# A spatiotemporal model with visual attention for video classification

Mo Shan and Nikolay Atanasov

Department of Electrical and Computer Engineering

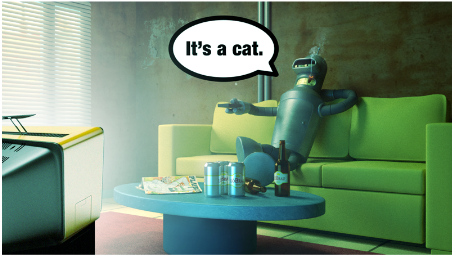
July 16, 2017

# Outline

# 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.

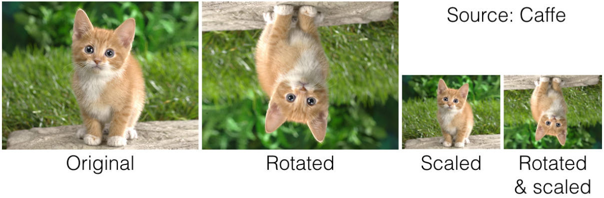



Source: Harvey M., Five video classification methods

# Motivation

## Rotation and scale

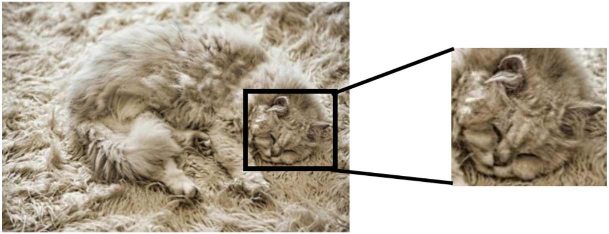
- Existing benchmark contains videos of daily scenes.
- Objects in real world could be rotated and scaled.



# Motivation

## Visual attention

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

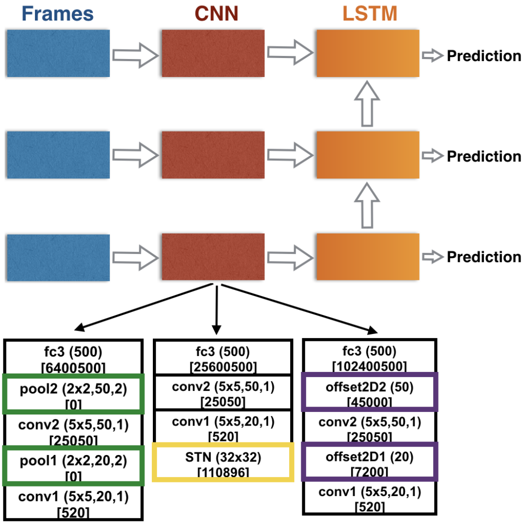


Source: cs231n, Stanford

# Proposed model

## Architecture

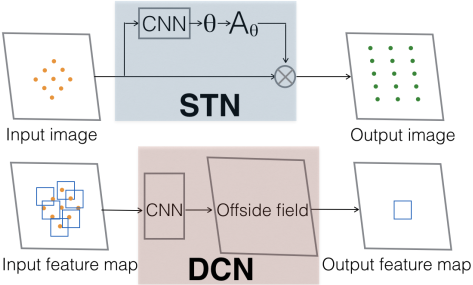
- The proposed model concatenates CNN to RNN.
- The CNN stage is augmented with attention modules.



# Proposed model

## Attention modules

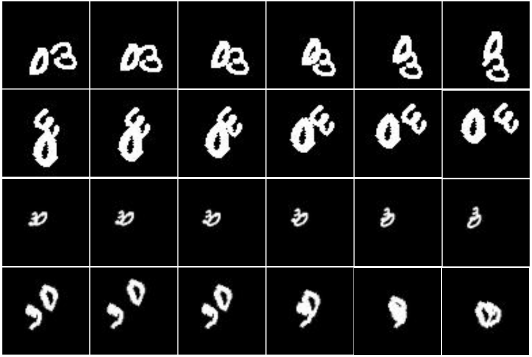
- STN (Jaderberg, 2015) learns a global affine transformation.
- DCN (Dai, 2017) learns offsets locally and densely.



# Experiment

## Dataset

- Moving MNIST is augmented with rotation and scaling.



# Experiment

## Quantitative analysis

- Results are shown in Table 1.
- DCN-LSTM consistently performs the best in all cases.

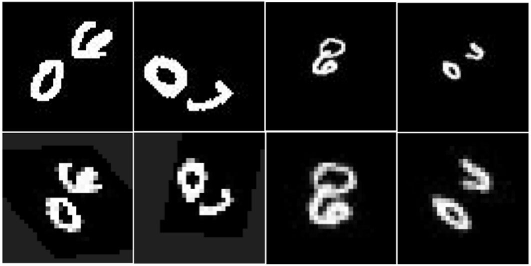
Table: Comparison of cross entropy loss and test accuracy for the proposed model and baseline.

| Moving MNIST | LeNet-LSTM | STN-LSTM | DCN-LSTM |
|---|---|---|---|
| Normal | 1.44, 97.96% | 1.98, 87.26% | 1.27, 99.62% |
| Rotation | 1.42, 98.43% | 1.97, 90.47% | 1.29, 99.70% |
| Scaling | 1.52, 96.28% | 1.99, 86.90% | 1.28, 99.41% |
| Rotation+Scaling | 1.51, 96.82% | 1.99, 89.10% | 1.25, 99.46% |

# Experiment

## Qualitative analysis

- STN could not attend to each digit individually.



# Experiment

## Digit gesture classification

- Elastic deformation simulates oscillations of hand muscles.
- Results are shown in Table 2.
- DCN could learn the deformation field explicitly.
- DCN-LSTM has the potential to handle articulated objects.

Table: Cross entropy loss and test accuracy for deformed digits.

| LeNet-LSTM | STN-LSTM | DCN-LSTM |
|---|---|---|
| 1.48, 97.19% | 1.48, 97.19% | 1.28, 99.30% |

# Conclusion

## Key insights

- DCN-LSTM achieves high accuracy compared to baseline.
- Attention isuseful to deal with rotation and scale changes.
- STN-LSTM performs poorly due to global transformation.
- Future work: how to train the entire model end to end.