Error Plots for Various Models, Learning Rates, and Activation Functions

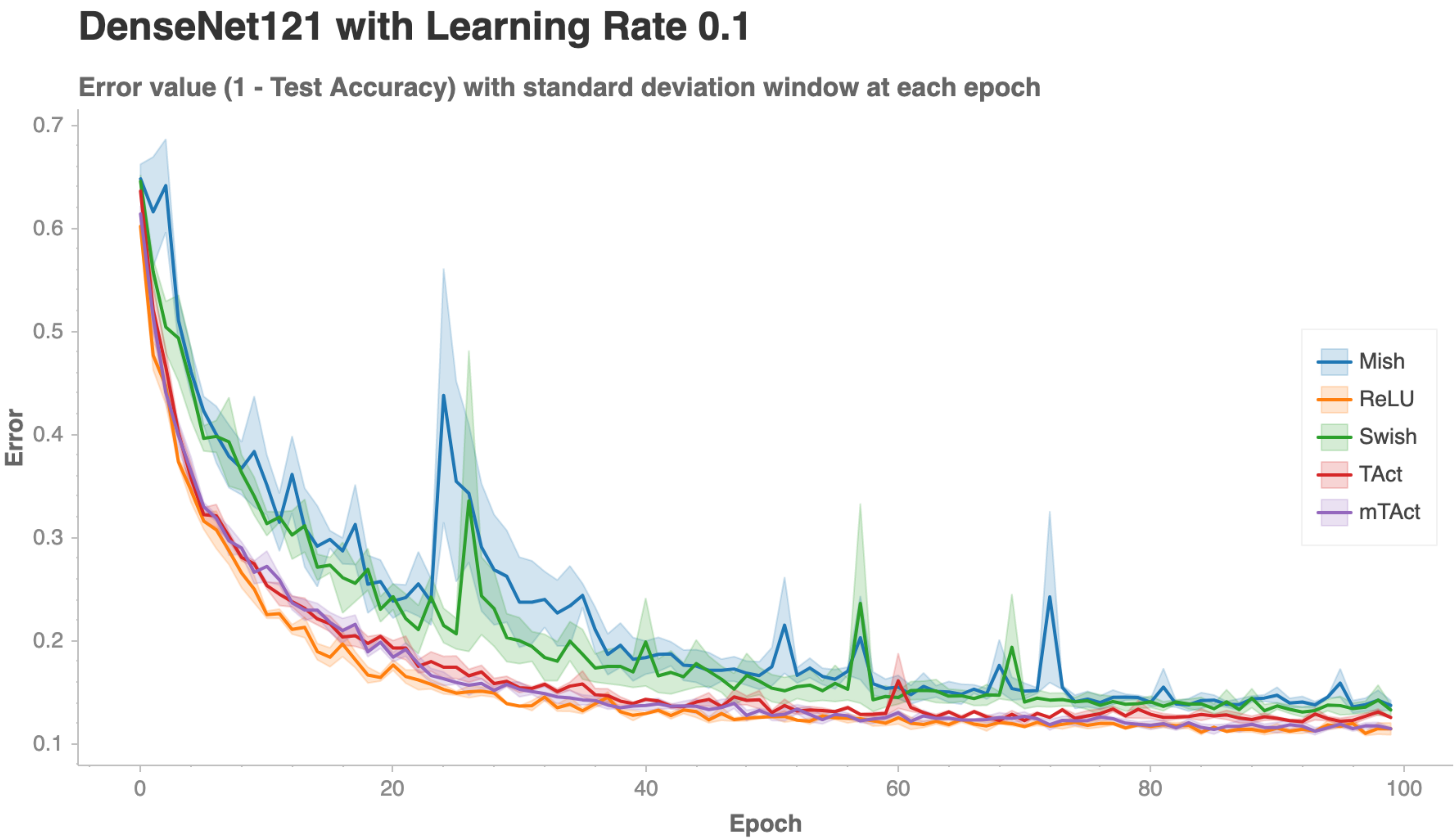
Gobind Puniani

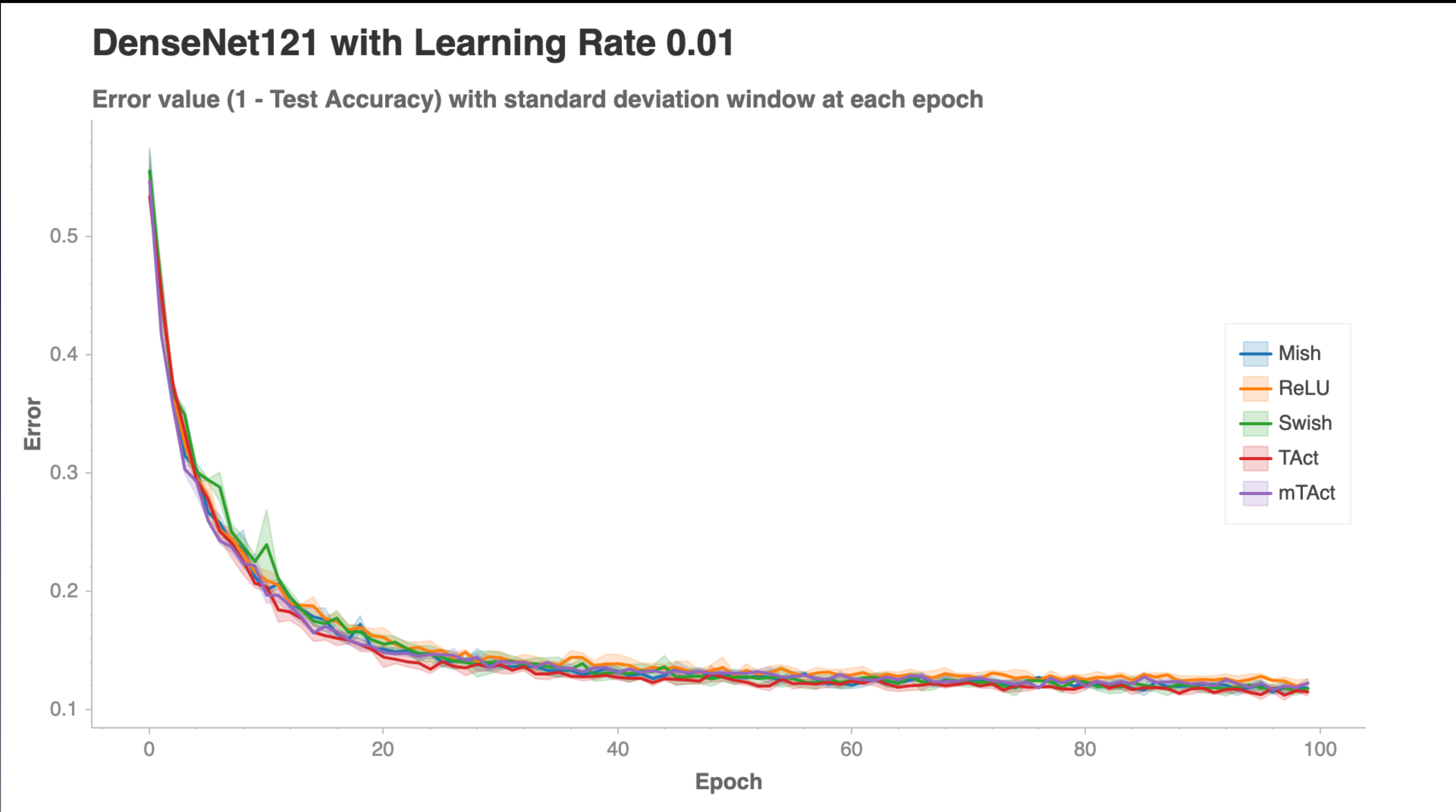
MATH 198: Spring 2020

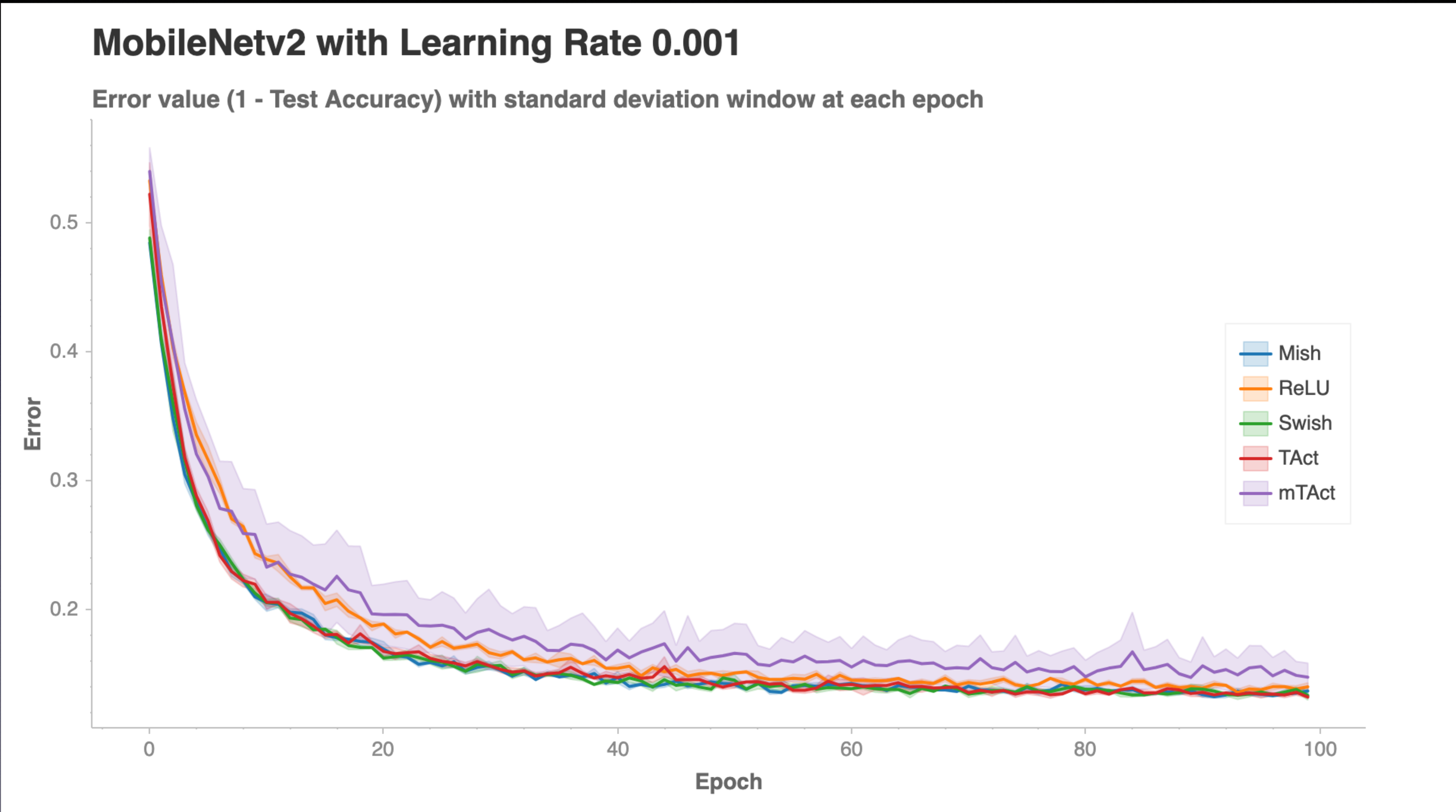
02/20/20

Plots

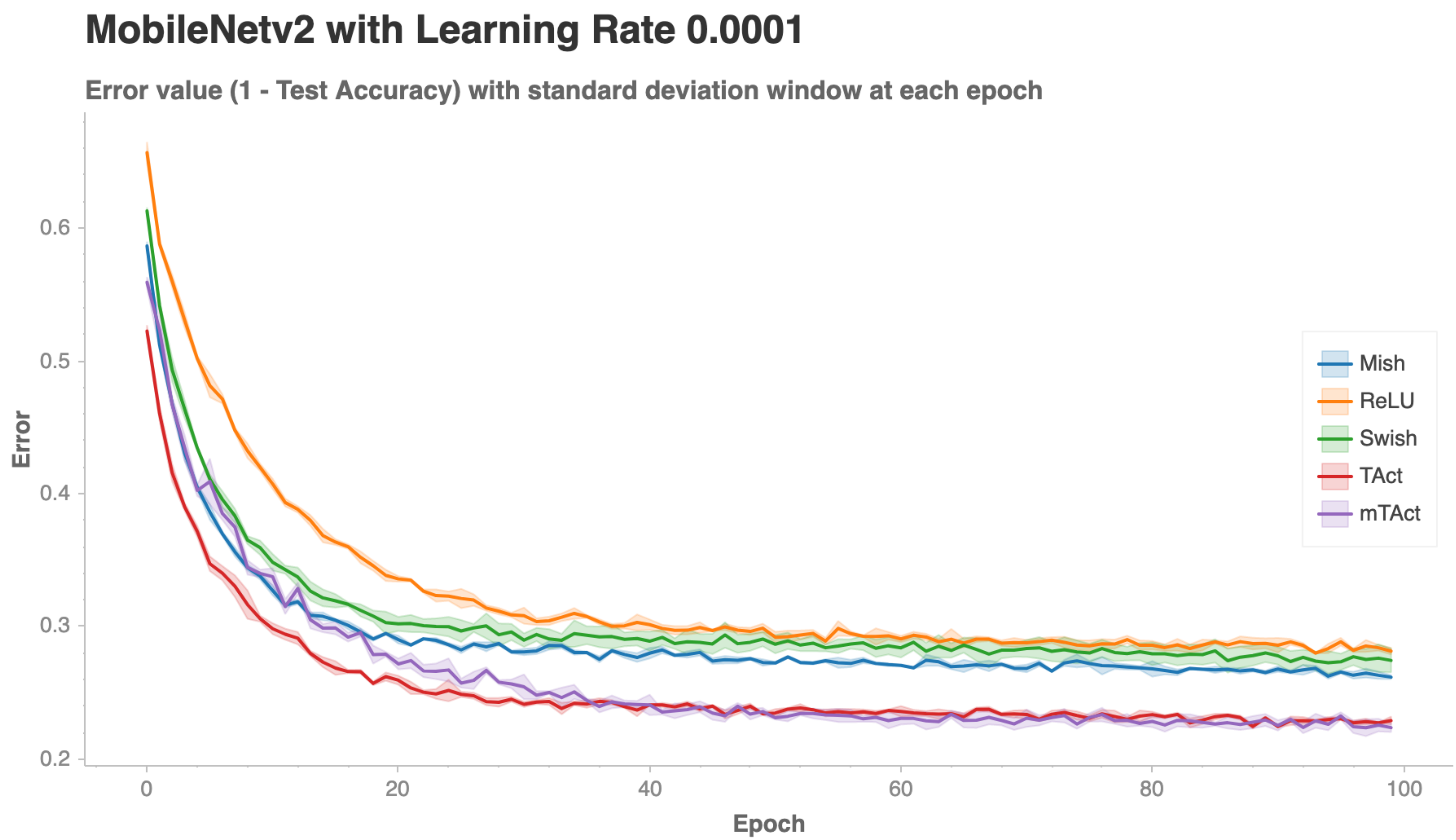
- 6 plots in total
  - Chartify library
  - Test top-1 accuracy at each epoch averaged over 3 runs
  - Error value on y-axis: 1 - (Averaged Test Accuracy)
- 5 activation functions on each plot: ReLU, Swish, Mish, TAct, and mTAct
- DenseNet-121
  - Learning Rates: 0.1 and 0.01
- MobileNetv2 and SE Net-18
  - Learning Rates: 0.001 and 0.0001


DenseNet121 with Learning Rate 0.1
Error value (1 - Test Accuracy) with standard deviation window at each epoch
Error
Epoch
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
40
60
80
100
Mish
ReLU
Swish
TAct
mTAct


DenseNet121 with Learning Rate 0.01
Error value (1 - Test Accuracy) with standard deviation window at each epoch
0.5
0.4
0.3
0.2
0.1
Error
0
20
40
60
80
100
Epoch
Mish
ReLU
Swish
TAct
mTAct

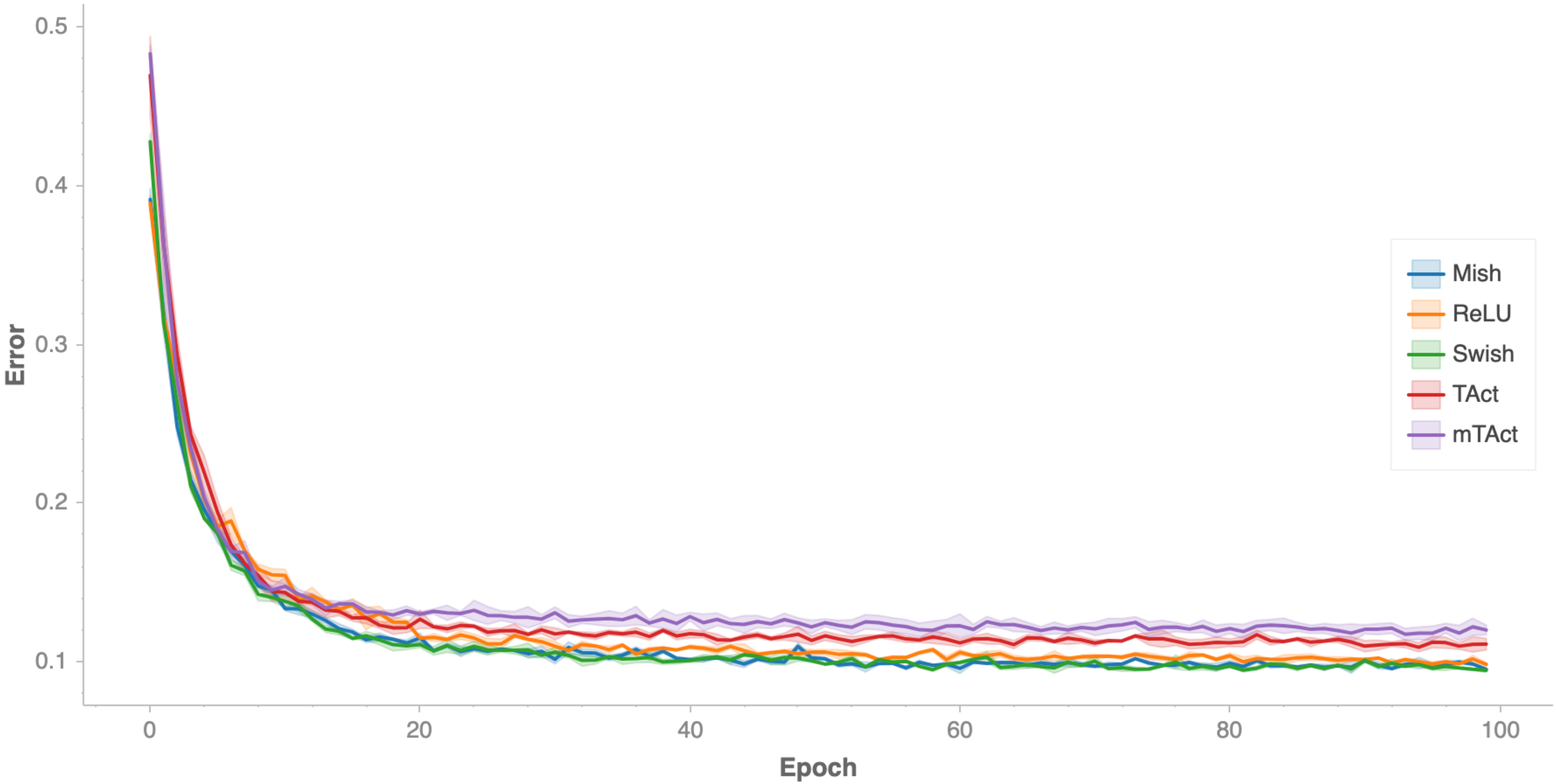


MobileNetv2 with Learning Rate 0.001
Error value (1 - Test Accuracy) with standard deviation window at each epoch
Error
0.5
0.4
0.3
0.2
Epoch
0
20
40
60
80
100
Mish
ReLU
Swish
TAct
mTAct


MobileNetv2 with Learning Rate 0.0001
Error value (1 - Test Accuracy) with standard deviation window at each epoch
Error
0.6
0.5
0.4
0.3
0.2
0
20
40
60
80
100
Epoch
Mish
ReLU
Swish
TAct
mTAct

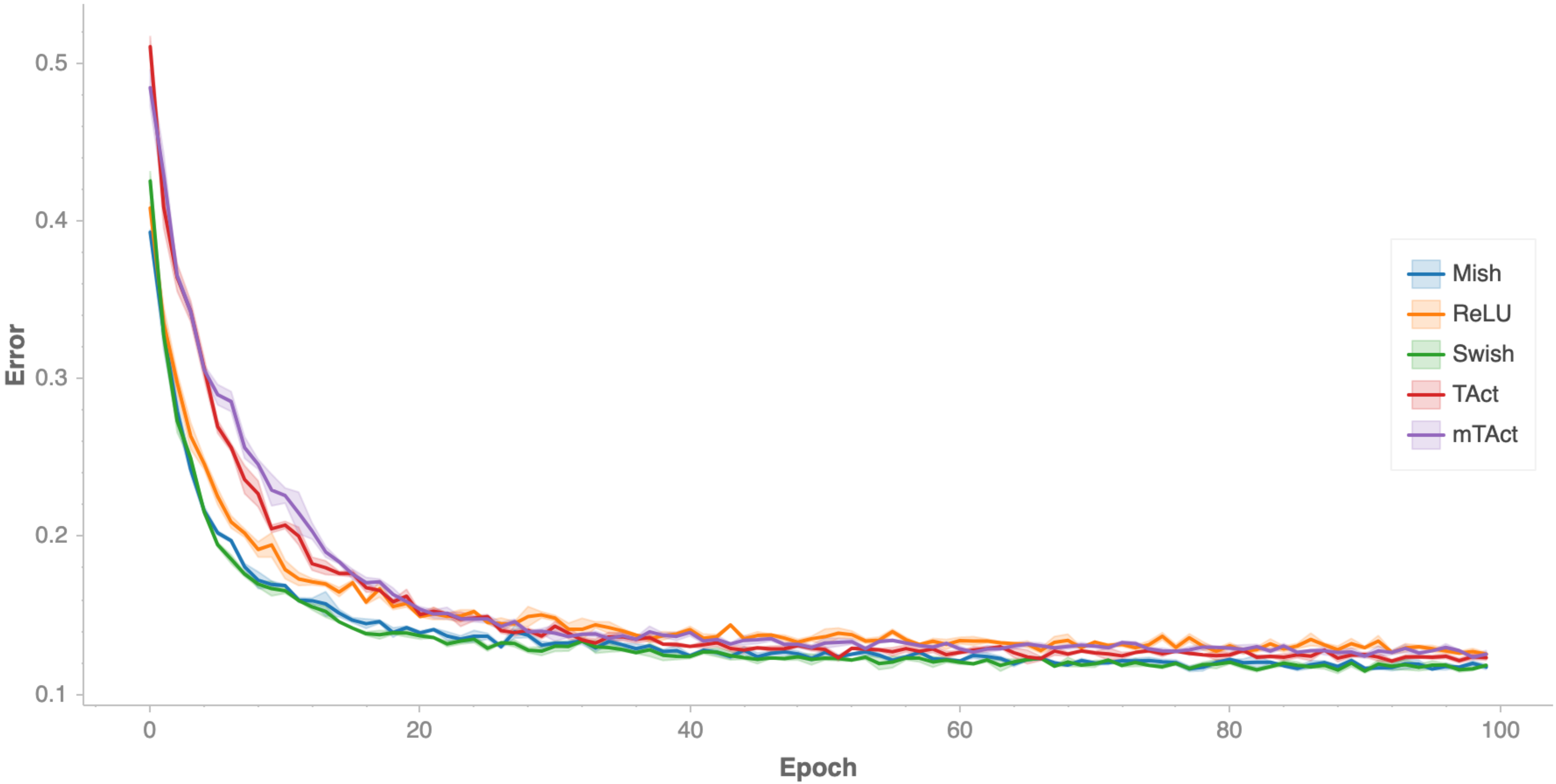
SENet18 with Learning Rate 0.001

Error value (1 - Test Accuracy) with standard deviation window at each epoch



SENet18 with Learning Rate 0.0001

Error value (1 - Test Accuracy) with standard deviation window at each epoch



Number of Parameters

- DenseNet-121: 7.0m parameters
  - https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803
- MobileNet v2: 3.4m parameters
  - https://towardsdatascience.com/review-mobilenetv2-light-weight-model-image-classification-8febb490e61c
- SE Net-18: 25.6m parameters
  - https://towardsdatascience.com/review-senet-squeeze-and-excitation-network-winner-of-ilsvrc-2017-image-classification-a887b98b2883

Squeeze-and-Excitation Networks

Gobind Puniani
MATH 198: Spring 2020
02/20/20

Convolution Operators

- Convolutional Neural Networks (CNNs) are useful for visual tasks
- CNNs apply filters to detect spatial connections along input channels
  - Finding spatial correlations between features strengthens representative power
- Other approaches to new architectures focus on spatial relations
- This paper investigates channel relations instead

Squeeze-and-Excite Blocks

- New architectural unit: Squeeze-and-Excite (SE) block
  - Higher representational quality through modeling interdependencies between channels of convolutional features
  - Learn global information to emphasize informative features and ignore less useful ones

Structure of SE Blocks

- SE blocks perform feature recalibration for a given convolution in two phases: *squeeze* and *excitation*
- Features passed through *squeeze* operation:
    - Channel descriptor aggregates feature maps across spatial dimensions
        - Use global average pooling to squeeze global info onto channel descriptor
            - Other aggregation techniques may be considered

Structure of SE Blocks (Cont.)

- Channel descriptor then passes through *excitation* phase
  - Embedding input and outputs channel weights applied to feature maps
    - Function must be flexible (able to learn non-linear relations) and must be able to learn inclusive relationships to allow emphasis of multiple channels
      - Simple gating mechanism
  - Paired FC (fully-connected) layers reduce and then increase, respectively, the dimensionality
  - Output of SE block fed to subsequent layers

SE Networks

- SE Network (SENet) is simply a collection of SE blocks linked together
- SE blocks can replace other blocks in other architectures
  - Inception, ResNet, ResNeXt, Inception-ResNet, MobileNet, ShuffleNet, etc.
  - Inserted after non-linearity following a convolution
  - Role changes depending on depth placement
    - Excites class-agnostic features in shallower levels
    - More specialized by class in deeper levels
  - Only slight increase in complexity and computational cost
- Explicit modeling of channel interdependencies increases sensitivity to informative features to be exploited in later layers

Model and Computational Complexity

- Accuracy gains offset slight increase in complexity
- Single forward pass for a 224 by 224 pixel input image
  - ResNet-50: ~3.86 GFLOPs
  - SE-ResNet-50 (reduction ratio $r = 16$): ~3.87 GFLOPs
  - Accuracy of SE-ResNet-50 is much greater
    - Approaches that of ResNet-101 (~7.58 GFLOPs)
- Slight runtime increase per block (tens of ms)
- Additional parameters from paired FC gating mechanism:
  - SE-ResNet-50 requires additional ~2.5m parameters on top of ~25m from ResNet-50 (~10% increase)

Related Work

- Increasing depth could improve learning quality (Inception and VGGNets models)
- Batch Normalization added stability by regulating distribution of inputs to each layer
- Identity-based skip connections allowed for deeper and stronger networks (ResNets)
- Gating mechanisms control the flow of info along shortcut connection (highway networks)
- Some research optimized for reducing computational and model complexity
    - Assumption: channel relationships can be formulated as composition of instance-agnostic functions with local receptive fields

Power of SENet

- Claim: Employ mechanism to model dynamic, non-linear dependencies between channels using global info
    - More efficient learning
    - Higher representational power
- SE blocks can be used as fundamental units for algorithmic architecture searches

Experiment

- ImageNet dataset
  - 1.28m training images and 50k test images
  - 1000 classes
- 100 epochs
- Initial LR = 0.6
  - Decreased to 0.06 at 30 epochs
  - Decreased to 0.006 at 60 epochs
  - Decreased to 0.0006 at 90 epochs
- Multiple blended architectures tested with positive results

Results

- First place in 2017 ILSVRC classification competition
  - Best result: 2.251% test top-5 error
  - 25% relative improvement over previous year's winner (2.991%
- Steady optimization on blended architectures
- SENets outperform all baseline architectures on CIFAR-10 and CIFAR-100

Conclusion

- SE blocks improve accuracy of network by enabling dynamic channel-wise feature recalibration
- SENets achieve state-of-the-art performance across multiple datasets and tasks
- Better understanding of failure of previous models to model channel-wise feature dependencies
    - Useful for tasks requiring strong discriminative features
- Feature importance values from SE blocks may help with other tasks (such as network pruning)

Questions

- What are downsampling operators?
- What are local and global theoretical receptive fields?
- What is a simple gating mechanism?
- What is single-crop error?