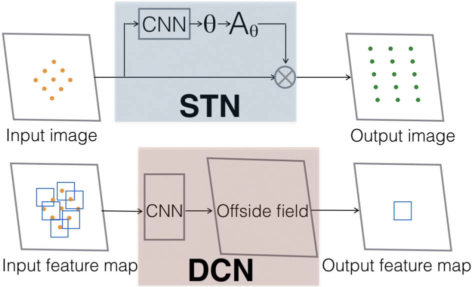
- The proposed model concatenates CNN to RNN.
- The CNN stage is augmented with attention modules.



Proposed model

Attention modules

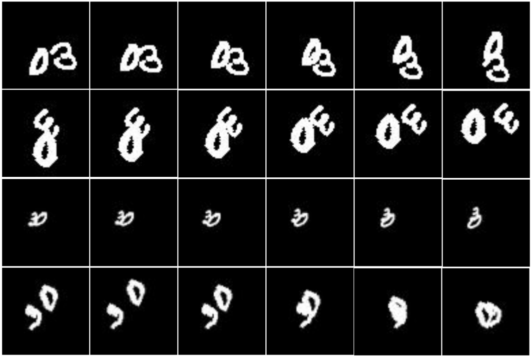
- Spatial transformer network learns a global affine transformation.
- Deformable convolutional networks learns offsets locally and densely.



Experiment

Dataset

- Moving MNIST is augmented with rotation and scaling (Demo).



Experiment

Quantitative analysis

- DCN-LSTM consistently performs the best in all cases.

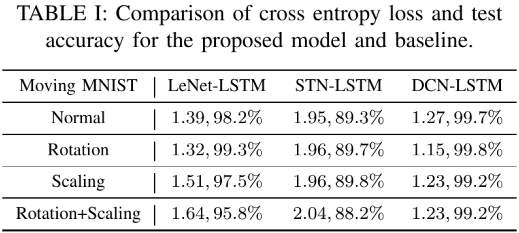


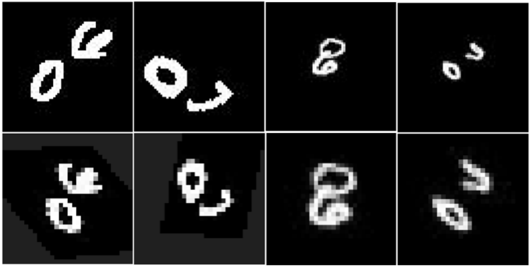
TABLE I: Comparison of cross entropy loss and test accuracy for the proposed model and baseline.

Moving MNIST	LeNet-LSTM	STN-LSTM	DCN-LSTM
Normal	1.39, 98.2%	1.95, 89.3%	1.27, 99.7%
Rotation	1.32, 99.3%	1.96, 89.7%	1.15, 99.8%
Scaling	1.51, 97.5%	1.96, 89.8%	1.23, 99.2%
Rotation+Scaling	1.64, 95.8%	2.04, 88.2%	1.23, 99.2%

Experiment

Qualitative analysis

- STN could not attend to each digit individually.



Conclusion

Key insights

- DCN-LSTM achieves high accuracy compared to baseline.
- Attention modules are useful to deal with rotation and scale changes.
- STN-LSTM does not perform well due to global transformation.
- How to train the entire model end to end remains a future work.