VR Mini Project

Image Captioning

Group Code - AAAS

Aarushi Goenka - IMT2018001
Ameya Kurme - IMT2018007
Ayush Mishra - IMT2018013
Soham Kolhe - IMT2018073

Theory -

Image Captioning - Image Captioning is the process of generating textual description of an image – based on the objects and actions in the image.

MobileNet - MobileNet is a CNN architecture model for Image Classification and Mobile Vision.There are other models as well but what makes MobileNet special is that it needs very less computation power to run or apply transfer learning to. This makes it a perfect fit for Mobile devices,embedded systems and computers without GPU or low computational efficiency compromising significantly with the accuracy of the results.

Architecture -

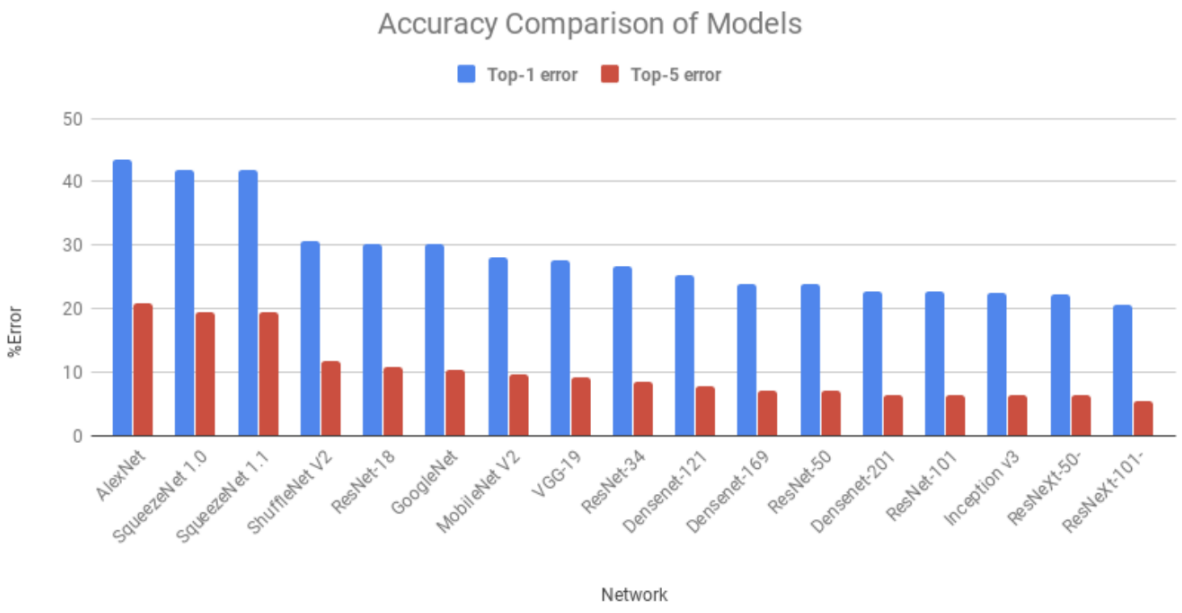
Preprocessing -

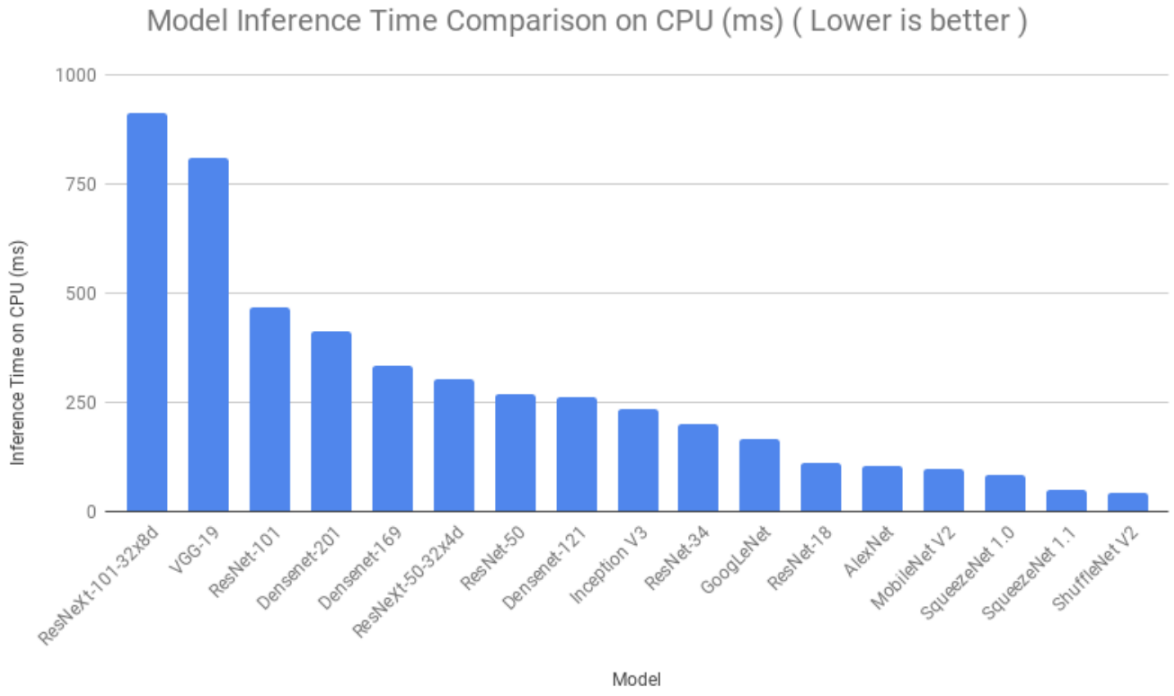
First, in preprocessing, basic Data Augmentation is done. The images are resized to 256. Then a random crop of 224 and then normalized further.
The Model used is Sequence to sequence model. That is, a method of encoder-decoder based machine translation and language processing. In the encoder we used both resNet50 and mobileNet models and then based on results we chose the more suitable one for the final captioning and bleu score.

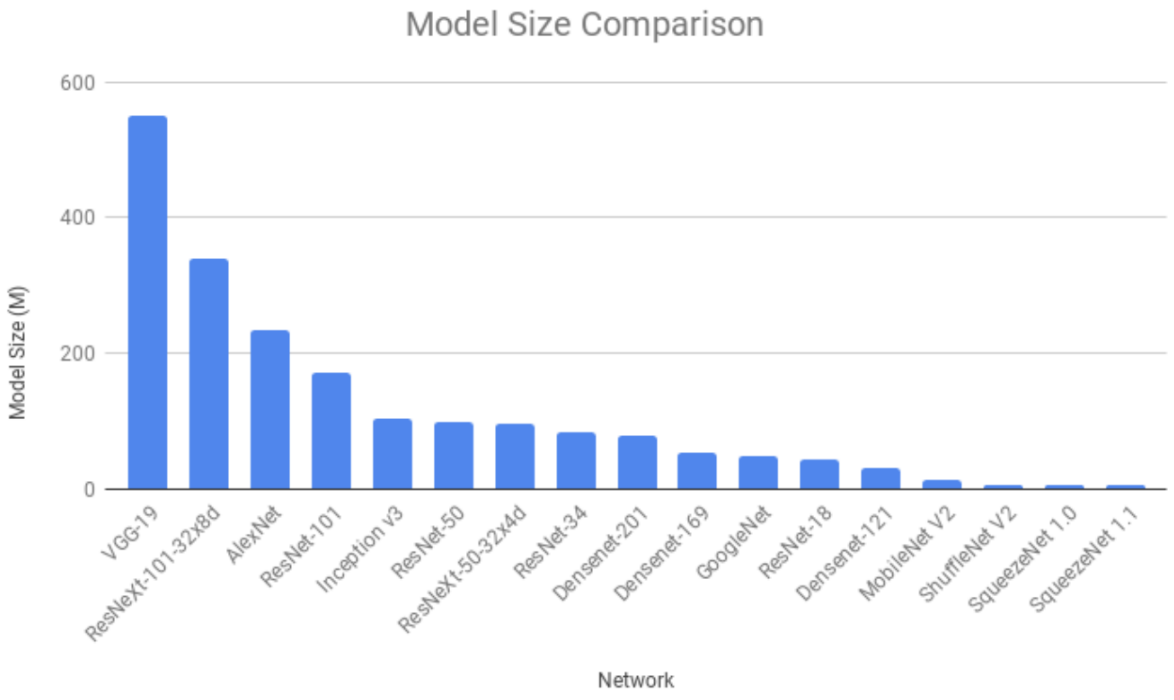
Comparing Model for Image Classification -

1. Top-1 Error: A top-1 error occurs if the class predicted by a model with highest confidence is not the same as the true class.

2. Top-5 Error: A top-5 error occurs when the true class is not among the top 5 classes predicted by a model (sorted in terms of confidence).







Our Decoder is the attention decoder. Attention in RNNs corresponds to context. Each generated output word is not just a function of just the

final hidden state but rather a function of ALL hidden states. It is not a simple concatenation or dot product, but an “attention” operation that, for every decoder output step, produces a distinct vector representing all encoder hidden states but giving different weights to different encoder hidden states. It is helpful with long length sequence input.

We chose using MobileNet because the model’s size was below 100 MB although resNet50 was performing better.

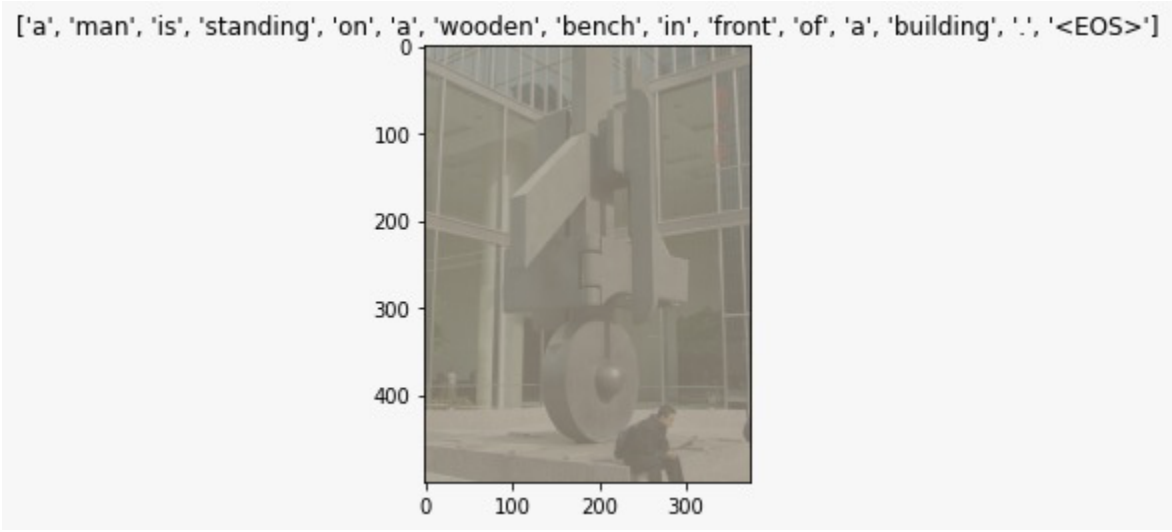
Metric -

The metric used for the evaluation of the accuracy of the Image Captioning obtained by our model is called Bleu Score. BLEU stands for Bi-Lingual Evaluation Understudy. It’s a popular and inexpensive way to automatically measure the performance of a Machine Translation model. In a nutshell, BLEU compares the machine’s translation — what’s known as the candidate translation — with existing human-generated translations, known as the reference translations.

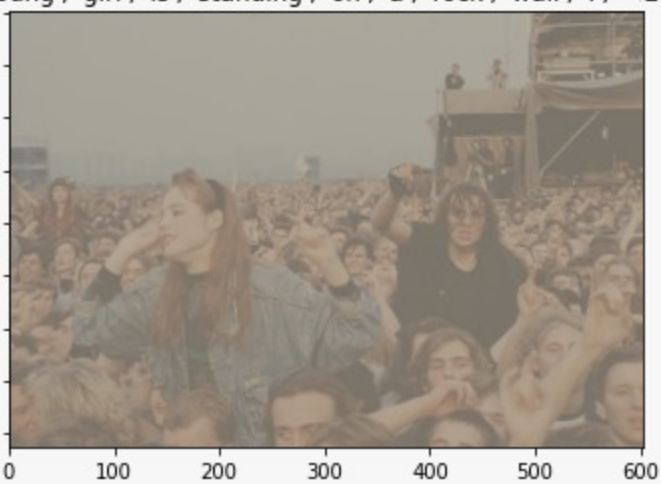
Observations -

Model	Epoch	Bleu Score	Model Size(in MB)
ResNet50	1	0.04	112
ResNet50	20	0.33	112
MobileNet	1	0.03	96
MobileNet	20	0.30	96

Shared Images with subjective results -



['a', 'young', 'girl', 'is', 'standing', 'on', 'a', 'rock', 'wall', '.', '<EOS>']



['a', 'man', 'in', 'a', 'black', 'shirt', 'is', 'standing', 'near', 'a', 'rock', 'wall', '.', '<EOS>']

['a', 'person', 'in', 'a', 'yellow', 'shirt', 'is', 'running', 'through', 'the', 'sand', '.', '<EOS>']