CMPS 240

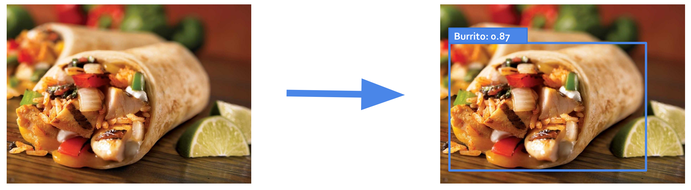
Object Classification and Localization on ImageNet

Matthew J. Howard · Alex Williamson · Arindam Sarma

University of California, Santa Cruz

Background Information

Motivation:
The motivation for our project stems from the desire to classify an object given any image of the object, along with localizing the object within a bounding box inside of the image.



Detecting objects in such a way is important because it can be applied to many real-world tasks, such as facial recognition and autonomous driving.

Convolutional Neural Networks:
For image detection tasks, nearly all state-of-the-art models utilize Convolutional Neural Networks (CNNs), which are able to spatially reason about relationships between pixels, effectively learning pixel-level features such as edges and shapes without manually encoding features.

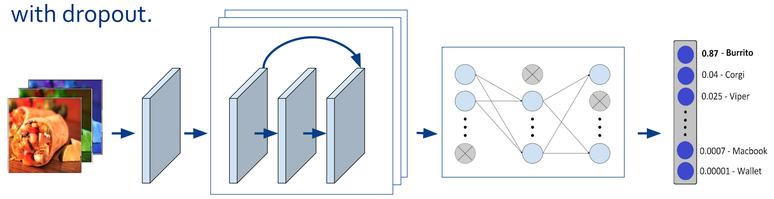
Challenges:
Despite their learning capacity, high-performing CNNs are limited by several factors: Data Quality, Model Architecture, Hardware (GPU, RAM).

Main Contributions:
1) We propose a prototype CNN architecture that combines components of several state-of-the-art models, and
2) Demonstrate the effectiveness of our model on both classification and localization tasks by extending the Single Shot MultiBox Detector (SSD)

Methodology

Neural Network Model
Our prototype architecture combines elements from several models in order to develop a more robust model for both classification and localization.

Specifically, our model consists of 3 stages: 1) a pooled convolutional layer, 2) residual learning blocks, and 3) a small, fully connected network with dropout.



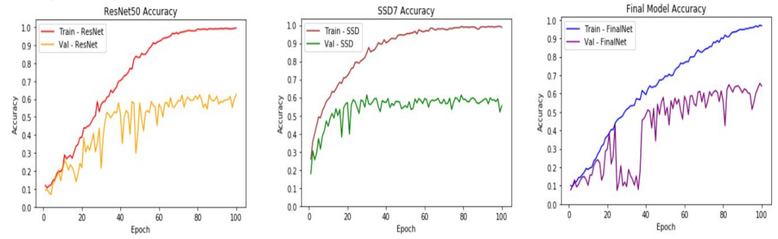
Stage 1 - Pooled Convolutional Layer
Convolutional layers effectively learn filters that activate when detecting spatial features of the input, and pooling down-samples the representation to reduce its size to avoid overfitting and improve robustness.

Stage 2 - Residual Learning Blocks
Residual Learning blocks were introduced by the ResNet model, and address the *vanishing gradient* problem by adding shortcut connections around convolutional blocks to allow gradients to flow back to the input layer with less impedance.

Stage 3 - Fully Connected Network with Dropout
The final (small) fully connected dropout network combines higher-level convolutional features to more accurately predict class labels without adding high complexity to preserve model generalizability.

Results and Analysis

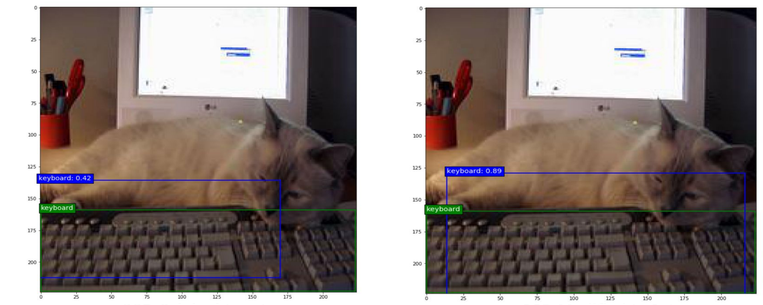
Image Classification Accuracy



Top-1 accuracy on the 25-class ImageNet data after 100 training epochs.

Our model exhibits high variance in early stages, but achieves higher accuracy with a less overfitting versus ResNet50 and SSD7.
- ➔ Top-1 accuracy: **66%,** 61% (SSD7), 63% (ResNet50).
- ➔ Top-5 accuracy: 91%, 90% (SSD), **93%** (Res)

Object Localization Performance



SSD Original (0.42) **SSD Modified (0.89)**

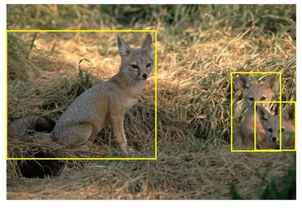
Predictions with class confidence values from a localization model trained on the 25-class data sample.
Blue: Predictions, Green: Hand Labeled.

Our model provides similar performance in terms of bounding box tightness; however, we are able to improve confidence in the predicted class value.

Dataset: **ImageNet**

ImageNet is a collection of quality-controlled, human-annotated images that maps concepts (synonyms and related words) from WordNet to corresponding images. In addition to labeled images, ImageNet also provides annotated bounding boxes on images.

For this project, we use a sample of **12,960** annotated images across **25 classes** (concepts) to train a prototype model for the tasks of classification and localization.



Experimental Setup

Classification: We compare our proposed model against two baselines - a SSD7 base classifier and ResNet50 - and evaluate validation accuracy.

Localization: We compare a Single Shot MultiBox Detector (SSD7) model against a modified SSD7 model that uses our proposed architecture as the base classifier network and showcase localization.

For both tasks, we split our data 80/20 for training/validation.

Hardware: We ran our tests over 100 epochs using an Nvidia 980TI GPU with 6GB of vRAM, coupled with 16GB of system RAM.

Conclusions and Future Work

Our prototype architecture exhibited marginal gains on the 25-class image classification task over the tested SSD7 and ResNet50 baselines, but shows potential for longer training durations, as exhibited by the high classification accuracy variance at early stages and upward trend at later epochs. Localization experiments validate the improvement in classification confidence; however, more validation is pending.

In the future, we plan to test our model on the full ImageNet (1000 classes) and allow longer training duration to study model generalizability.