Item Removal Detection for Retail Environments with Neural Networks

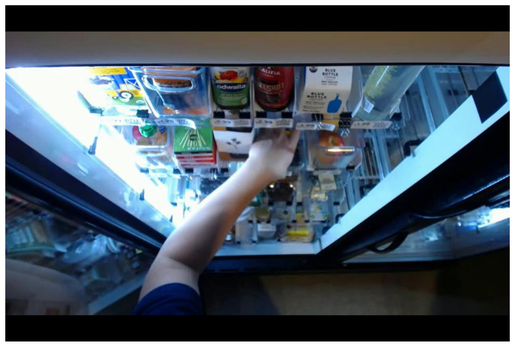
Lingjie Kong, Xingchen Fan

Introduction

- Motivation: Enhance shopping experience through item removal detection, so people can just shop and go
- Problem Definition: Given a set of video frames, determine whether people are adding items or removing items.
- Input: Video snippets of people's shopping actions
- Output: Classes of adding item, removing item, null action, as well as item classes
- Challenge: Video classification requires preprocessing the raw video and needs a more complicated network structure
- Result: Best late fusion neural network structure with SqueezeNet transfer learning model while modifying the last layer to be fully connected gives an accuracy of more than 70%.

Pre-process Data

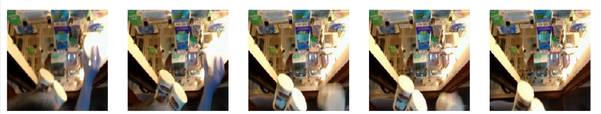
- Raw Data: Camera is mounted on top of retail refrigerator to collect raw video. Classes and frames of interest are labelled for training and prediction.
- One Frame of Video



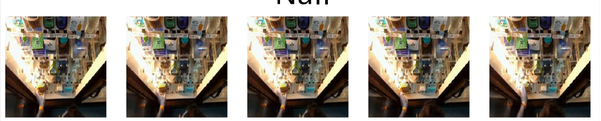
- Clip Frames from Video:

Add

Remove

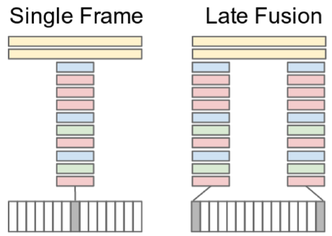


Null



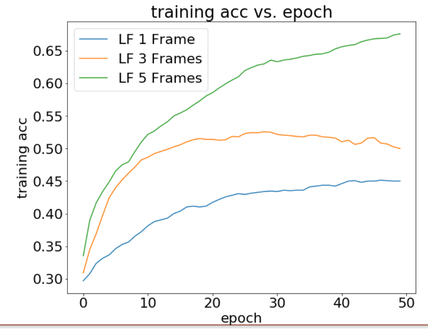
Late Fusion with CNN

- Late Fusion: Late fusion model places two or more separated single-frame networks with shared parameters. Its result will be concatenated together into a fully connected layer for prediction.



- Transfer Learning: We used SqueezeNet model pretrained on ImageNet for our late fusion. We replaced the last convolution layer with a fully connected layer and only train the last layer to make prediction.
- Training, Validation and Test: We split 1980 collected sets of video frames into 1580, 200, 200 for training, validation, and testing.
- Loss Function and Optimizer: We used the Softmax cross-entropy loss and Adam optimizer to optimize the loss function. Regularization and learning rate is fine tuned through 100 random search.
- Result: We tried 1-frame, 3-frame, and 5-frame late fusion model and 5-frame yields the best accuracy as below.

Method	Rand	LF 1 Frame	LF 3 Frames	LF 5 Frames
Training Accuracy	N/A	45%	56%	71%
Testing Accuracy	25%	44%	55%	70%



3D ConvNets (C3D)

- C3D: C3D model incorporates temporal information directly by cascading several 3D convolution layers and 3D pooling layers. Video frames are concatenated directly along the temporal dimension.

- Transfer Learning: We started with a C3D model pretrained on Sports-1M and fine-tuned the last fully-connected layer for prediction
- Training, Validation and Test: We used the same dataset as described in late fusion.
- Loss Function and Optimizer: We used the Softmax cross-entropy loss and gradient descent optimizer to optimize the loss function.
- Result: Our C3D model currently yields results no better than random guess and requires more hyperparameter tuning.

Discussion

- Discussion: Late fusion with pretrained image models have shown great performance for our problem. C3D with transfer learning also shows some promise in achieving good performance. Both models are easy to implement and fast to run, which could be beneficial for real-time implementation in retail environments

Future

- Future: Both late fusion and C3D models could be further tuned and evaluated on more of our own data. There are also other video classification methods like two-stream convolutional network and long-term recurrent convolutional network (LRCN) that are worth trying for our problem.

Reference

[1] A. Karpathy, et al, "Large-scale video classification with convolutional neural networks".
[2] D. Tran, et al, "Learning spatiotemporal features with 3D convolutional networks,"