PennState College of Education | CSATS Center for Science and the Schools

Research Experiences for Teachers and Young Scholars in Advanced Self-Powered Systems of Integrated Sensors and Technologies (ASSIST)

SSIST

NSF

Developing a Multi-Channel Fine-Grained Image Classification Model

Justin Bush[1], Chonghan Lee[2], Zeinab Hakimi[2], Vijaykrishnan Narayanan[2]

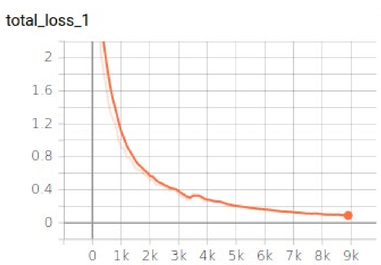
[1] Philipsburg-Osceola Area School District, Philipsburg, PA; [2] School of Electrical Engineering and Computer Science, The Pennsylvania State University

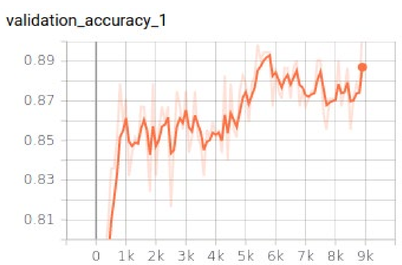
Objective

To develop and compare the relative effectiveness of a multi-channel fine-grained image classification model against a single-channel model on the CUB-200-2011 Dataset using predefined attention mechanisms.

Motivation

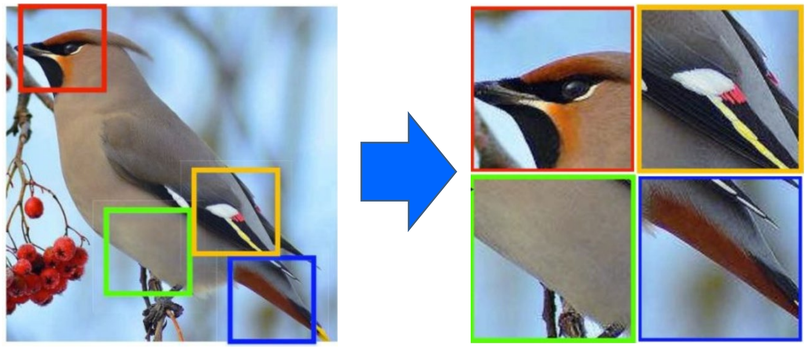
Current general classification models are often robust, but fine-grained classification models tend to be far less accurate. To improve the accuracy of the Inception v1 model's classification of bird species using the CUB-200-2011 Dataset, we proposed the use of a multi-channel approach with specific parts of the bird such as the beak, wing, tail, and breast as the attention mechanisms. This was inspired by the gains previously made on the accuracy of general classification of the ModelNet40 Dataset using a multi-channel multiple view model training process. The multiple view solution increased the performance on the ModelNet40 Dataset for general classification due to fixed point-of-view image training limiting the accuracy of the model for real world applications.



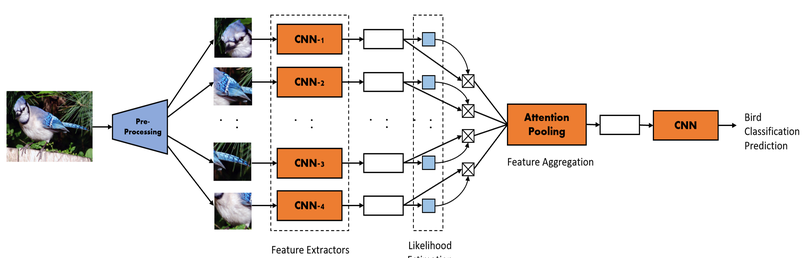


Making the Model

Our model was implemented using the TensorFlow machine learning library. We utilized the GoogLeNet (Inception v1) convolutional neural network architecture. Before training the model, we pre-processed each of the 11,788 images that represent the 200 different species of birds in the CUB-200-2011 Dataset to a uniform size. For the baseline single-channel model, training was performed using the bird images without specifying an attention mechanism. This required the model to use machine learning when inferencing the local attention mechanisms to highlight for classification. When training our multi-channel model, the bird's beak, wing, tail, and breast were identified as the specific attention mechanisms for the model to weigh.

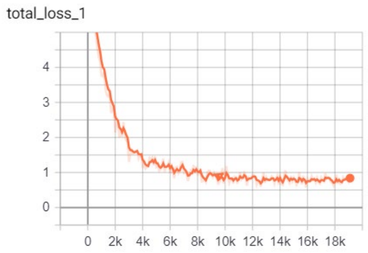


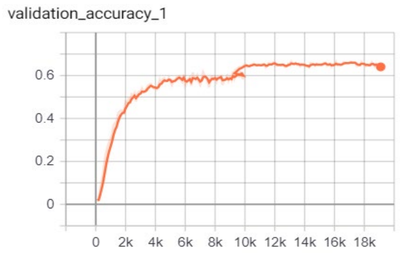
The model was trained over 19,000 iterations using a batch size of 32 images per iteration. With every iteration, the machine learning convolutional neural network convolved the learned features by passing each image through several convolutional layers and adjusting weights based on its learning.



Results

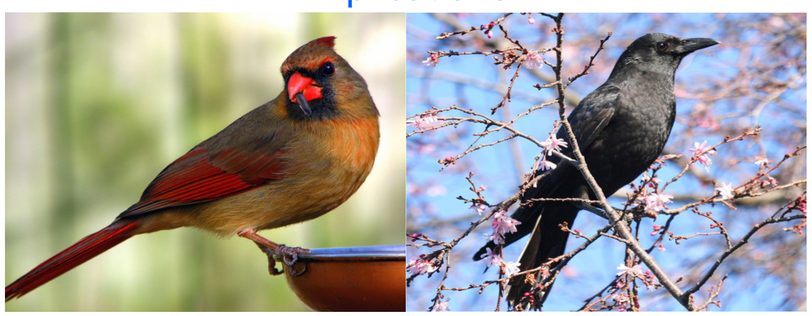
The model showed an 11% increase in accuracy of bird classification when compared against the single-channel model's performance on the same classification task.





Even though the CUB-200-2011 Dataset contained nearly 12,000 images, the results indicated that the presence of 200 different bird species classification options caused the model to begin overfitting somewhere between the 4,000th and 6,000th iterations. The total loss decreased rapidly prior to the overfitting, after which it remained relatively constant. Similarly, the accuracy of classification rapidly increased prior to the 4,000th iteration before leveling off.

Implications



The multi-channel approach to image classification shows promise as both the multiple view model and the fine-grained focused attention model showed improved accuracy over the single channel model on their respective datasets. With a sufficiently large training set and additional resources, the model could be improved to match the accuracy of state-of-the-art proprietary models.

The ASSIST RET and Young Scholar programs are funded by NSF Nanosystems Engineering Research Center Grant EEC 1160483.