American Sign Language Video Captioning System with Object Detection and Transfer Learning

Team 18: Michael Wang & Shuzhan Sun

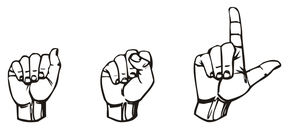
Abstract

- American Sign Language (ASL), used by the deaf community, can hardly be understood by the general public
- To bridge the communication barrier between the deaf community and the general public, in this work, a real-time video captioning system was developed
- Our video captioning system consists of object detection network and ASL classification network, and is trained on the ASL Alphabet dataset as well as a self-recorded dataset
- Our video captioning system has achieved 95%+ accuracy on the test dataset and is able to caption live stream

Introduction

American Sign Language

American Sign Language (ASL) is the 3rd most taught language in the U.S. with over 500,000 speakers. However, this is still only 0.15% of the U.S. population. This is extremely limiting on the deaf population in terms of with whom they can easily communicate. The alternative, written communication, is impersonal and sometimes even impractical during emergencies.



Previous work

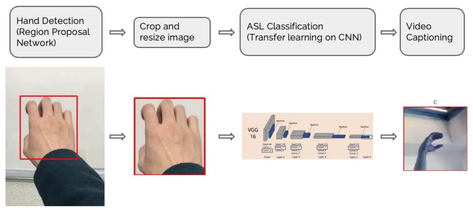
- Microsoft Kinect device to classify ASL signs in real time (Pugeault & Bowden 2011).
- Using depth-based hand tracking for ASL fingerspelling classification (Taylor 2018).
- Gloves for real-time ASL fingerspelling translation (Haydar et al. 2012).

Contributions

- Uses RGB data for training to allow for captioning to already existing videos
- Hand detection and cropping input video for better classification and reduction of overfitting to backgrounds of training data

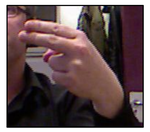
Methodology

Framework

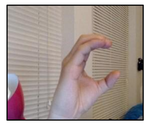


Datasets

ASL Alphabet Dataset:
- 87,000 images
- 200 x 200 pixels, RGB
- 26 classes (A-Z)

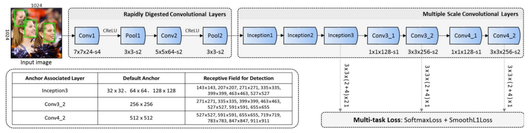


Self-Recorded Dataset:
- 2,060 images
- 640 x 480 pixels, RGB
- 26 classes (A-Z)



Object Detection

Pretrained Model (Zhang et.al. 2017):



Approach	CPU-model	mAP(%)	FPS
ACF [40]	i7-3770@3.40	85.2	20
CasCNN [16]	E5-2620@2.00	85.7	14
FaceCraft [26]	N/A	90.8	10
STN [5]	i7-4770K@3.50	91.5	10
MTCNN [44]	N/A@2.60	94.4	16
Ours	E5-2660v3@2.60	**96.0**	**20**

Transfer Learning with CNNs

Data Augmentation
- Rotating, flipping, noise, random center crop
- Expanded training data and reduced overfitting

ResNet50 & VGG16

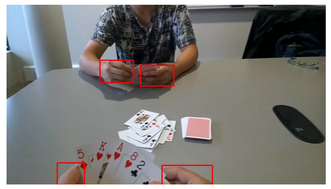
	ResNet50	VGG16
Non-Trainable Parameters	23,508,032	134,260,514
Trainable Parameters	531,226	1,055,514
Avg training time / epoch (s)*	636s / 78s	784s / 101s

*ASL Alphabet dataset and Self-recorded dataset

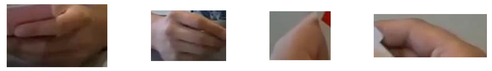
Transfer Learning Retraining
- Two fully connected (FC) layers with ReLU, Dropout = 0.2, Adam optimizer, and log softmax loss.
- 1000 classes ImageNet -> 256 output > 26 ASL classes

Results

Object Detection

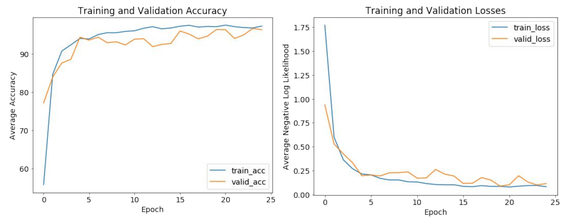


Cropped Images:



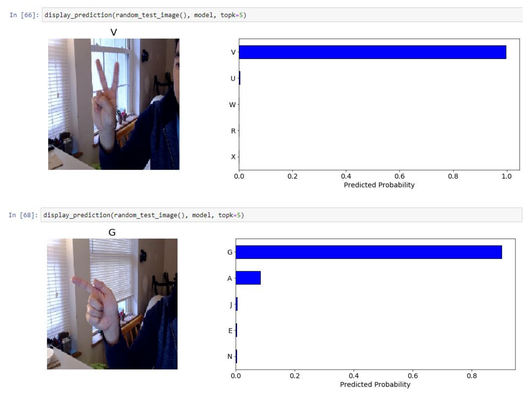
ASL Classification

Model Training:
- On ASL Alphabet data (**Model 1**):
  - 85.23% training accuracy, 78.39% validation accuracy
  - 8 epochs to converge with early stopping
- On self-recorded data (**Model 2**):
  - Best epochs: 97.07% training acc, 96.94% validation acc
  - 24 epochs to converge with early stopping



Test Data (self-recorded images):
- Model 1: 35.42 % accuracy
- Model 2: 95.26 % accuracy

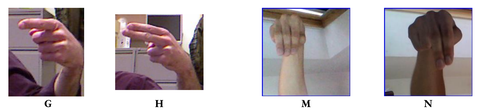
Example Output:



Discussion

Significance of Results

- 95%+ accuracy on 26 classes is sufficient for intelligible video captioning
- Difficult signs to distinguish include **G** and **H**, **M** and **N**



- Model 1 (trained on ASL Alphabet data) struggled with self-recorded test data (~35% accuracy) even with data augmentation and normalization methods, suggesting problems with overfitting and generalizing to new data
- ResNet50 trained faster and resulted in higher accuracies in test data than VGG16. Likely due to reduced number of parameters and avoiding overfitting to the training data

Possible Improvements & Future Work

- Annotate data with bounding boxes to train a Faster R-CNN model, combining hand detection and ASL classification in one
- Inclusion of SPACE and NOTHING classes
- Increase size and diversity of dataset(s) to account for varying lighting conditions and different backgrounds
- Retrain more layers of pretrained model during transfer learning
- Expand work to more signs and gestures outside of fingerspelling

Conclusion

- This proposed ASL fingerspelling classification framework has high performance but additional work should be done to avoid overfitting to the training dataset
- Larger datasets with different lightings and settings are necessary for a robust video captioning system based on RGB data alone
- ASL has many more gestures and complex facial expressions. There is still a long way to go before ASL translation is useful to the deaf community

References

[1] Pugeault, N. Bowden, R. Spelling it out: Real-time ASL fingerspelling recognition. IEEE International Conference, 2011.
[2] Taylor, J et al. Efficient and Precise Interactive Hand Tracking Through Joint, Continuous Optimization of Pose and Correspondences. Microsoft Research, 2018.
[3] Haydar, J. et al. ASL Fingerspelling Translator Glove. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 1, 2012.
[4] Zhang, S. et al. FaceBoxes: A CPU Real-time Face Detector with High Accuracy. arXiv:1708.05234v4, 2017.

ReThink IE