Sign Language Translation Using Ladder Networks

Guanzhou Ye
markye@stanford.edu

Introduction

It is time consuming to label data for image classification, especially for sign language translation where each word is a separate model class. My project demonstrates how convolutional ladder networks can perform this task using 80% fewer labelled data than traditional models.

Input: images of hand signs
Output: one-hot encoded vector specifying the translated word

Results: ladder network achieved superior performance while using 80% less data, compared to a baseline CNN

Data

Data source: Kaggle

Data are greyscale and normalized 28x28 images of sign language alphabet

I remove most of the labels, as goal of this experiment is to use as few labels as possible

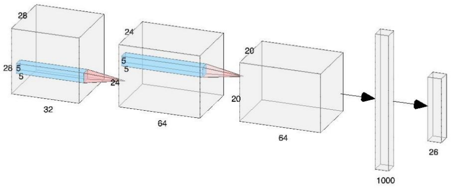
Features

Image recognition problem, therefore image pixels become the features

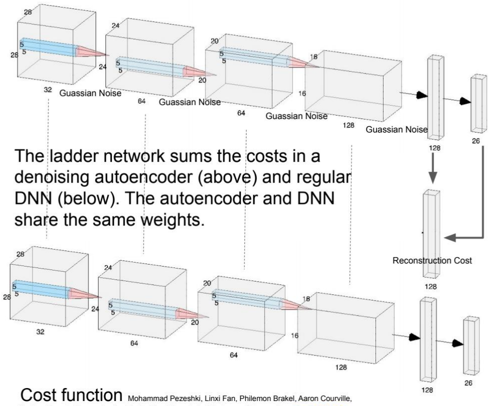
784 features (28x28 image), each feature is the greyscale pixel value (originally from 0-255, normalized to 0-1)

Models

Baseline CNN (used for comparison):



Ladder Network:



The ladder network sums the costs in a denoising autoencoder (above) and regular DNN (below). The autoencoder and DNN share the same weights.

Cost function Mohammad Pezeshki, Linxi Fan, Philemon Brakel, Aaron Courville, Yoshua Bengio. Deconstructing the Ladder Network Architecture. *arXiv preprint* arXiv:1511.06430, 2015.

$$Cost = -\Sigma_{n=1}^{N} \log P\big(\tilde{y}(n) = y^{*}(n)|x(n)\big) + \Sigma_{n=N+1}^{M} \Sigma_{l=1}^{L} \lambda_l \, \text{ReconsCost}(z^{(l)}(n), \hat{z}^{(l)}(n))$$

Results

Architecture	# of Labels Per Class	Accuracy	Precision	Recall
Baseline CNN	10	0.5297	0.5313	0.5114
Ladder network		0.6606	0.6382	0.6481
Baseline CNN	40	0.7379	0.7573	0.7496
Ladder network		0.8447	0.7874	0.8021
Baseline - CNN	70	0.7799	0.7666	0.7679
Ladder network		0.8898	0.8256	0.8436
Baseline - CNN	100	0.8118	0.8030	0.8000
Ladder network		0.9274	0.8801	0.8940
Baseline - CNN	200	0.8302	0.8190	0.8134
Ladder network		0.9533	0.9048	0.9166

Discussion

The ladder network required only 40 labels to achieve a higher accuracy than the CNN did using 200 labels (reduction of 80%). Overall, the ladder network outperformed the baseline CNN by at least 15%. This is expected given that the ladder network uses the denoising autoencoder to learn weight representations that better generalize to all data.

Future

- Expand the number of supported words
- Build an RNN ladder network for translating videos
- Productionization - build image localization to identify signs in a larger image and classify them in real time

Link to presentation video:
https://drive.google.com/file/d/1r7Y7oS3DPbva2EOL4bVYBqXOIPlAEUSc/view?usp=sharing