CNN Image Recognition Architecture Simplification using Patch-Based Data Reduction Techniques

Jiying Zou jiyingz@stanford.edu

Rui Yan ruiyan@stanford.edu

Yuan Liu linda921@stanford.edu

Motivation

Problem

Traditional CNN image recognition methods are both time- and memory-consuming. State-of-the-art architectures such as ResNet set benchmark accuracies but often do not optimize for reducing computational intensity.

Question

- How can we modify CNN architecture in ways that will lead to a reduced need for computational resources?
- How will such network architecture modifications affect model performance?

Data / Features

Source Kaggle Flowers Recognition Competition

Description

- 5 classes: daisy, dandelion, rose, sunflower, tulip
- 4242 images total, roughly 800 images/class

Preprocessing (Features)

- Color images are randomly cropped (256x256 pixels) and normalized

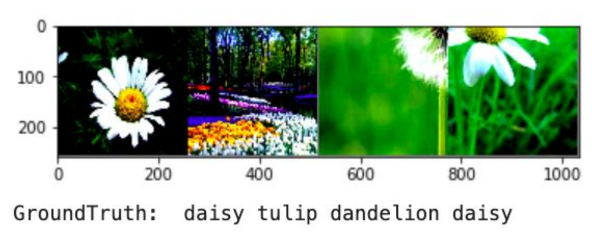


Fig1. Transformed data and ground truth labels

Methods

Baseline Models

1. ResNet-50: *50 layers*
2. BagNet-33: *Linear aggregation of results of independent modified ResNet-50's*
   - Pretrained on ImageNet, last layer modified to fit # classes

Experiments

- Trained BagNet-33 model on full dataset
- Changed very last aggregation layer weights
  - "Blackout" a patch by manually setting its weight to 0

Evaluation

Top 1 Accuracy $= \frac{1}{n}\sum_{i=1}^{n} 1\{top\ pred\ for\ obs\ i =\ ground\ truth\}$

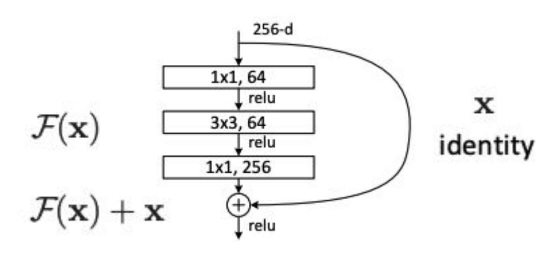
Loss = Average Cross-Entropy = $-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_k^{(i)} \log \hat{y}_k^{(i)}$ for *k* classes

Models

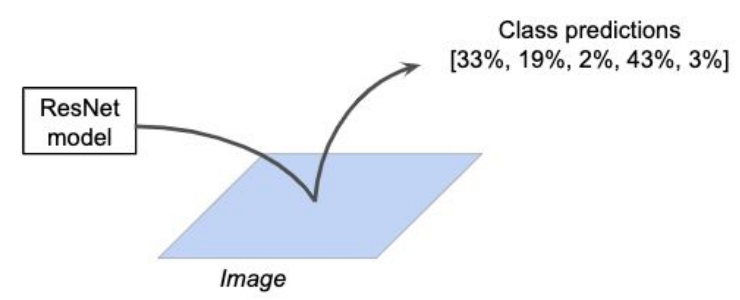
ResNet-50

- State-of-the-art CNN model
- Infers image filters by moving patch-by-patch through image
- Lower layers learn basic structures like edges, curves, and corners
- Deeper layers learn more complicated patterns

Utilizes "identity connections" to fix vanishing gradient issue



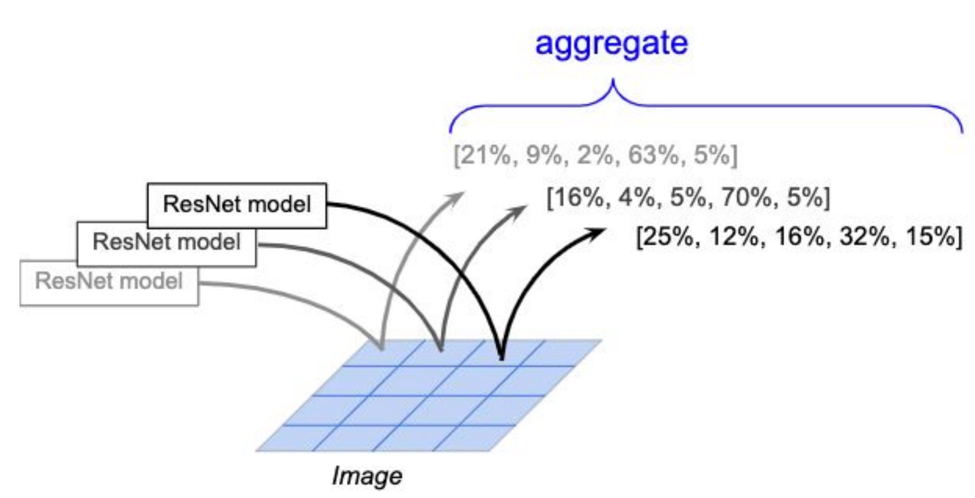
Credit: Image modified from https://arxiv.org/pdf/1512.03385.pdf



BagNet-33

- More computationally efficient by design, architecture simplified without losing much performance
- Runs a modified ResNet-50 (3x3 layers replaced by 1x1) over every patch (33x33 pixels/patch) of image
  - Effectively limits receptive field to ~64 patches
  - Patch results independent of others

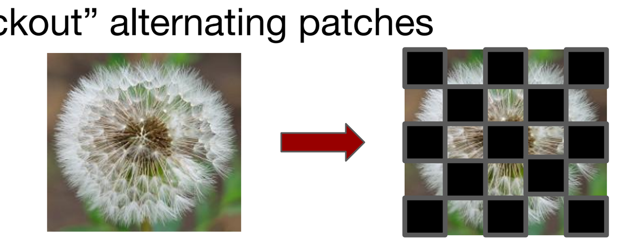
Derives patch-wise class evidence and aggregates (averages) for final prediction



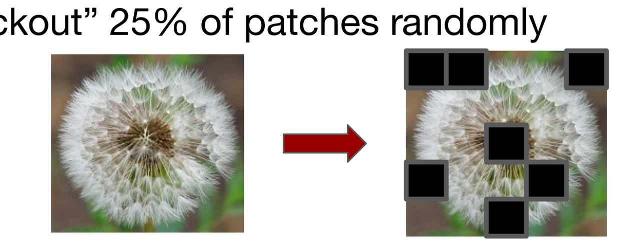
Experiments

- Replace the average aggregation method with different aggregation techniques:

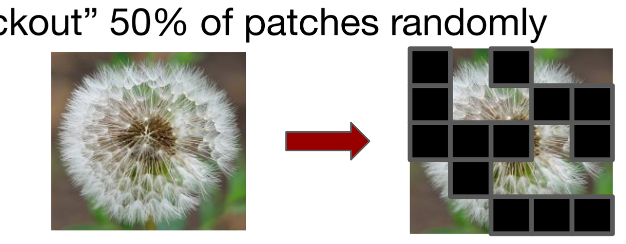
"Blackout" alternating patches



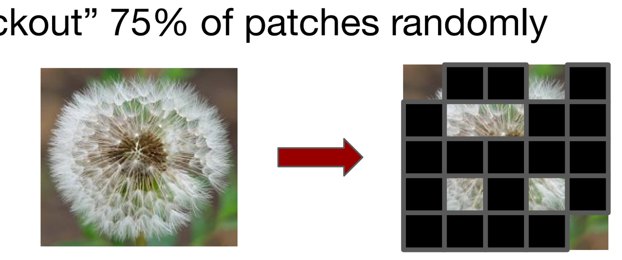
"Blackout" 25% of patches randomly



"Blackout" 50% of patches randomly



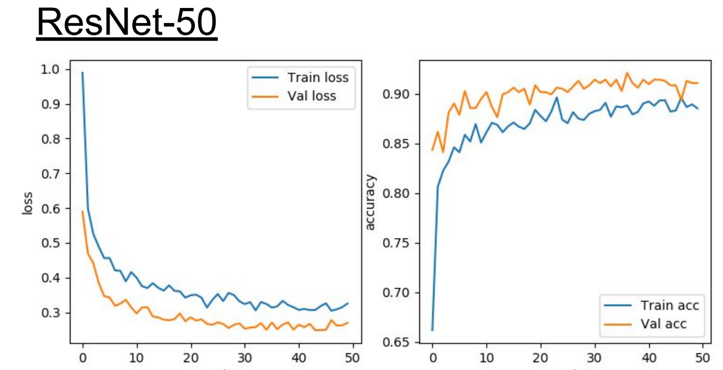
"Blackout" 75% of patches randomly



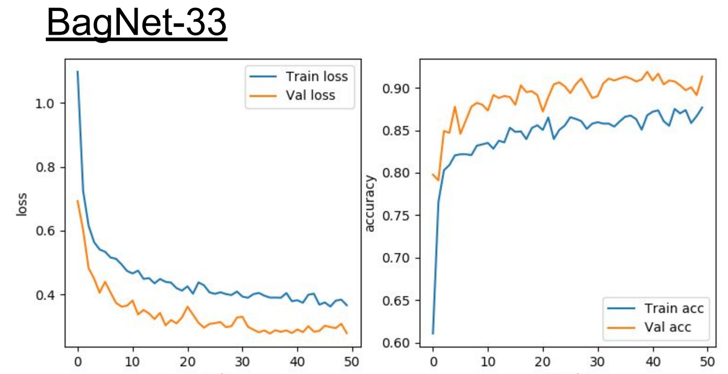
Results

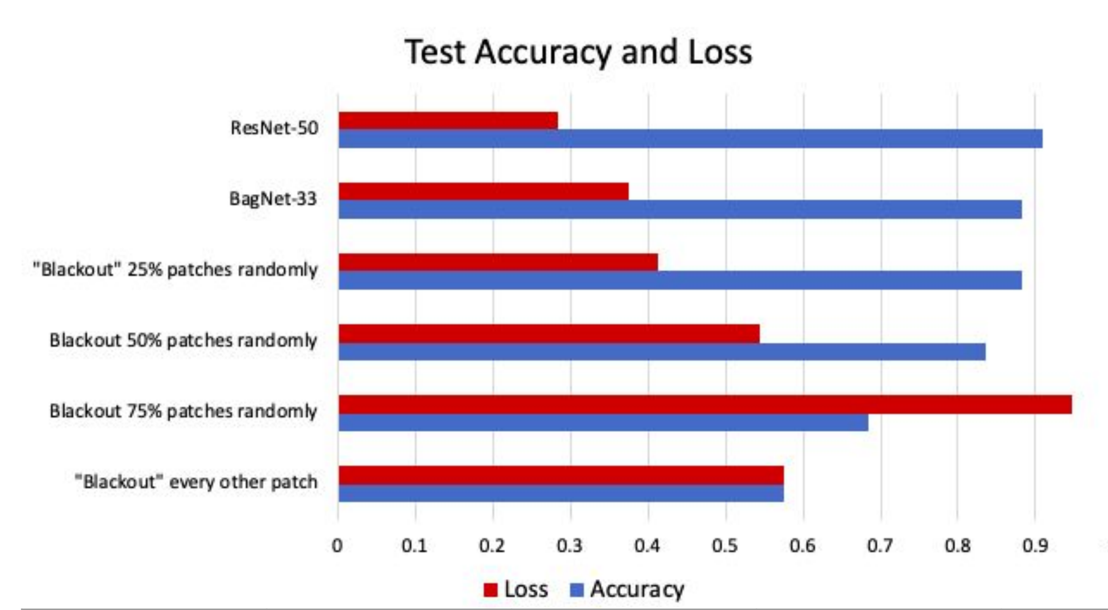
Learning Curves

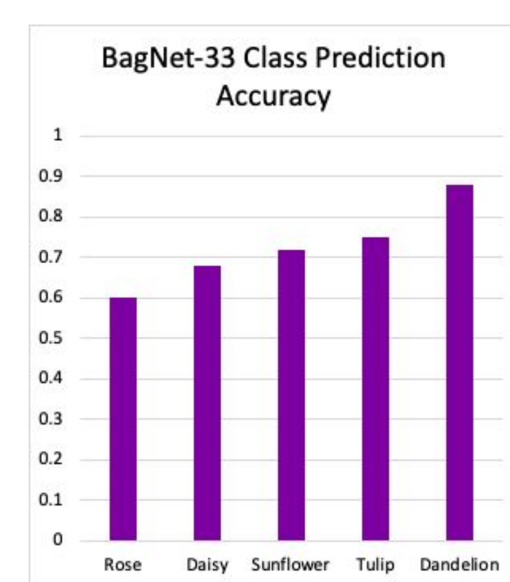
ResNet-50



BagNet-33







	ResNet-50	BagNet-33	Experiment: Alternating "blackout"	Experiment: Random 25% "blackout"	Experiment: Random 50% "blackout"	Experiment: Random 75% "blackout"
Train Acc / Loss	0.9211 / 0.2517	0.9288 / 0.2799				
Test Acc / Loss	0.9093 / 0.2847	0.8821 / 0.3745	0.8293 / 0.5761	0.8827 / 0.4123	0.8327 / 0.5445	0.6853 / 0.9464

Discussion

Main Findings

- It is possible to randomly dispose of ~25% of patch results on an image and maintain prediction accuracy comparable to that of full BagNet-33.
- Dropping out patches causes loss to increase (and accuracy to decrease) in a non-linear fashion.
- Random 50% blackout is comparable to alternating blackout.
- Reproduced BagNet's comparable performance to ResNet, suggesting that CNN might care about patch details/patterns in an order-agnostic way.

Limitations / Challenges

- Dataset size relatively small, concerns about generalizability.
- Model pretrained on ImageNet seems to underfit our dataset.
- Computational limitations:
  - BagNet-33 took ~4.6x the runtime compared to ResNet-50
  - Trained with only 50 epochs and batch size = 32

Future Directions

1. **Generalizability analysis**
   - Try to confirm results using other BagNet models
   - Replicate results on larger datasets (e.g. ImageNet)
   - Stochasticity/robustness analysis on model results
2. **Other aggregation schemes**
   - Weight patch results by proximity to image center
   - Nonlinear aggregations
   - Predict if patch is mainly background, exclude in aggregation
3. **Application**
   - Results are proof-of-concept, how to apply "blackout" idea to save memory and runtime warrants further consideration

Credits / References

[1] Brendel, W. and Bethge, M. *Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet.* ICLR 2019 Conference Paper. Mar 2019.

[2] He, K., Zhang, X., Ren, S. and Sun, J. *Deep Residual Learning for Image Recognition*. IEEE CVPR 2016 Conference Paper. June 2016.

[3] Ramprasaath R. Selvaraju. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. ICCV 2017 Conference Paper. Oct 2016.

- Special thanks to Dr. Avner May and Dr. Jared Dunnmon for introducing the project topic and occasional guidance throughout the workflow.