In this ablation study, we evaluate 5 popular deep learning architectures for a binary image classification task using F1-score as our primary metric. The models tested include ResNet50, Inception, InceptionResNetV2, MobileNetV2 and YOLOv8n. Using dataset from "Kubric: A Scalable Dataset Generator" presented at CVPR 2022.

Our dataset contains images of many objects but for binary classification we will try shoes, toys, bags and mice each separately. The frequency of objects varies, with shoes appearing most frequently and mice least in between toys and bags respectively. We trained and evaluated each model 4 times on our data using Google Collab and Python 3.10.12 with PyTorch 2.1.0 on a Tesla V100 GPU.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ResNet50 | Inception | InceptionResNetV2 | MobileNetV2 | | Yolov8n |
| Total params (Millions) | 24.6 M (93.98 MB) | 22.852 M (87.17 MB) | | | | 2.7 M (6 MB) |
| Trainable params (Millions) | 1.049 M (4.00 MB) | 22.818 M (87.04 MB) | | | | 1.44 M |
| Non-trainable params (Millions) | 23.58 M (89.98 MB) | 34432 M (134.50 KB) | | | | 1.26 M |
| Training Time (Seconds) | 40.83 s | 66.65 s | 78.8 s | | 79.6 s | 26.7 s |
| Evaluation Time [Test Data] (Seconds) | 0.37 s | 0.385 s | 0.55 s | | 0.55 s | 0.8 s |
| Predict Time [Test Data] (Seconds) | 0.368 s | 0.368 s | 0.54 s | | 0.54 s |

Model performance was assessed also based on training time, evaluation time, and parameter counts (trainable and non-trainable) as shown in Table [1]. The goal of this ablation study is to identify the most suitable architecture for our binary classification problem by rigorously profiling each model according to our evaluation procedure and metrics. Insights gained will aid in selecting the optimal model configuration to deploy for this application.The shoe (object 14) was the most frequently occurring object in the dataset, appearing in 2208 out of 2160 total frames. For classification purposes, 1160 images showed the shoe and were given the positive label "shoe". The other 1000 frames did not feature the shoe and were thus assigned the negative label "no shoe".

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | ResNet50 | | Inception | | InceptionResNetV2 | | MobileNetV2 | | Yolov8n | |
| Testing | F1-Score | | 70.92% | | 99.39% | | 99.7% | | 99.7% | | 100% | |
| Accuracy | | 71.91% | | 99.38% | | 99.69% | | 99.69% | | 100% | |
| Loss | | 0.567 | | 0.01528 | | 0.0268 | | 0.025 | | 0.0023 | |
| TN | FP | 122 | 35 | 157 | 0 | 157 | 0 | 156 | 1 | 157 | 0 |
| FN | TP | 56 | 111 | 2 | 165 | 1 | 166 | 0 | 167 | 0 | 167 |
|  | Epoch Till Early Stop  (Out of 32) | | 32 | | 20 | | 32 | | 32 | | 6 | |

This labeling scheme created a semantically balanced binary classification task for the shoe object, with approximately half the data belonging to each class. Evaluating model performance on this common vs. rare object distribution provides insight into how the architectures handle class imbalance. The results also reflect how well models can distinguish between images containing the shoe versus all other backgrounds and objects.

The toy (object 16) was the second most frequently occurring object in the dataset, appearing in 888 out of 2160 total frames. Where 760 images featured the toy and were assigned the positive class "toy". The remaining 1400 frames did not show the toy and were given the negative label "no toy".

This labeling process created an imbalanced binary classification problem for the toy object, with approximately 35% of data belonging to the "toy" class and 65% to "no toy". Evaluating model performance in this minority class scenario provides insight into how the architectures handle class imbalance when distinguishing toy images from others. The results reflect the ability of models to accurately classify images as containing the toy or not under an imbalanced class distribution.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | ResNet50 | | Inception | | InceptionResNetV2 | | MobileNetV2 | | Yolov8n | |
| Testing | F1-Score | | 23.07% | | 99.53% | | 97.2% | | 98.14% | | 100% | |
| Accuracy | | 69.13% | | 99.69% | | 98.14% | | 97.24% | | 100% | |
| Loss | | 0.57 | | 0.005 | | 0.042 | | 0.064 | | 0.01158 | |
| TN | FP | 209 | 8 | 217 | 1 | 214 | 3 | 213 | 4 | 217 | 0 |
| FN | TP | 92 | 18 | 1 | 106 | 3 | 104 | 3 | 104 | 0 | 107 |
|  | Epoch Till Early Stop  (Out of 32) | | 32 | | 29 | | 32 | | 32 | | 6 | |

The bag (object 1) appeared least frequently of the objects studied, present in 216 out of the total 2160 frames. For classification. Where 250 images showed the bag and received the positive class label "bag". The remaining 1910 frames did not feature the bag and were assigned the negative class "no bag".

This labeling process created a highly imbalanced binary problem for the bag object, with approximately 12% of data belonging to the "bag" class and 88% to "no bag". Evaluating model performance in this extreme minority class scenario provides insight into how well architectures can classify bag images compared to others under significant class imbalance. The results indicate each model's ability to accurately discriminate between frames containing the bag versus all other content.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | ResNet50 | | Inception | | InceptionResNetV2 | | MobileNetV2 | | Yolov8n | |
| Testing | F1-Score | | 11.42% | | 100% | | 95.23% | | 93.5% | | 100% | |
| Accuracy | | 90.43% | | 100% | | 99.07% | | 98.7% | | 100% | |
| Loss | | 0.234 | | 0.003 | | 0.0197 | | 0.022 | | 0.00028 | |
| TN | FP | 219 | 0 | 291 | 0 | 291 | 0 | 291 | 0 | 291 | 0 |
| FN | TP | 25 | 3 | 0 | 33 | 3 | 30 | 4 | 29 | 0 | 33 |
|  | Epoch Till Early Stop  (Out of 32) | | 31 | | 29 | | 32 | | 32 | | 2 | |

The mouse (object 12) was the least frequently occurring object, appearing in only 48 out of the total 2160 frames.

For classification, 48 images depicted the mouse and were labeled with the positive class "mouse". The remaining 2112 frames that did not show the mouse were given the negative class label "no mouse".

This labeling scheme produced an extremely imbalanced binary problem for the mouse, with approximately 2% of data belonging to the minority "mouse" class and 98% to "no mouse". Evaluating model performance in this rare class scenario provides insight into how well the architectures handle highly skewed distributions. The results indicate each model's capability to accurately classify mouse images compared to other content when dealing with such extreme class imbalance.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | ResNet50 | | Inception | | InceptionResNetV2 | | MobileNetV2 | | Yolov8n | |
| Testing | F1-Score | | 0% | | 100% | | 100% | | 100% | | 100% | |
| Accuracy | | 97.53% | | 100% | | 100% | | 100% | | 100% | |
| Loss | | 0.1 | | 0.00008 | | 0.00016 | | 0.002 | | 0.00010 | |
| TN | FP | 316 | 0 | 316 | 0 | 316 | 0 | 316 | 0 | 316 | 0 |
| FN | TP | 8 | 0 | 0 | 8 | 0 | 8 | 0 | 8 | 0 | 8 |
|  | Epoch Till Early Stop  (Out of 32) | | 21 | | 32 | | 32 | | 32 | | 2 | |