3D Model-Based Data Augmentation for Hand Gesture Recognition

Ben Limonchik, Guy Amdur
Computer Science, Stanford University

Background

Sébastien Marcel's paper:
- Hand segmentation using a space discretisation based on face location and body anthropometry
- Fully Connected Neural Network
- **93%** accuracy on uniform background, **84.4%** accuracy on uniform background

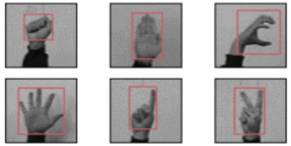

Figure 1: Marcel's method

Problem Statement

The dataset used in Marcel's paper is relatively small in size and not generalized to linear transformations, complex background and different lighting. In this paper we explore how new data based on 3D models can augment a dataset to achieve better accuracy.

Datasets

Figure 2.1: Real training images: A, Five, C, V

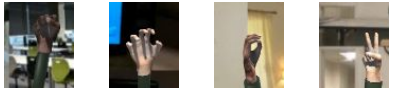

Figure 2.2: Virtual training images: A, Five, C , V

Gesture	Real Training	Virtual Training	Testing (Uniform)	Testing (Complex)
A	1329	1329	58	39
B	487	487	61	41
C	572	572	65	47
Five	654	654	76	58
Point	1395	1395	65	54
V	435	435	57	38
Total	**4872**	**4872**	**382**	**277**

Real data:
4872 training images were collected with a variety of backgrounds. There are six different types of gestures. The testing set is divided into images with uniform white background and images with a complex background making the classification problem harder

Virtual data:
We generated a virtual set of training images using Unity. For every real training image we generated an equivalent virtual image

Methods

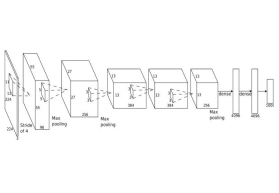

Figure 3: CNN Architecture

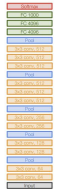

Figure 4: VGG-16

Figure 5: Inception-ResNet-v2

- **Figure 3:** Customized CNN architectures differing in the type and number of layers chained together followed by a softmax loss to output class scores, i.e: [conv-relu-bn-pool]x2 → [affine]x1 → [softmax]
- **Figure 4:** VGG is a 16-layer CNN architecture developed at the University of Oxford in 2014. We added another fully-connected layer at the end of the network to output the desired number of classes. We also transferred learning from an existing model and fine-tuned its hyperparameters.
- **Figure 5:** The Inception-ResNet-v2 network is a CNN architecture developed at Google in 2016. The network combines the residual connections and the latest version of Google's Inception architecture. We transferred learning from an existing model and fine-tuned its hyperparameters.

Experimental Findings

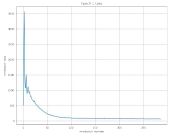

Figure 6: Loss per minibatch

Figure 7: Weights visualization of layer1 of

Losses: To visualize the learning of the best 2conv layer model, we graphed the loss against the number of iterations. To reduce the magnitude of the spikes we introduced L2 regularization and larger batch sizes. However, due to the small dataset, the variability of the losses couldn't be completely eliminated.

Weight Visualization: To visualize the learning we printed the weights of the the 64 filters of the first layer. Some of the filters show more intense colors around the center of the weighted-image which could correlate with the position of hands in the training images

Architecture	Testing accuracy
[FC]-[FC]-[Softmax]	18.2%
[Conv-Relu]-[FC]x2-[Softmax]	37.9%
[Conv-Relu]x2-[FC]x3-[Softmax]	52.3%
[Conv-Relu]x3-[FC]x3-[Softmax]	41.2%
[Conv-Relu]x4-[FC]x1-[Softmax]	39.2%
[Conv-Relu-BN-pool]x2-[FC]-[Softmax]	56.9%
[Conv-Relu-BN-dropout-pool]x2-[FC]-[Softmax]	46.0%

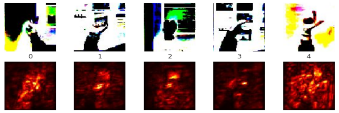

Figure 3: Saliency map of 5 examples

Architecture design:
We experimented with a variety of custom network architecture. Due to the small data set size we tried to minimize the number of fully connected layers and increase the number of Conv layers. Nevertheless, we found that two Conv layers yielded the best accuracies for our data.

Hyperparameter tuning:
Epochs: 15
Regularization: L2 0.003
Learning rate : 0.001

saliency maps:
generating Saliency maps of our best custom model confirmed that hand region was the most important region for the classification decision.

Analysis and Future Work

Method	Training set	Testing set	Accuracy
VGG-16	**Real**	**Uniform**	0.800
		Complex	0.501
	Virtual	**Uniform**	0.620
		Complex	0.450
	Combined	**Uniform**	0.642
		Complex	0.624
Inception-ResNet-v2	**Real**	**Uniform**	0.820
		Complex	0.848
	Virtual	**Uniform**	0.938
		Complex	**0.966**
	Combined	**Uniform**	**0.962**
		Complex	0.958

- We surpassed Marcel's benchmark with 96% on both uniform and complex sets
- 3D model-based data can augment datasets to improve CNN models.
- Future work includes training larger amount of virtual data on this model to see if we can improve the model even more, and test virtual data augmentation on new, more complex datasets.