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:

\[
\begin{align*}
\text{Cost} &= -\sum_{n=1}^{N} \log P(y(n) = y^*(n) | x(n)) + \\
&\quad \sum_{m=1}^{M} \gamma_{m,n} \text{ReconsCost}(z_{m}(n), z_{m}^{(0)}(n))
\end{align*}
\]

Results

<table>
<thead>
<tr>
<th>Architecture</th>
<th># of Labels Per Class</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline CNN</td>
<td>10</td>
<td>0.5297</td>
<td>0.5313</td>
<td>0.5114</td>
</tr>
<tr>
<td>Ladder network</td>
<td>60</td>
<td>0.6006</td>
<td>0.6392</td>
<td>0.6481</td>
</tr>
<tr>
<td>Baseline CNN</td>
<td>40</td>
<td>0.7379</td>
<td>0.7573</td>
<td>0.7496</td>
</tr>
<tr>
<td>Ladder network</td>
<td>80</td>
<td>0.8447</td>
<td>0.7674</td>
<td>0.8021</td>
</tr>
<tr>
<td>Baseline - CNN</td>
<td>70</td>
<td>0.7799</td>
<td>0.7666</td>
<td>0.7679</td>
</tr>
<tr>
<td>Ladder network</td>
<td>80</td>
<td>0.8898</td>
<td>0.8256</td>
<td>0.8436</td>
</tr>
<tr>
<td>Baseline - CNN</td>
<td>100</td>
<td>0.8118</td>
<td>0.8030</td>
<td>0.8000</td>
</tr>
<tr>
<td>Ladder network</td>
<td>100</td>
<td>0.9274</td>
<td>0.8801</td>
<td>0.8940</td>
</tr>
<tr>
<td>Baseline - CNN</td>
<td>200</td>
<td>0.8302</td>
<td>0.8190</td>
<td>0.8134</td>
</tr>
<tr>
<td>Ladder network</td>
<td>200</td>
<td>0.9533</td>
<td>0.9048</td>
<td>0.9166</td>
</tr>
</tbody>
</table>

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/1Y7Z8fAaEDo/bYv3y3xpXovUx.webp/view?usp=sharing