Machine Learning II Final Project - Temporal Shift Module Evaluation

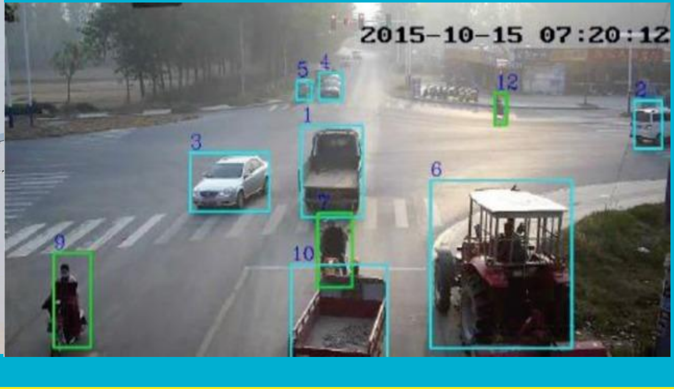
Peter Shagnea and Sarvesh Bhagat

Agenda

- Motivation
- Methods
- Data
- Training
- Testing
- Conclusion

Video Understanding - Applications

- Edge /On premise Devices
  - Drone/Surveillance
  - Medical devices
  - Self driving Cars
- Cataloging large databases
  - >10^5 hours of videos uploaded to youtube daily



Video Understanding - Methods

- Activation function for a video model represented as $A \epsilon R^{N \times C \times T \times H \times W}$
  - N: Batch Size
  - C: Channels
  - T: Temporal Dimension
  - HxW: Pixels

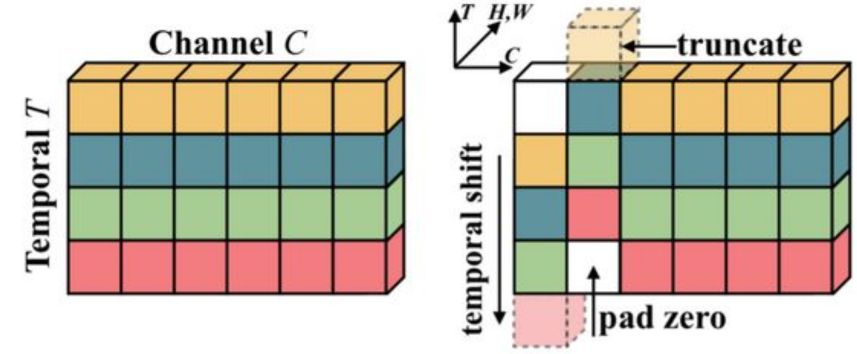
- 2D CNN operates independently over T
  - Relatively efficient but cannot infer temporal order
  - Can be combined w/LSTM , temporal relation network, etc.

- 3D CNN inflates 2D kernels to 3D
  - Computationally intensive, larger number of parameters

Temporal Shift Module

- Developed at MIT
  - “TSM: Temporal Shift Module for Efficient Video Understanding” [ICCV 2019]
- First place on something-something V2 leaderboard

- Computational efficiency comparable to a 2D CNN



(a) The original tensor without shift.

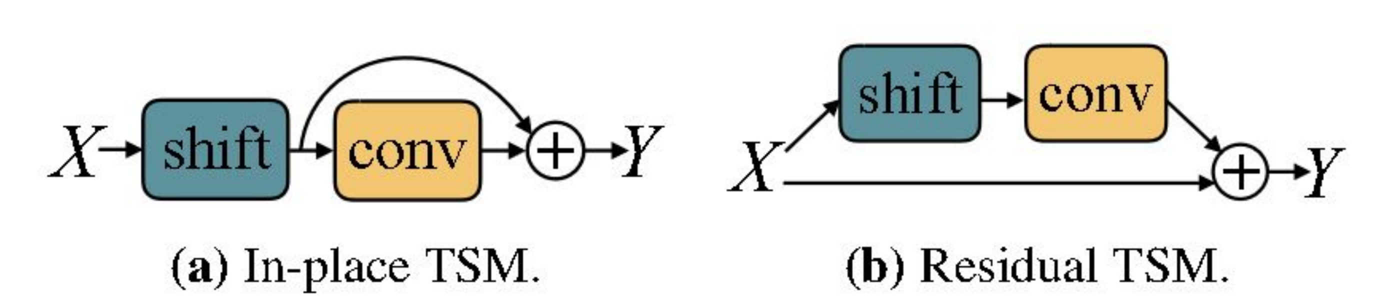
(b) Offline temporal shift (bi-direction).

Temporal Shift Module

- TSM Shifts channels along temporal dimension
  - Shift occurs in residual branch
  - ResNet50 and ResNet101 used here

- Optimal value for shift is ⅛ of the channels

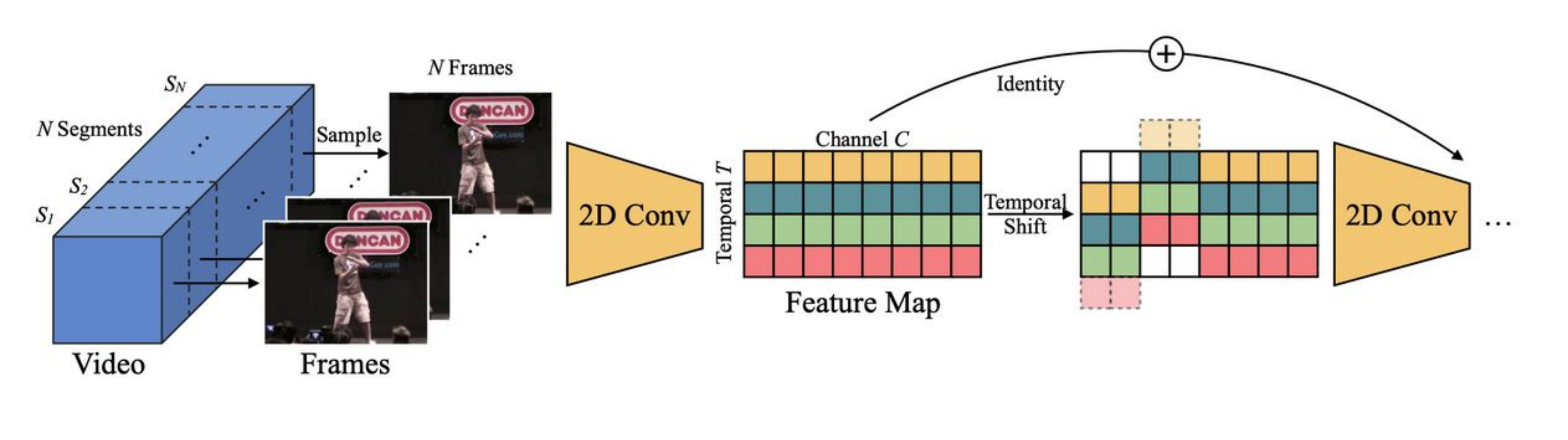
- Values can be shifted both forward and backwards in time
  - Online version shifts only backwards



(a) In-place TSM.

(b) Residual TSM.

Temporal Shift Module



Dataset Description

- 20BN-SOMETHING-SOMETHING V2 is a large, densely labeled videoset
  - Much larger than similar datasets
- Videos of humans performing pre-defined actions
- 174 Classes - individual actions
  - Additional object notations

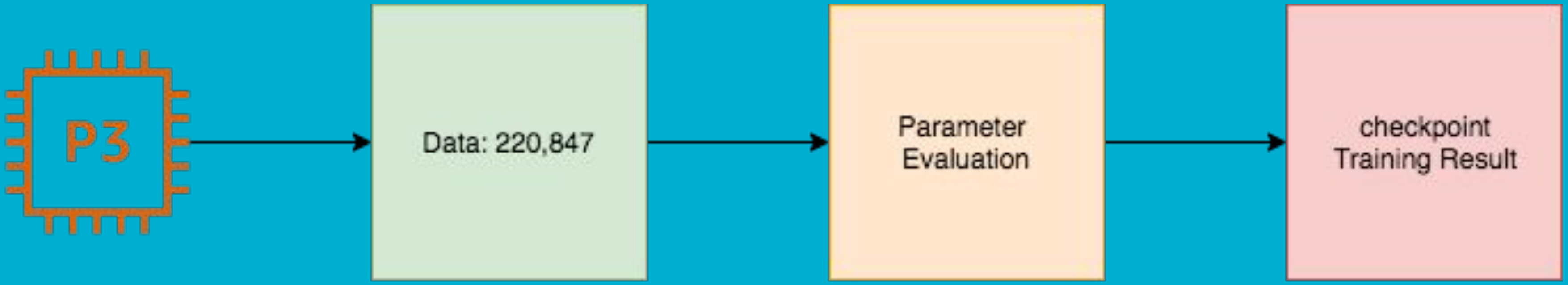
Train	Validation	Test	Total
168,193	24,777	27,157	220,847

Sample Data

Label: Trying to pour water into a glass, but missing so it spills next to it

Label: Pulling two ends of a rubber band so that it gets stretched

Training Set-UP



P3

Data: 220,847

Parameter Evaluation

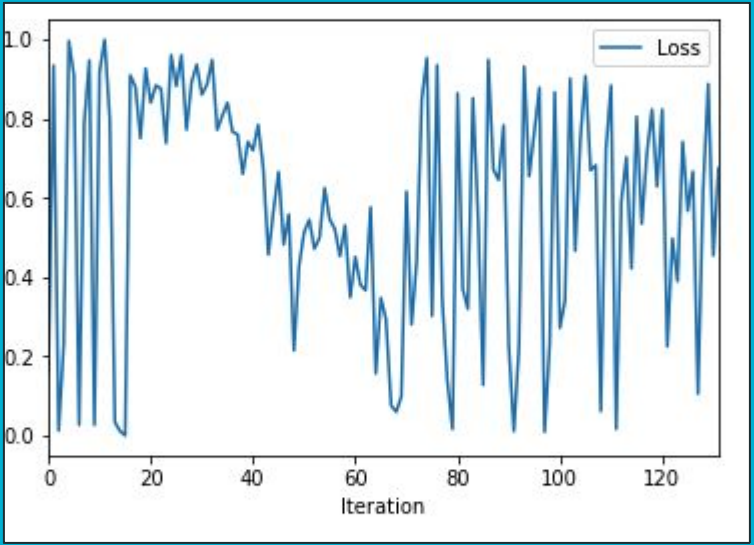
checkpoint Training Result

ml.p3.8x;arge
32 vcpu
4xV100
244 GiB
64 GB (GPU)

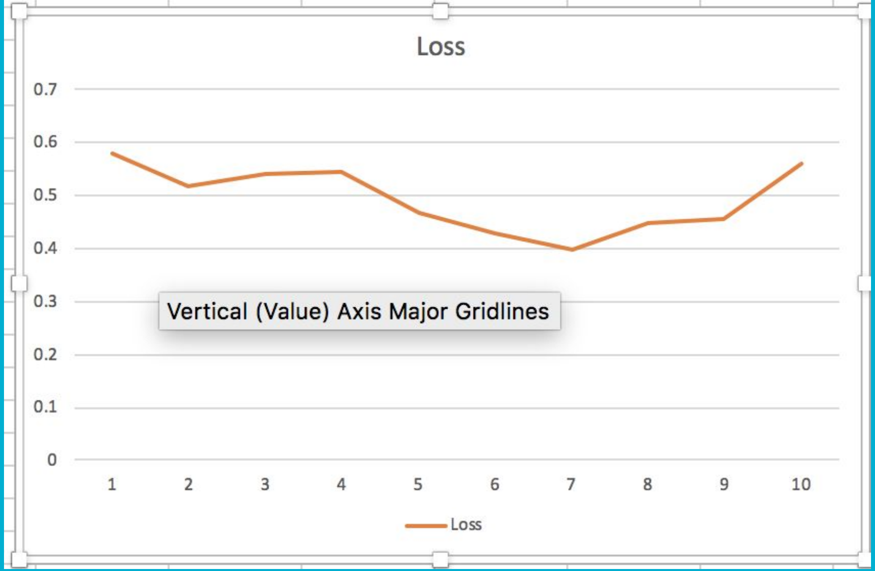
Train: 168,193
Validation: 24, 777
Test: 27,157

- Number of segments
- Learning rate
- Dropout Layer
- Epochs
- Batch size

Training Results



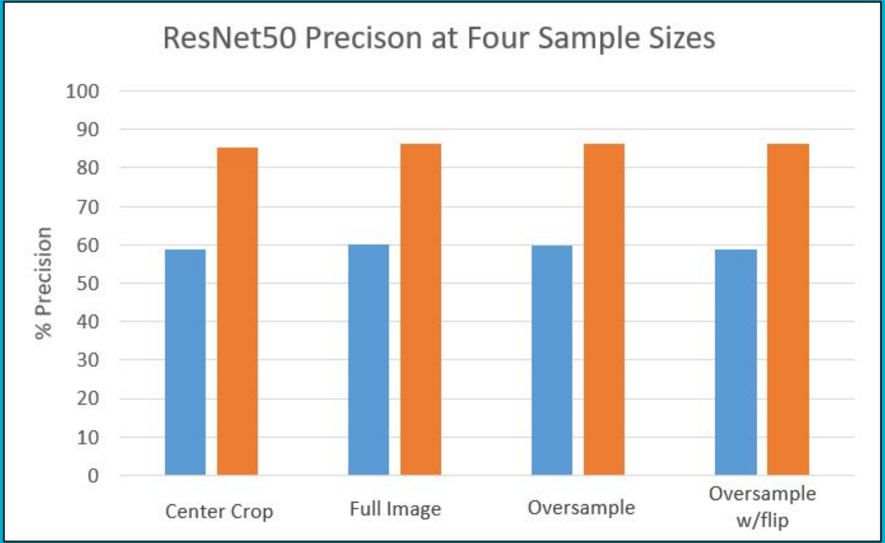
Number of epochs = 1



Number of epochs = 10

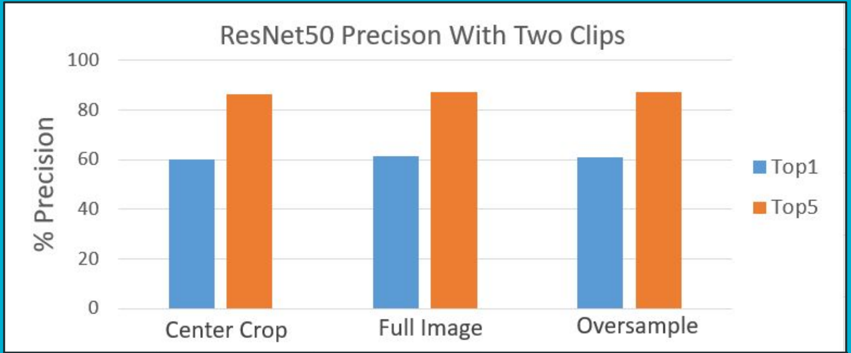
Test Set Results

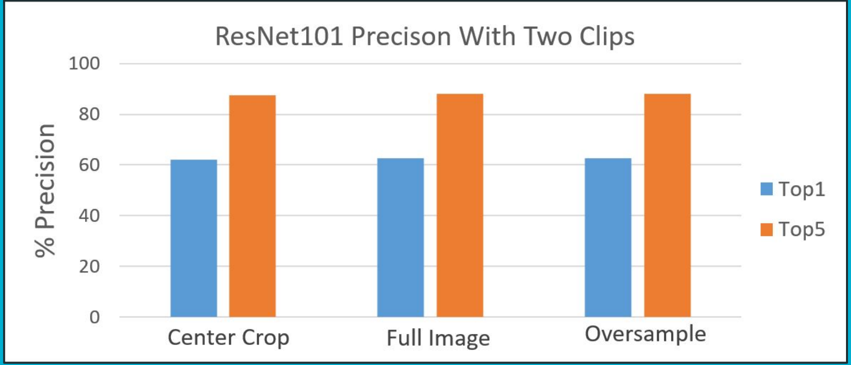
- Tested on Resnet50 and Resnet 101
- Various sampling methods for evaluation
  - Outlined in something-something paper
- Cropping had limited effect on results



Test Set Results - Network Comparison

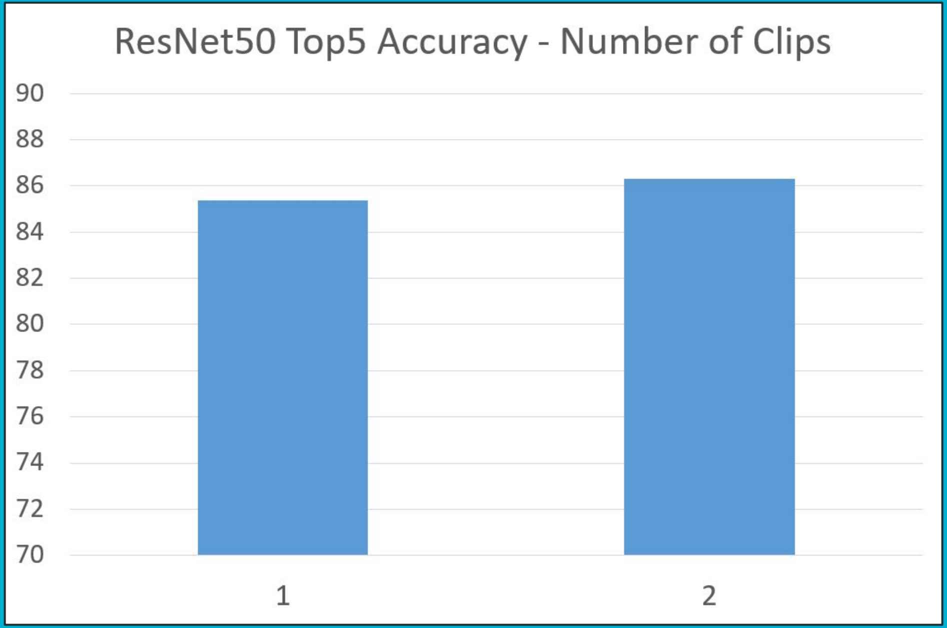
- Sample size was increased to two clips
  - Two subsets of randomly selected fraes
- Compared Resnet 50 and Resnet 101
- ResNet 101 performs better
  - Leads by small margin
  - May not be worth increase in model size





Test Set Results - Multiple Clips

- Literature suggests loading multiple clips and averaging softmax results
- Doubles evaluation time
- Minor performance gains (~0.75%)



Conclusions

- TSM enables hardware-efficient video recognition
- Can use a 2D CNN backbone to enable joint spatial-temporal modelling
- Enables low-latency video recognition on edge devices with low cost compared to 3D CNN.

Sources

[1] Lin, Ji, TSM: “Temporal Shift Module for Efficient Video Understanding”, *ICCV 2019*, August, 2019

[2] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaim-ing He.Non-local neural networks.arXiv preprintarXiv:1711.07971, 10, 2017. 1, 2, 5, 6, 7

[3] Raghav Goyal, “The "something something" video database for learning and evaluating visual common sense.”

Questions???