# Scene Classification with Residual CNN's

Josh King, Vayu Kishore, Filippo Ranalli

Problem Statement

Given a picture of a scene, determine which of 365 classes it belongs to.
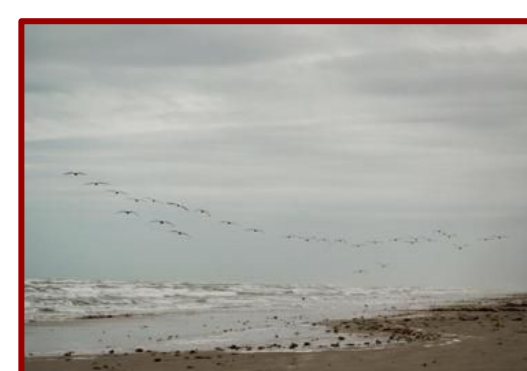
Dataset

- 365 classes
- 5000 training examples per class
- 100 validation examples per class
- 900 test examples per class
- Resolution of 256 x 256
  - We scale down to 64 x 64 due to memory limitations
- Extra "Challenge" data exists
  - 25,000 training examples per class
- Higher Resolution images exist
  - 512+ x 512+
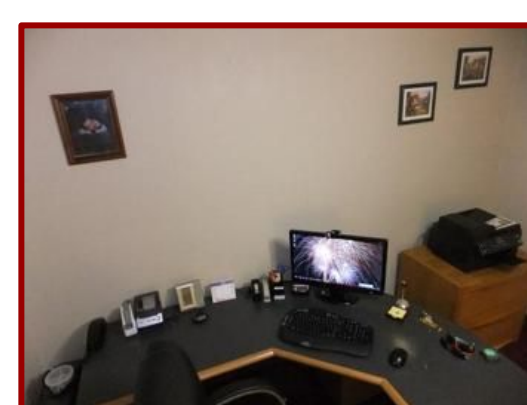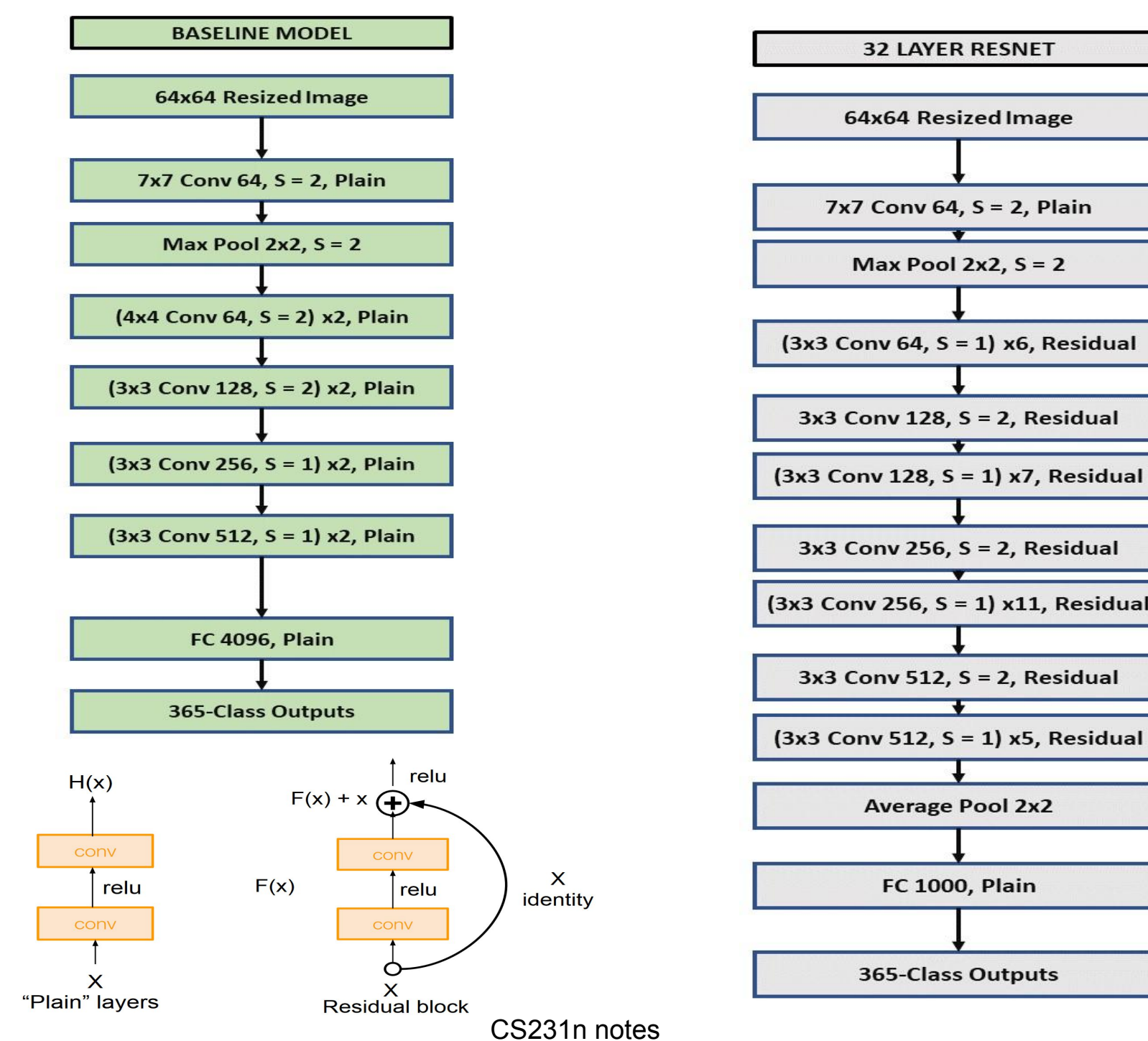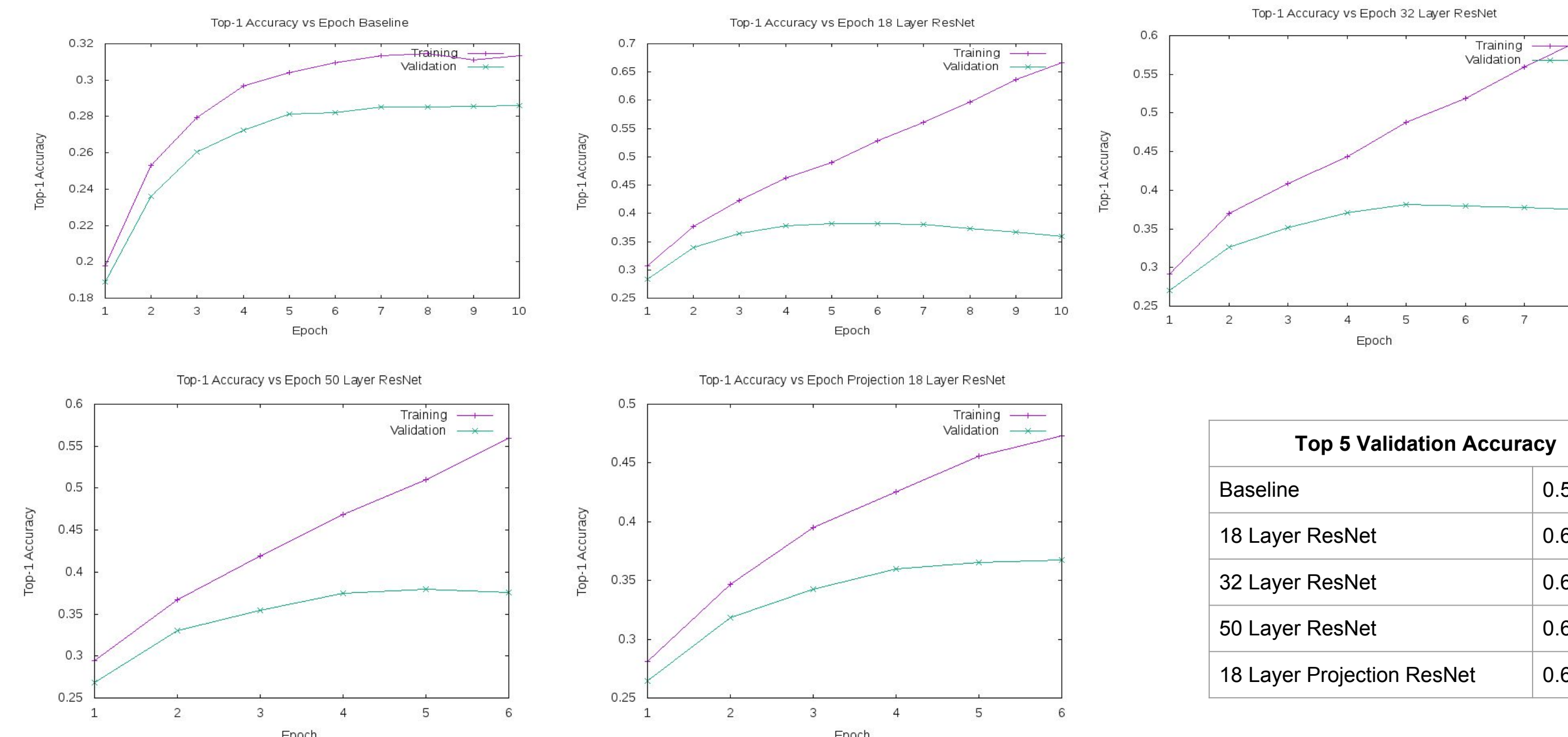

Home Office

Archaeological Excavation

Beach

Sandbar

References

- CS231n 2017 class notes.
- He, Zhang, Ren, Sun, *"Deep Residual Learning for Image Recognition"*.
- Zhou, Khosla, Lapedriza, Torralba, Oliva, *"Places: An Image Database for Deep Scene Understanding"*
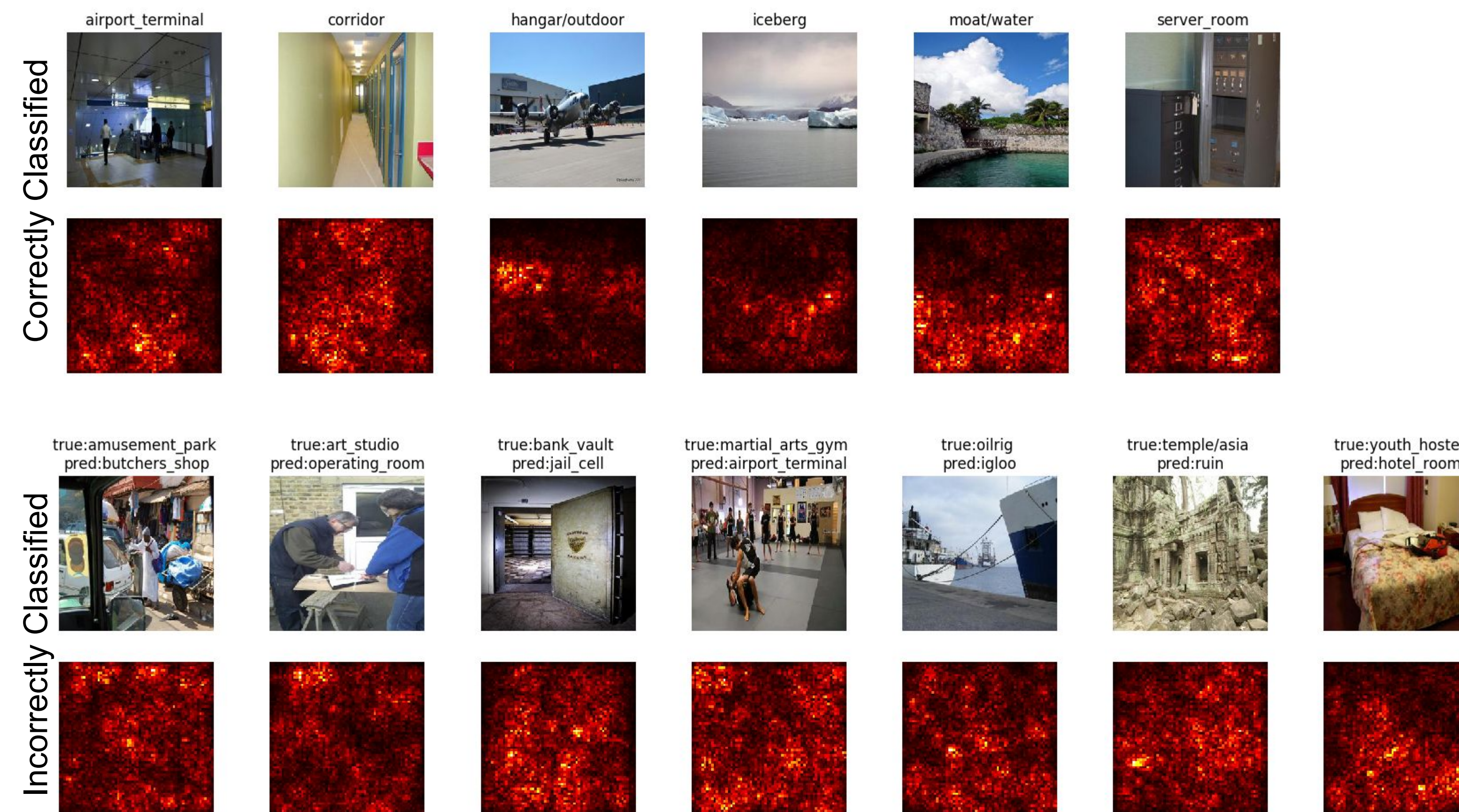- http://places2.csail.mit.edu/

Models

BASELINE MODEL
- 64x64 Resized Image
- 7x7 Conv 64, S = 2, Plain
- Max Pool 2x2, S = 2
- (4x4 Conv 64, S = 2) x2, Plain
- (3x3 Conv 128, S = 2) x2, Plain
- (3x3 Conv 256, S = 1) x2, Plain
- (3x3 Conv 512, S = 1) x2, Plain
- FC 4096, Plain
- 365-Class Outputs

32 LAYER RESNET
- 64x64 Resized Image
- 7x7 Conv 64, S = 2, Plain
- Max Pool 2x2, S = 2
- (3x3 Conv 64, S = 1) x6, Residual
- 3x3 Conv 128, S = 2, Residual
- (3x3 Conv 128, S = 1) x7, Residual
- 3x3 Conv 256, S = 2, Residual
- (3x3 Conv 256, S = 1) x11, Residual
- 3x3 Conv 512, S = 2, Residual
- (3x3 Conv 512, S = 1) x5, Residual
- Average Pool 2x2
- FC 1000, Plain
- 365-Class Outputs

"Plain" layers / Residual block / CS231n notes

Training and Validation Accuracy

Top-1 Accuracy vs Epoch Baseline

Top-1 Accuracy vs Epoch 18 Layer ResNet

Top-1 Accuracy vs Epoch 32 Layer ResNet

Top-1 Accuracy vs Epoch 50 Layer ResNet

Top-1 Accuracy vs Epoch Projection 18 Layer ResNet

Top 5 Validation Accuracy	
Baseline	0.573
18 Layer ResNet	0.676
32 Layer ResNet	0.682
50 Layer ResNet	0.683
18 Layer Projection ResNet	0.673

Example Results and Saliency Maps

Correctly Classified

airport_terminal | corridor | hangar/outdoor | iceberg | moat/water | server_room

Incorrectly Classified

true:amusement_park pred:butchers_shop | true:art_studio pred:operating_room | true:bank_vault pred:jail_cell | true:martial_arts_gym pred:airport_terminal | true:oilrig pred:igloo | true:temple/asia pred:ruin | true:youth_hostel pred:hotel_room

Discussion

- State of the Art classifiers get 90% top 5 accuracy, much higher than our models.
- We attribute this our inability to train on entire higher resolution dataset.
- Learning curves indicate overfitting, even in the presence of strong L2 regularization and dropout.
- Found top-5 val. accuracy to be more informative than top-1 as classes may be ambiguous (hotel vs hostel) and are not mutually exclusive (ex: temple and ruin).
- Based on visual inspection, significant portion of misclassifications seem reasonable.
- Some labelling issues in validation dataset.
- Saliency maps seem generally more diffuse than saliency maps for object detection, indicating that the classification generated by model is sensitive to the entire image. Some exceptions where presence of a particular object defines a scene.