Xceptional Landmark Recognition

Tyler Yep (tyep@stanford.edu), Heidi Chen (hchen7@stanford.edu)

Problem & Task Definition

The Google Landmark Recognition Challenge asks competitors to classify popular landmarks from a massive dataset of images, with few training examples for any one landmark. Due to the extreme class imbalance and scope of the dataset, such landmark recognition is a difficult problem.

Given a 256x256 RGB image, our task is to output a landmark ID (blank if there is no landmark in the image) as well as a confidence score. For our model, to specify that there is no landmark, we still output a landmark id, but use a confidence of 0.

Dataset & Metric

Dataset: The Landmark dataset [1] contains over 5 million images and over 100,000 unique landmark classes.

- Train: classes with 100+ examples (6512 classes, 1.2 million images)
- Dev: random sample from remaining images in full train set
- Test: withheld stage 2 submission set on official Kaggle competition page

Metric: Global Average Precision (GAP). Given a list of predicted landmark labels and confidence scores, the evaluation takes a weighted average over the landmarks:

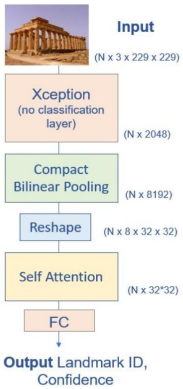
$$GAP = \frac{1}{M}\sum_{i=1}^{N} P(i)rel(i)$$

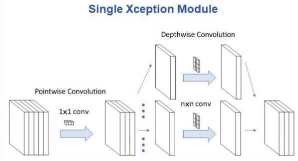
Main Approach

Our final model architecture consists of four major components:

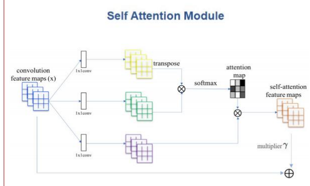
1. **Xception Network**
   - Depthwise Separable Convolutions and Residual Modules for high baseline performance that cuts down on computation and parameters [2]
2. **Compact Bilinear Pooling**
   - Encodes second order feature statistics by calculating outer products of Xception output vectors
   - Minimizes computational costs with Count Sketch dimension reduction [3]
3. **Soft Self-Attention**
   - Performs 1D Convolutions on three branches of input maps
   - Outputs sum of scaled attention and original input map
   - Captures global dependencies by eliminating padding and adjusting for earlier minimal kernel sizes [4]
4. **Fully Connected Layer w/ Softmax**
   - Shrinks or expands final representation into shape (num_classes, 1) and finds the most likely landmark id.

Final Model Architecture



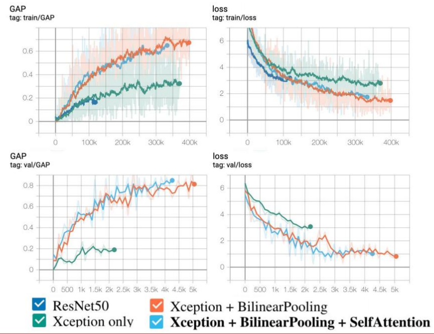


[2]



[4]

Training & Development Set Curves

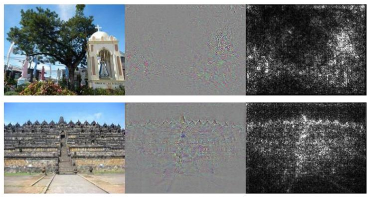


Results & Analysis

Model Metrics on Dev Set	GAP	Loss
ResNet50	0.241	3.183
Xception only	0.674	1.301
Xception + SpatialAttention	0.1188	3.079
Xception + BilinearPooling	0.812	0.908
Xception + BilinearPooling + SelfAttention	**0.841**	**0.932**

- All Xceptions performed as well or significantly better than ResNet baseline
- Xception + Bilinear Pooling performed best, with or without Self-Attention

Saliency Maps (Xception)



Attention likely added little to performance because Xception alone already heavily encodes spatial features, as depicted.

Future Work

- More advanced attention models and/or concatenated attentions
- Confidence reranking algorithms to maximize GAP
  - Spatial feature matching using Google's Deep Local Features (DeLF)
  - Fast Nearest Neighbors search using Faiss algorithm
- Increased model complexity (additional parameters, etc.)
- Miscellaneous: indoor/outdoor filtering, training on more classes

References

[1] Bor-Chun Chen and Larry Davis. Deep representation learning for metadata verification. IEEE Winter Applications of Computer Vision Workshops, 2019.

[2] François Chollet. Xception: Deep learning with depthwise separable convolutions. CoRR , abs/1610.02357, 2016.

[3] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. CoRR, abs/1511.06062, 2015.

[4] Dimitris Metaxas Augustus Odena Han Zhang, Ian Goodfellow. Self-attention generative adversarial networks. 2019.

Xceptional Landmark Recognition

Tyler Yep (tyep@stanford.edu), Heidi Chen (hchen7@stanford.edu)

Problem & Task Definition

The Google Landmark Recognition Challenge asks competitors to classify popular landmarks from a massive dataset of images, with few training examples for any one landmark, which is necessary for image captioning or geotagging.

Given a 256x256 RGB image, our task is to output a landmark id (blank if there is no landmark in the image) as well as a confidence score. For our model, to specify that there is no landmark, we still output a landmark id, but use a confidence of 0.

Dataset & Metric

Dataset: The Landmark dataset [1] contains over 5 million images and over 100,000 unique landmark classes.

- Train: classes with 100+ examples (6512 classes, 1.2 million images)
- Dev: random sample from remaining images in full train set
- Test: withheld stage 2 submission set on official Kaggle competition page

Metric: Global Average Precision (GAP). Given a list of predicted landmark labels and confidence scores, the evaluation takes a weighted average over the landmarks:

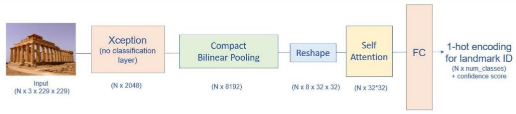
$$GAP = \frac{1}{M}\sum_{i=1}^{N} P(i)rel(i)$$

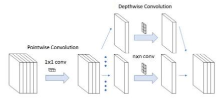
Main Approach

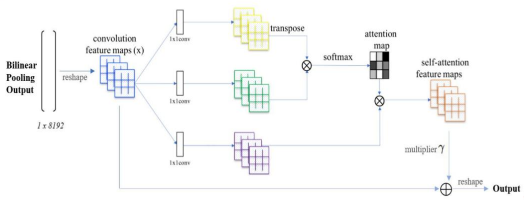
Our final model architecture consists of four major components:

1. **Xception Network**
   - Depthwise Separable 2D Convolutions and Residual Modules for high baseline performance that cuts down on computation and parameters [2]
2. **Compact Bilinear Pooling**
   - Encodes second order feature statistics by calculating outer products of Xception output vectors
   - Minimizes computational costs with Count Sketch dimension reduction [3]
3. **Soft Self-Attention**
   - Performs 1D Convolutions on three branches of input maps
   - Outputs sum of scaled attention and original input map
   - Captures global dependencies by eliminating padding and adjusting for earlier minimal kernel sizes [4]
4. **Fully Connected Layer w/ Softmax**
   - Shrinks or expands final representation into shape (num_classes, 1) and finds the most likely landmark id.

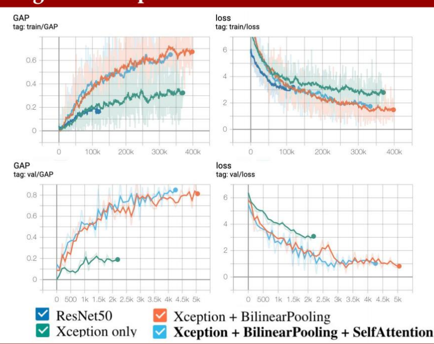
Final Model Architecture







Training & Development Set Curves

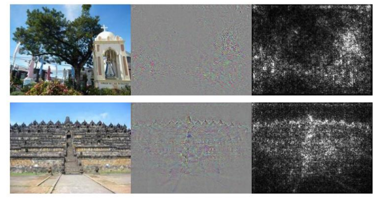


Results & Analysis

Model Metrics on Dev Set	GAP	Loss
ResNet50	0.241	3.183
Xception only	0.674	1.301
Xception + SpatialAttention	0.1188	3.079
Xception + BilinearPooling	0.812	0.908
Xception + BilinearPooling + SelfAttention	**0.841**	**0.932**

- Xception + Bilinear Pooling performed best, with or without Self-Attention
- All Xceptions perform as well or significantly better than ResNet baseline

Saliency Maps (Xception)



Attention likely added little to performance because Xception alone already heavily encodes spatial features, as depicted.

Future Work

- More advanced attention models and/or concatenated attentions
- Confidence reranking algorithms to maximize GAP
  - Spatial feature matching using Google's Deep Local Features (DeLF)
  - Fast Nearest Neighbors search using Faiss algorithm
- Increased model complexity (additional parameters, etc.)
- Miscellaneous: indoor/outdoor filtering, training on more classes

References

[1] Bor-Chun Chen and Larry Davis. Deep representation learning for metadata verification. IEEE Winter Applications of Computer Vision Workshops, 2019.
[2] François Chollet. Xception: Deep learning with depthwise separable convolutions. CoRR , abs/1610.02357, 2016.
[3] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. CoRR, abs/1511.06062, 2015.
[4] Dimitris Metaxas Augustus Odena Han Zhang, Ian Goodfellow. Self-attention generative adversarial networks. 2019.