Capstone Project: Image Caption

Cheryl Leong

Outline

Problem Statement and Background

Exploratory Data Analysis

Preprocessing / Feature Engineering

Hyperparameter Tuning and Modelling

Results

Summary and Improvements

Problem Statement and Background

Problem Statement

"A picture is worth a thousand words".

This adage was first coined in the 1900s. The idea behind it was that complex and sometimes *multiple ideas* can be conveyed by a single still image. This suggests that perception is subjective and there are numerous interpretations to a single image.

In this capstone project, I will be exploring the use of Neural Networks to generate captions in English to best describe an image. The generated captions can then be converted to audio output. This will be beneficial to provide context to the visually-impared.

Background

Dataset consists of 8,091 images and 40,455 captions.

Each image has 5 captions tagged to it and identified by the image path

A little girl in a pink dress going into a wooden cabin .
A little girl climbing the stairs to her playhouse .
A little girl climbing into a wooden playhouse .
A girl going into a wooden building .
A child in a pink dress is climbing up a set of stairs in an entry way .

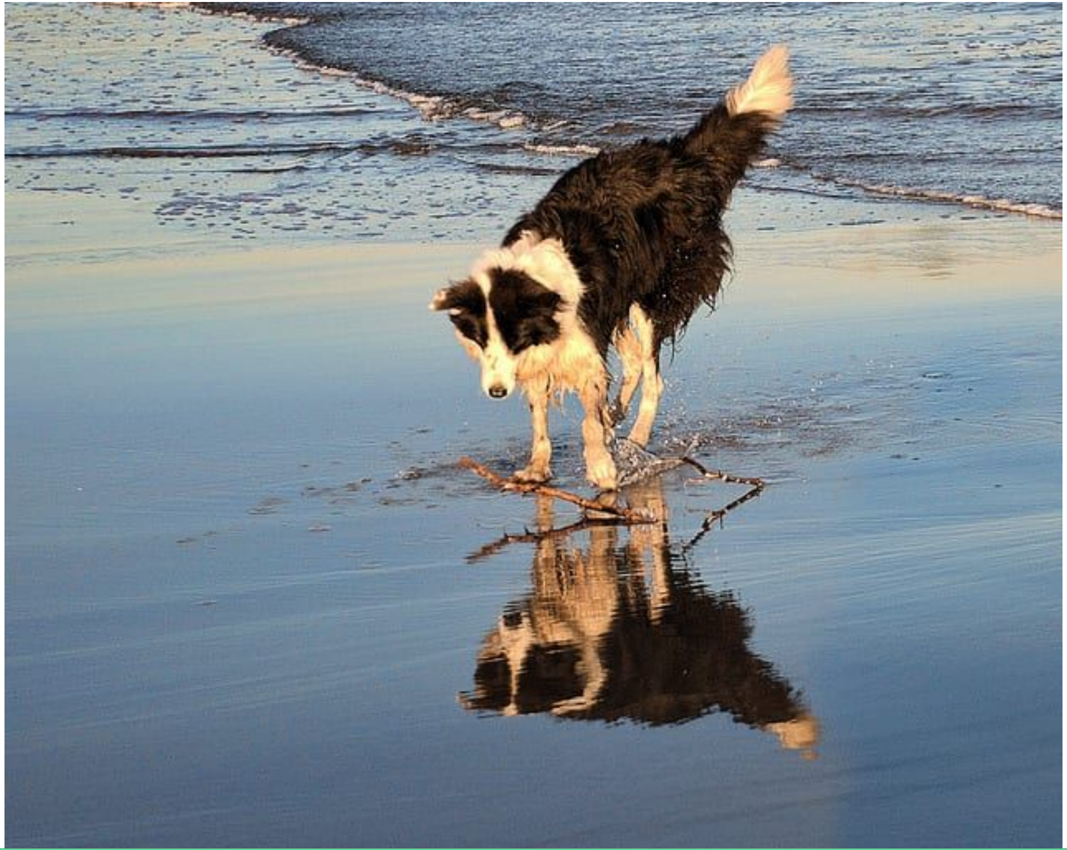
Two dogs on pavement moving toward each other .
Two dogs of different breeds looking at each other on the road .
A black dog and a white dog with brown spots are staring at each other in the street .
A black dog and a tri-colored dog playing with each other on the road .
A black dog and a spotted dog are fighting

Young girl with pigtails painting outside in the grass .
There is a girl with pigtails sitting in front of a rainbow painting .
A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
A little girl is sitting in front of a large painted rainbow .
A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .

man laying on bench holding leash of dog sitting on ground
A shirtless man lies on a park bench with his dog .
a man sleeping on a bench outside with a white and black dog sitting next to him .
A man lays on the bench to which a white dog is also tied .
A man lays on a bench while his dog sits by him .

The man with pierced ears is wearing glasses and an orange hat .
A man with glasses is wearing a beer can crocheted hat .
A man with gauges and glasses is wearing a Blitz hat .
A man wears an orange hat and glasses .
A man in an orange hat starring at something .

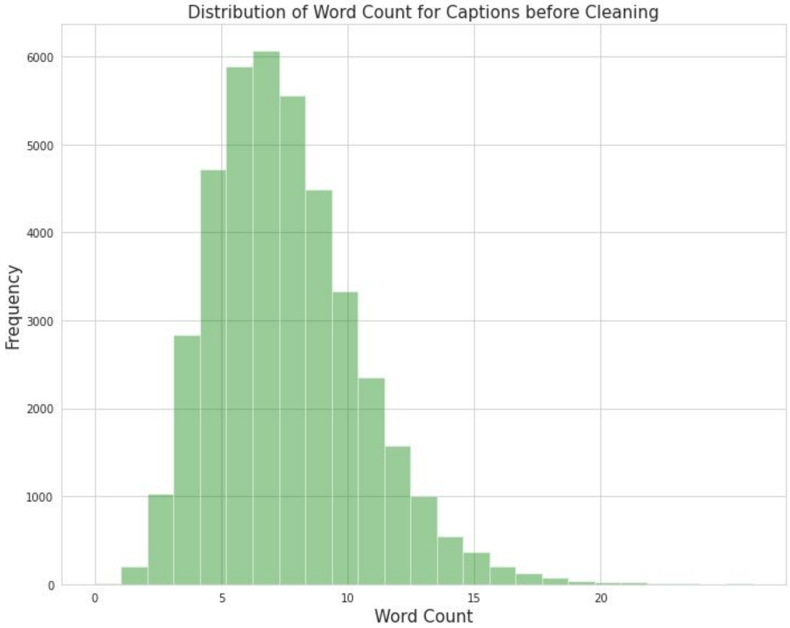
Caption This!

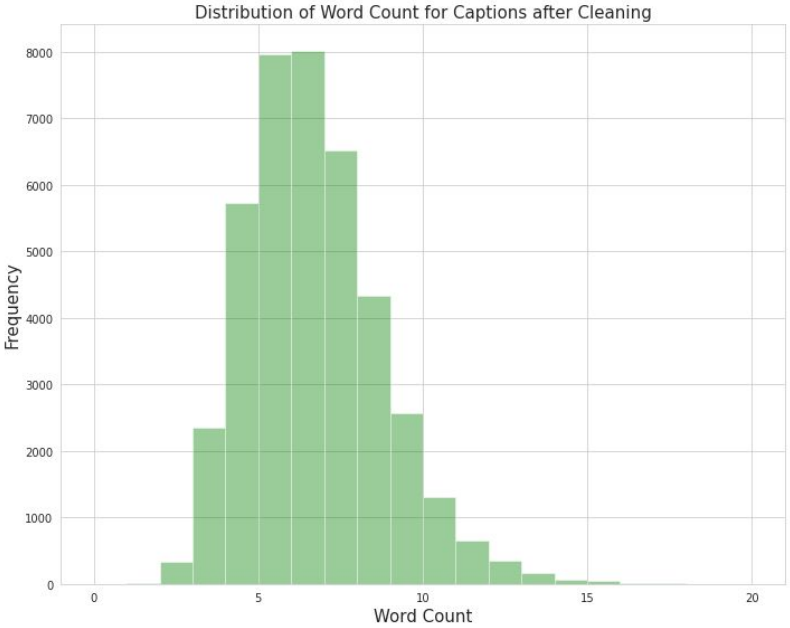


Exploratory Data Analysis

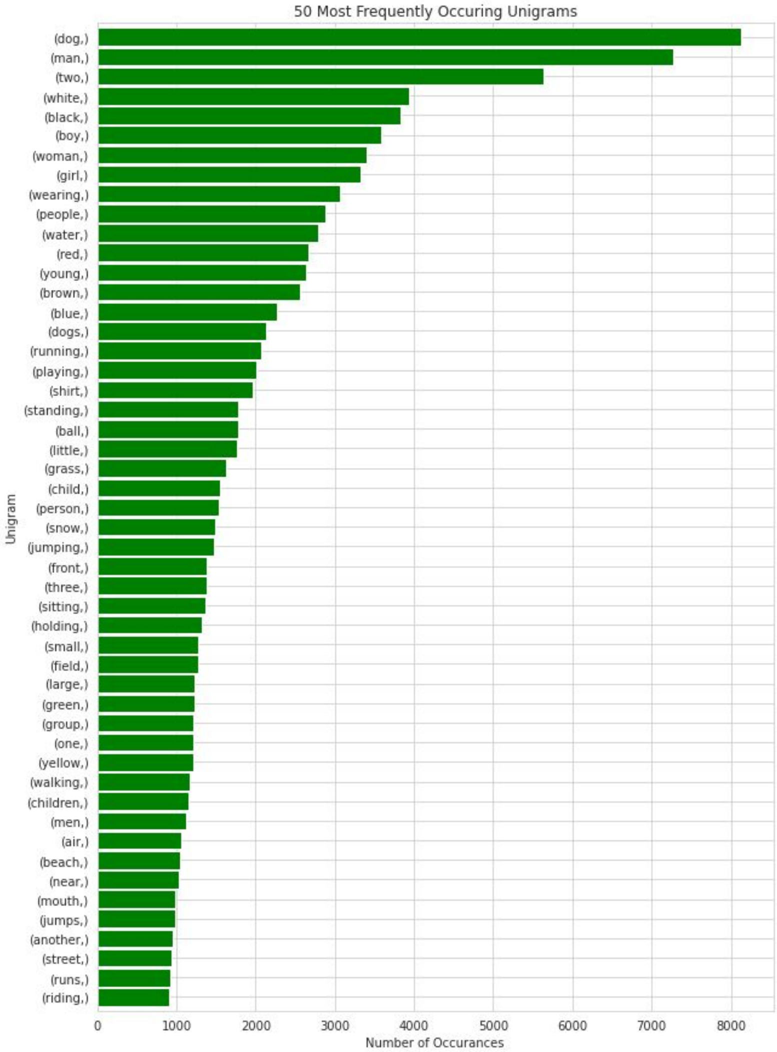
Distribution of Word Count for Captions

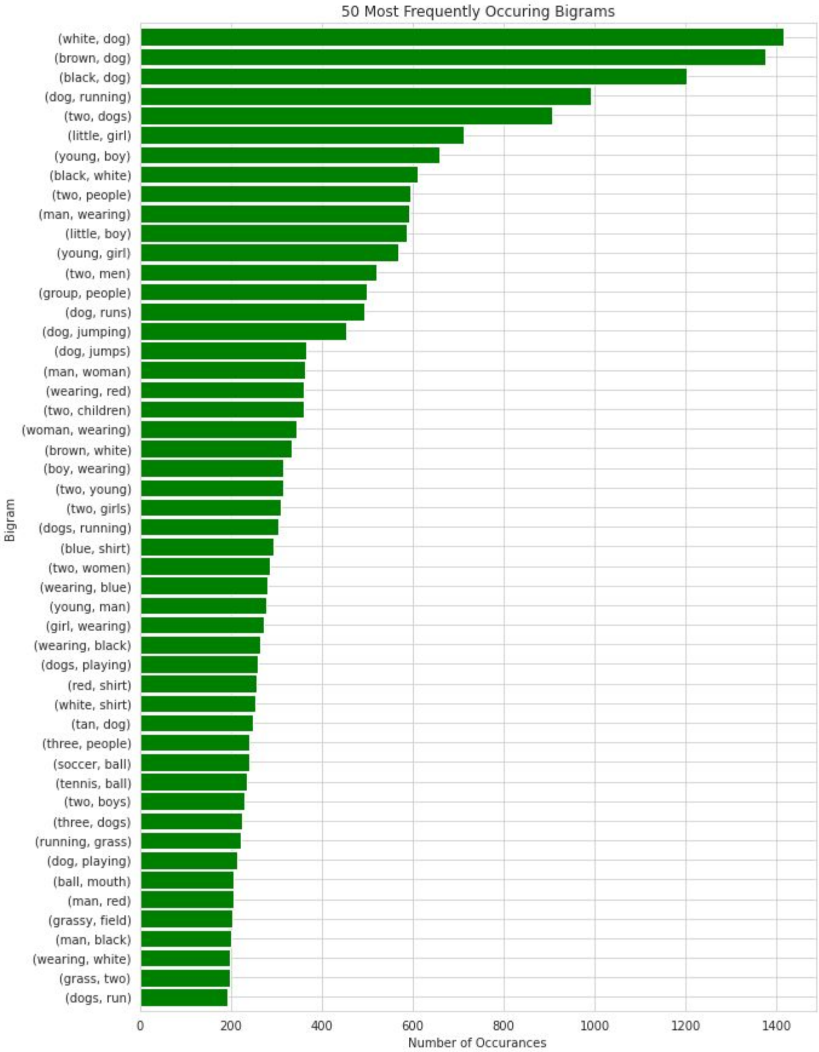
After cleaning, word count per caption decreased but the frequency of the median word count has increased





50 Most Frequent Occurring N-grams





Preprocessing / Feature Engineering

Pre-trained Models

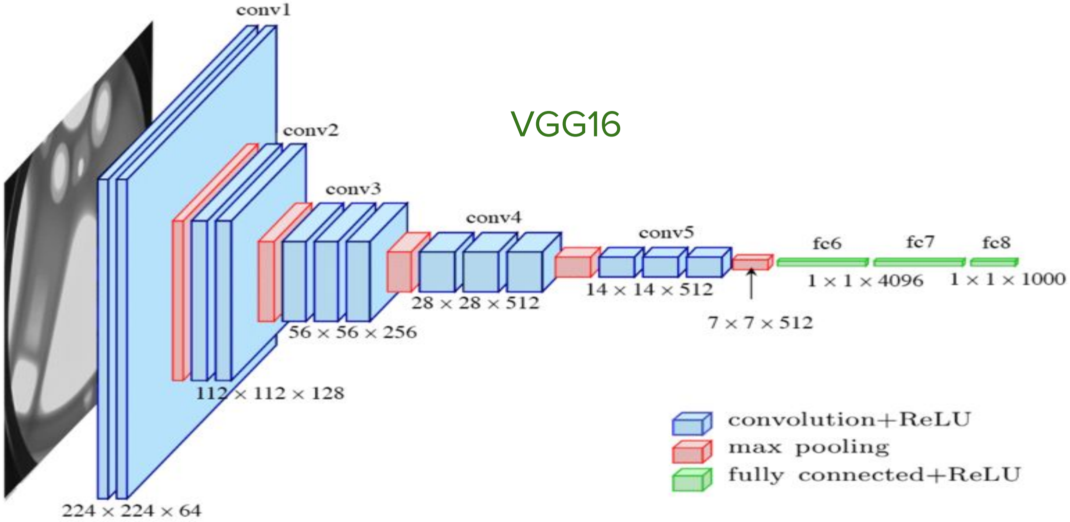
VGG16: 16 layers

Inception v3: 48 layers

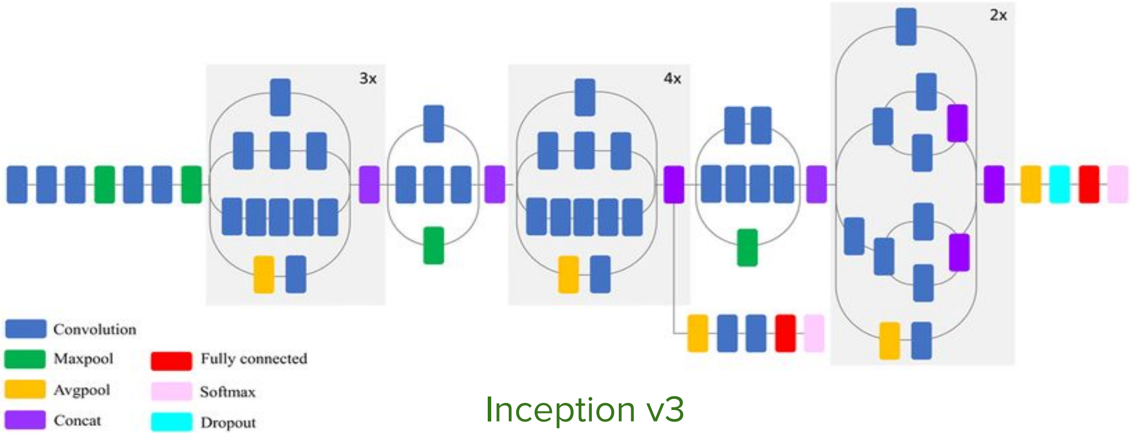
Both models trained with ImageNet dataset

Transfer Learning: Features from images dataset are extracted

Prediction Layers for both models were removed



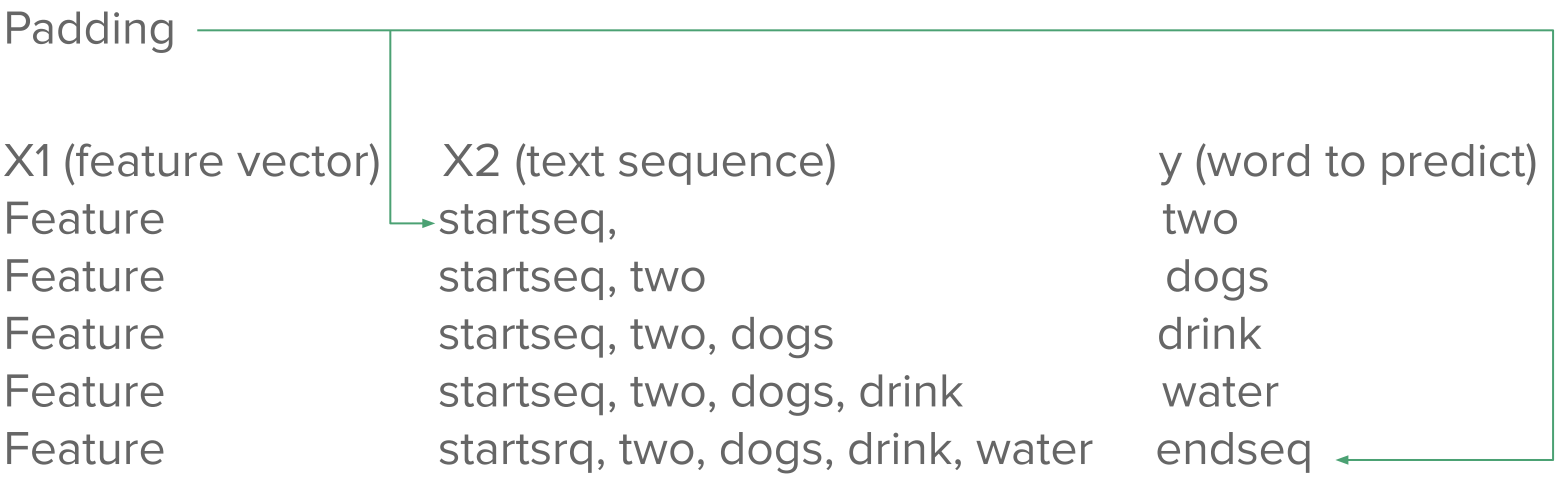
VGG16



Inception v3

Feature Engineering

Embedding by mapping of descriptions to image



Padding

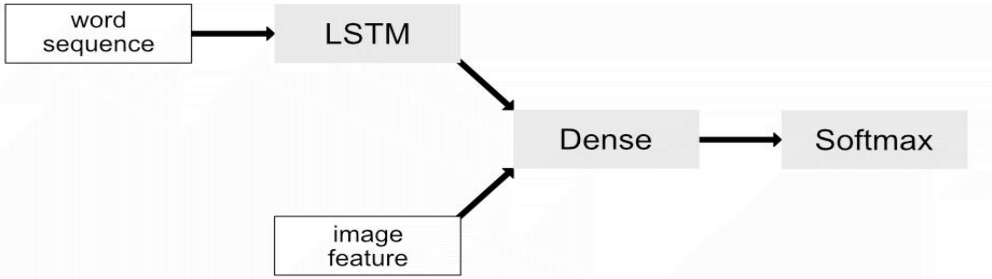
X1 (feature vector)	X2 (text sequence)	y (word to predict)
Feature	startseq,	two
Feature	startseq, two	dogs
Feature	startseq, two, dogs	drink
Feature	startseq, two, dogs, drink	water
Feature	startsrq, two, dogs, drink, water	endseq

Hyperparameter Tuning and Modelling

Hyperparameter Tuning

3 models each for VGG16 and Inception v3

1. Base Model (Dropout = 0.4)
2. Base Model + Dropout (0.6)
3. Base Model + Dropout (0.6) + Kernel Regularizer (Ridge Regression, L2 = 0.01)



word sequence

LSTM

Dense

Softmax

image feature

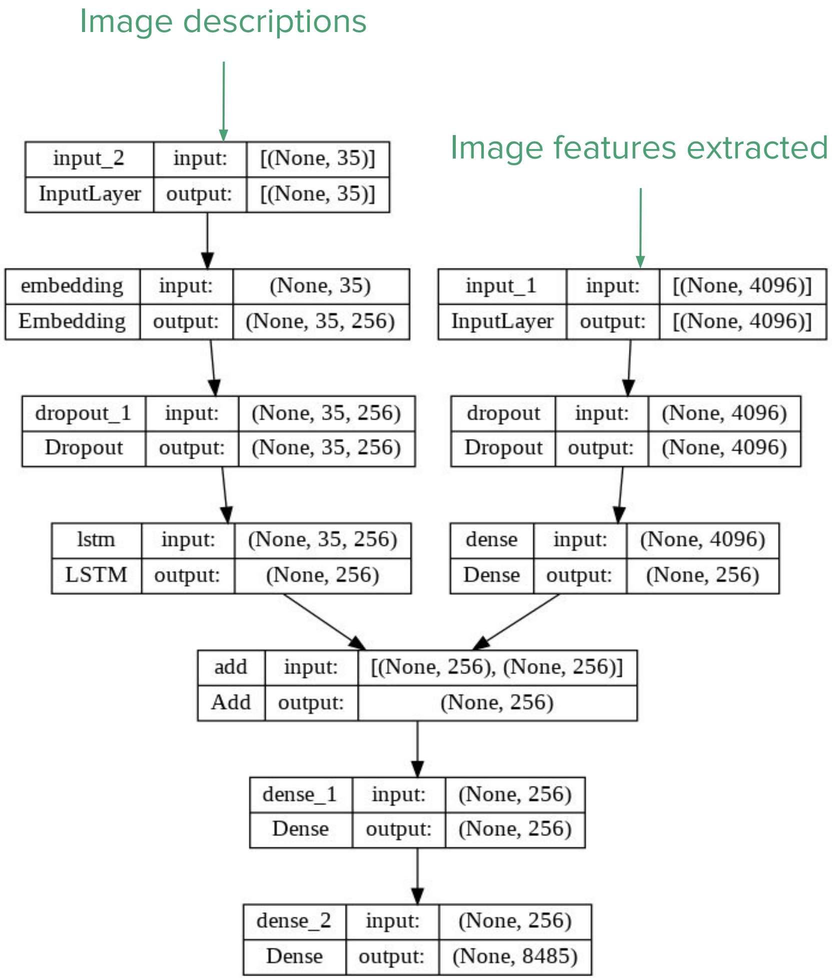


Image descriptions

Image features extracted

input_2 InputLayer | input: [(None, 35)] | output: [(None, 35)]

embedding Embedding | input: (None, 35) | output: (None, 35, 256)

input_1 InputLayer | input: [(None, 4096)] | output: [(None, 4096)]

dropout_1 Dropout | input: (None, 35, 256) | output: (None, 35, 256)

dropout Dropout | input: (None, 4096) | output: (None, 4096)

lstm LSTM | input: (None, 35, 256) | output: (None, 256)

dense Dense | input: (None, 4096) | output: (None, 256)

add Add | input: [(None, 256), (None, 256)] | output: (None, 256)

dense_1 Dense | input: (None, 256) | output: (None, 256)

dense_2 Dense | input: (None, 256) | output: (None, 8485)

Modelling

Train/Test Split = 84% train, 8% val, 8% test

Total Epochs = 300

Early stopping (patience) = 3 epochs, val accuracy

Optimizer = Adam

Batch Size = 32

Learning Rate = 0.001

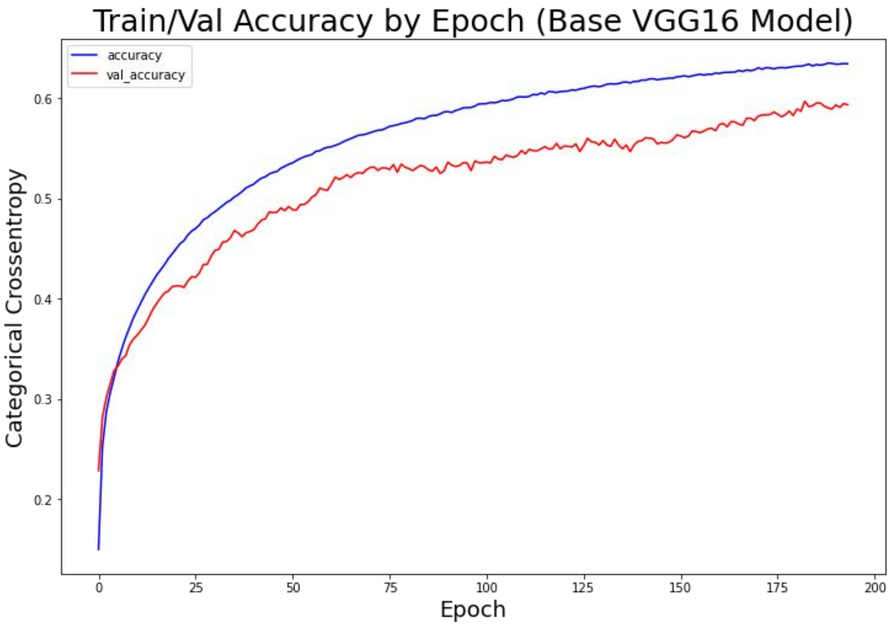
Loss = Categorical Crossentropy

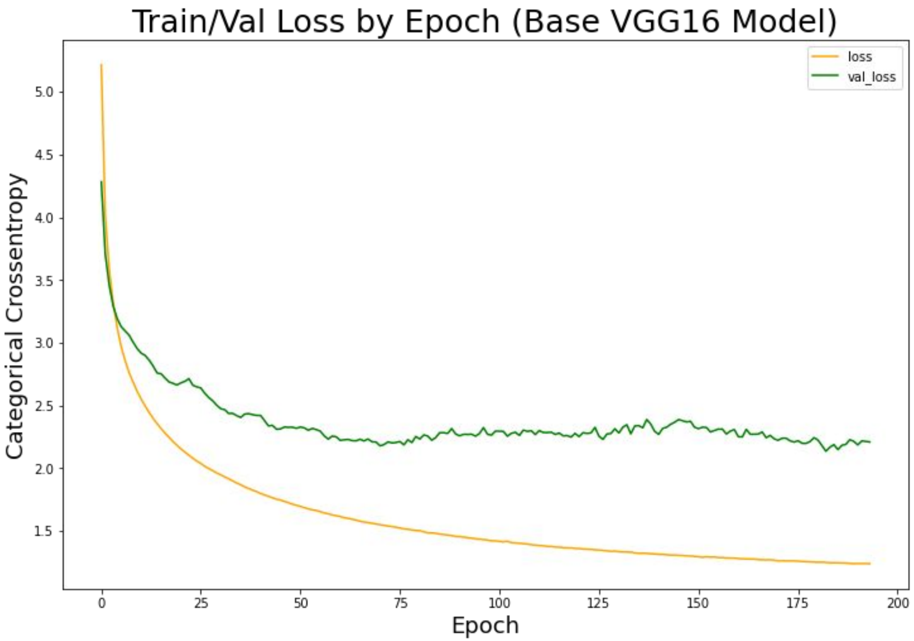
Metrics = Accuracy

Results

Performance Evaluation

Results for VGG16 Base Model (Dropout = 0.4)



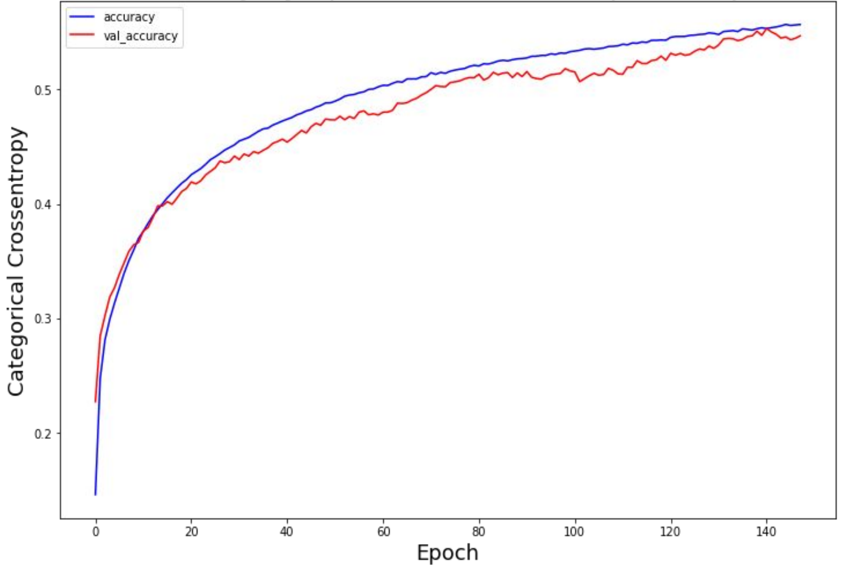


- **Accuracy**
- **Val Accuracy**

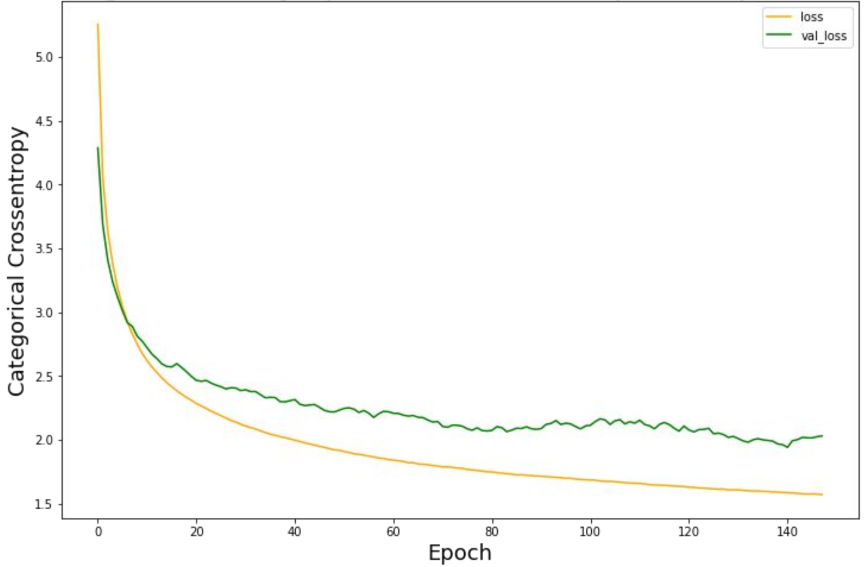
- **Loss**
- **Val Loss**

Results for VGG16 Base Model + Dropout (0.6)

Train/Val Accuracy by Epoch (Base VGG16 plus Dropout Model)



Training/Val Loss by Epoch (Base VGG16 plus Dropout Model)

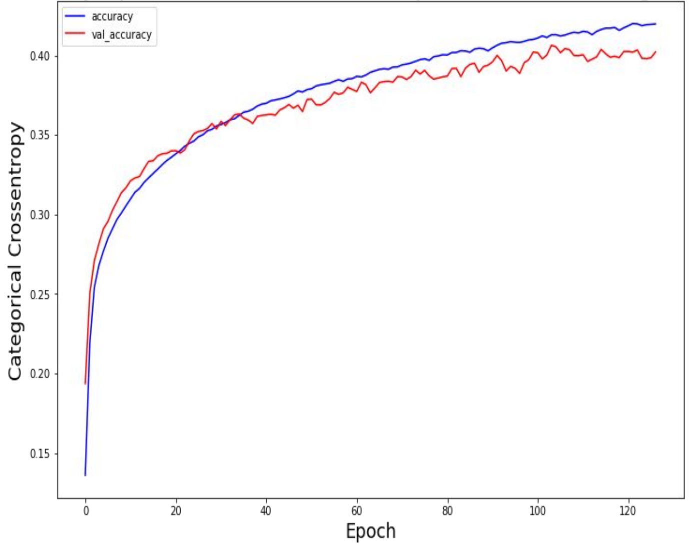


- Accuracy
- Val Accuracy

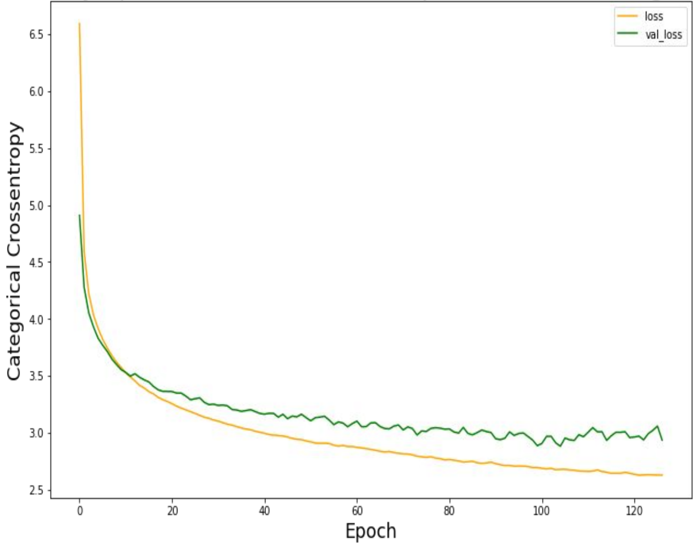
- Loss
- Val Loss

Results for VGG16 Base Model + Dropout (0.6) + Kernel Regularizer (L2 = 0.01)

Train/Val Acc by Epoch (Base VGG16 + Dropout and Kernel Regularizer Model)



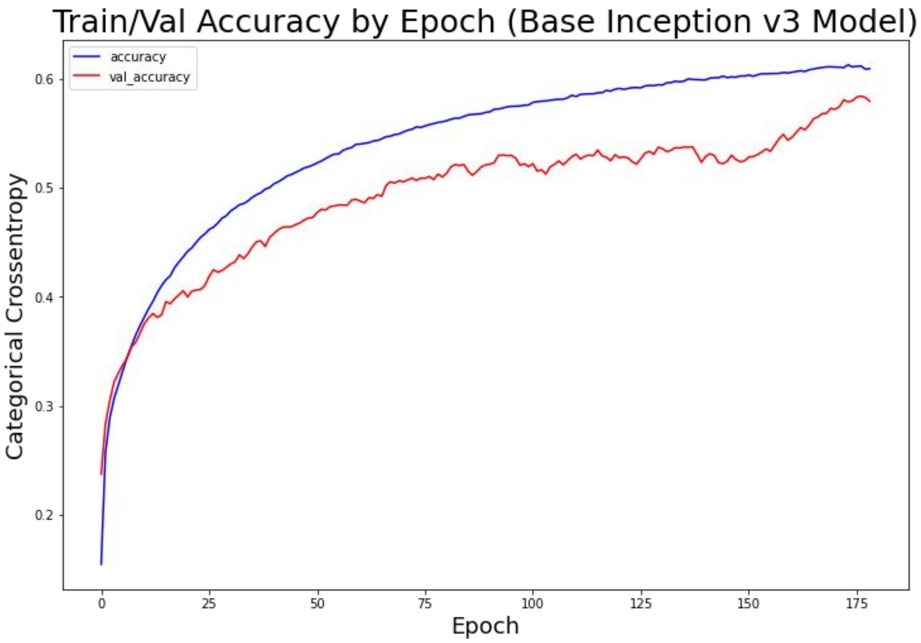
Train/Val Loss by Epoch (Base VGG16 + Dropout and Kernel Regularizer Model)

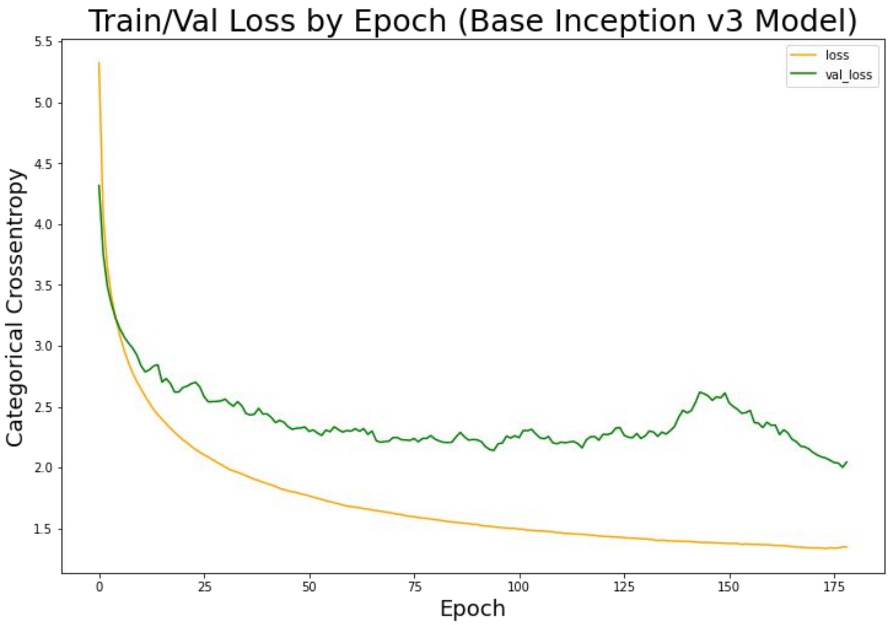


- Accuracy
- Val Accuracy

- Loss
- Val Loss

Results for Inception v3 Base Model (Dropout = 0.4)

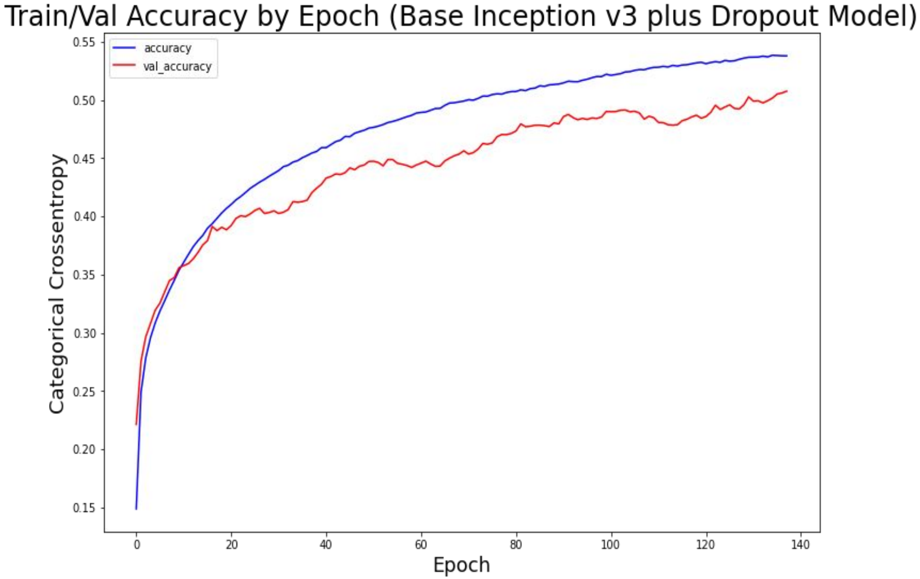


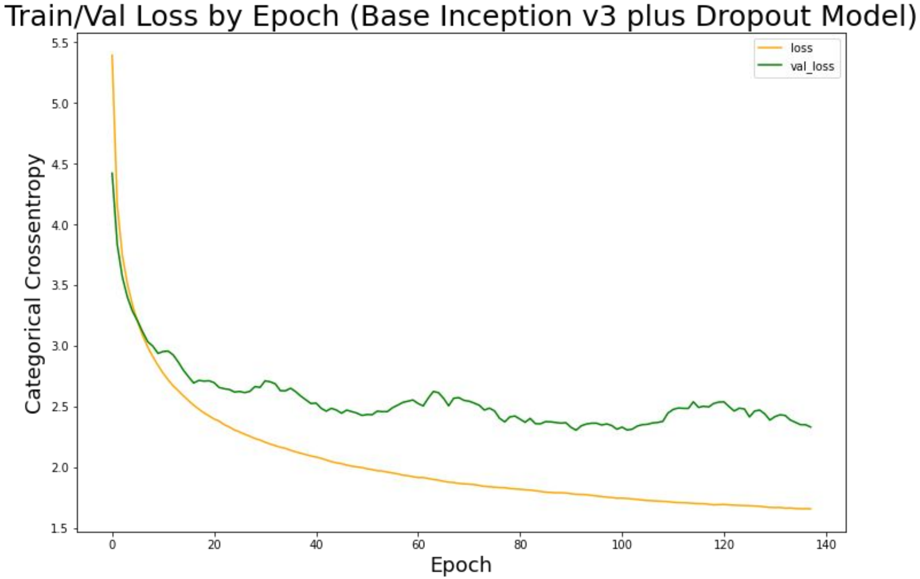


- **Accuracy**
- **Val Accuracy**

- **Loss**
- **Val Loss**

Results for Inception v3 Base Model + Dropout (0.6)

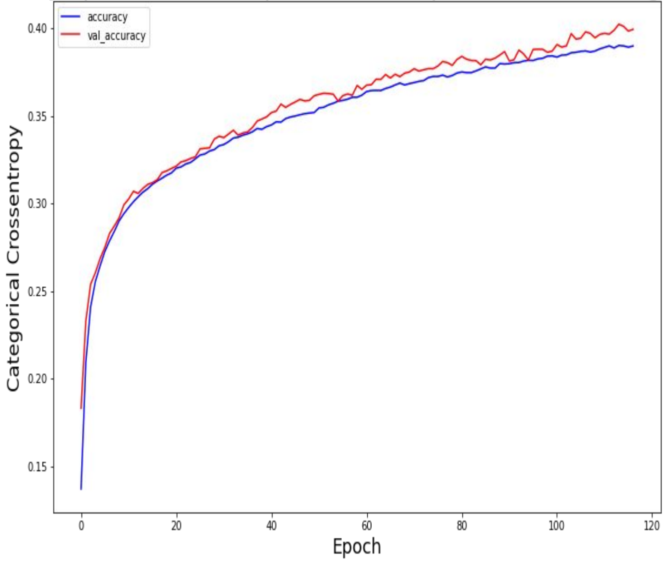

Train/Val Accuracy by Epoch (Base Inception v3 plus Dropout Model)

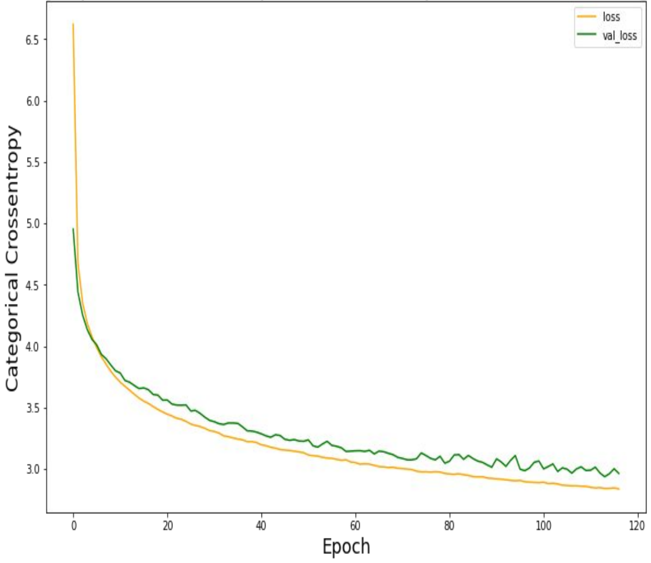

Train/Val Loss by Epoch (Base Inception v3 plus Dropout Model)

- **Accuracy**
- **Val Accuracy**

- **Loss**
- **Val Loss**

Results for Inception v3 Base Model + Dropout (0.6) + Kernel Regularizer (L2 = 0.01)


Train/Val Acc by Epoch (Base Inception v3 + Dropout and Kernel Regularizer Model)


Train/Val Loss by Epoch (Base Inception v3 + Dropout and Kernel Regularizer Model)

- **Accuracy**
- **Val Accuracy**

- **Loss**
- **Val Loss**

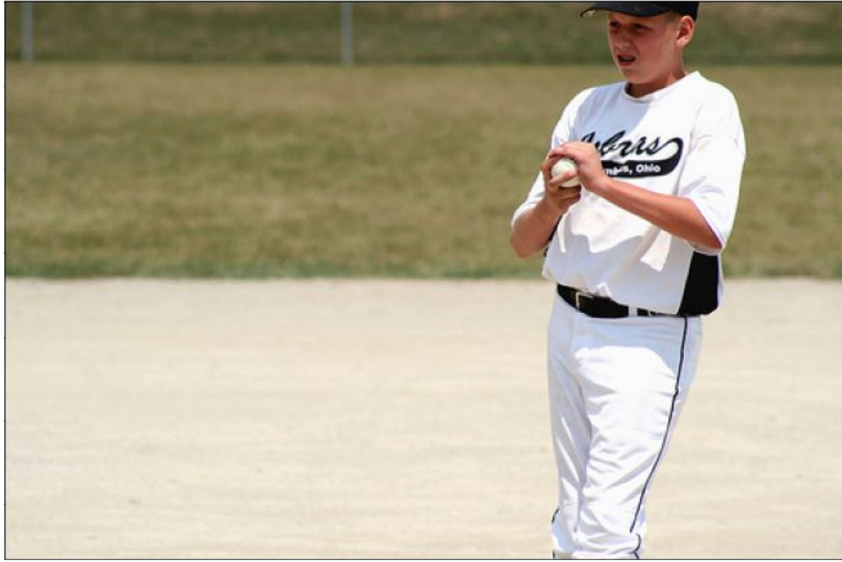
Bilingual Evaluation Understudy (BLEU) SCORE

	VGG16 Base	VGG16 Base + Dropout	VGG16 Base + Dropout + Kernel Regularizer	Inception v3 Base	Inception v3 Base + Dropout	Inception v3 Base + Dropout + Kernel Regularizer
BLEU-1	0.65	0.62	0.57	0.59	0.58	0.52
BLEU-2	0.47	0.43	0.33	0.39	0.38	0.28
BLEU-3	0.36	0.32	0.21	0.27	0.26	0.17
BLEU-4	0.28	0.24	0.12	0.19	0.18	0.1

Scores are calculated using modified n-gram precision score that clips the number of times to count a word, based on the maximum number of time it appears in the reference translation

Visualisation Evaluation

Actual vs Predicted Captions



Actual:

boy in white plays baseball

young boy is getting ready to through baseball

little leaguer getting ready for pitch

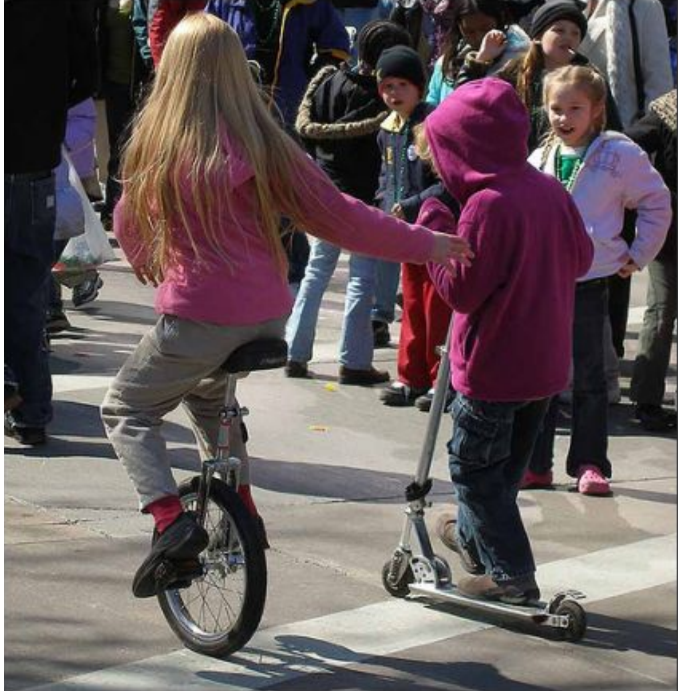
the boy is wearing white baseball uniform and holding baseball

the young ohio baseball player contemplates his pitch

Predicted:

VGG16 Base:	boy in blue and white holds ball
VGG16 Base + Dropout:	little leaguer getting ready to pitch
VGG16 Base + Dropout + Kernel Regularizer:	man in black shirt and white shirt is running on the grass

Actual vs Predicted Captions



Actual:

girl rides unicycle as another rides scooter next to her

girl rides unicycle with child who rides scooter

young girl rides unicycle next to another riding scooter on busy street

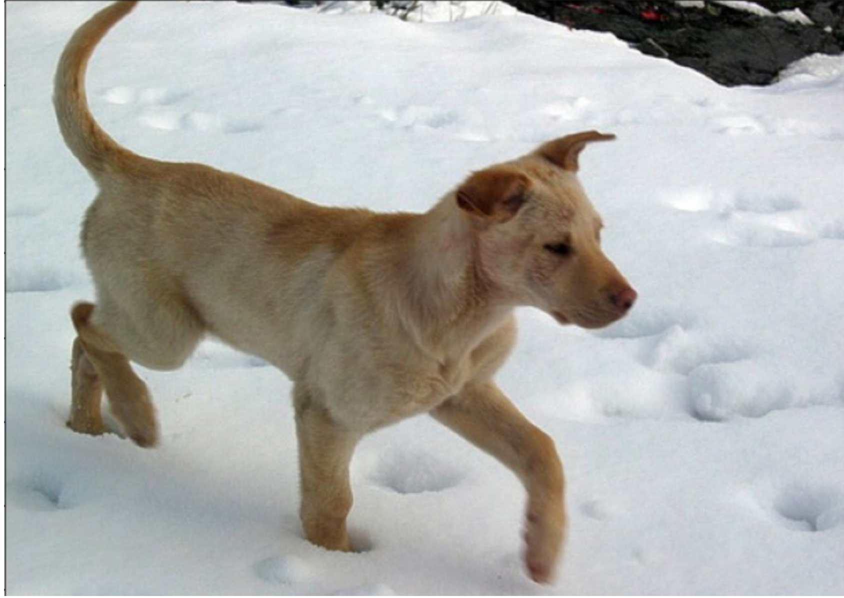
the girl on the unicycle reaches out for the child on the scooter

there is girl on unicycle and child on scooter

Predicted:

VGG16 Base:	girl rides scooter while another stands next to her
VGG16 Base + Dropout:	man in wheelchair rides bicycle as man takes picture
VGG16 Base + Dropout + Kernel Regularizer:	man in red shirt and hat is riding unicycle down the street

Actual vs Predicted Captions



Actual:

brown dog walks through snow

brown puppy walking through the snow

dog looks curious at adventures lying ahead in the snow

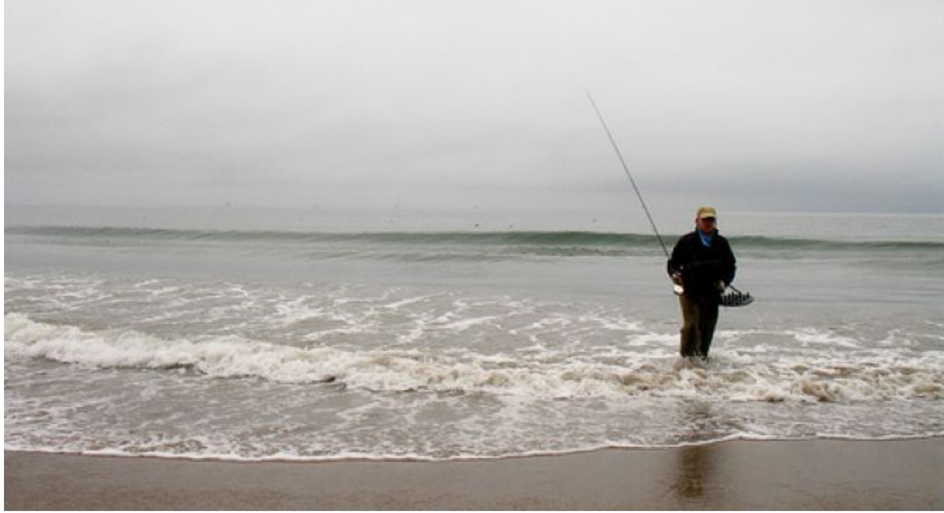
dog walking in the snow

yellow puppy walking through the snow

Predicted:

Inception v3 Base:	dog is digging steaks into the thick snow
Inception v3 Base + Dropout:	two dogs are running through the snow
Inception v3 Base + Dropout + Kernel Regularizer:	dog is running through the snow

Actual vs Predicted Captions



Actual:

fisherman stands on the beach on gray day

man in yellow cap is on the beach carrying fishing pole

man with fishing pole standing on beach

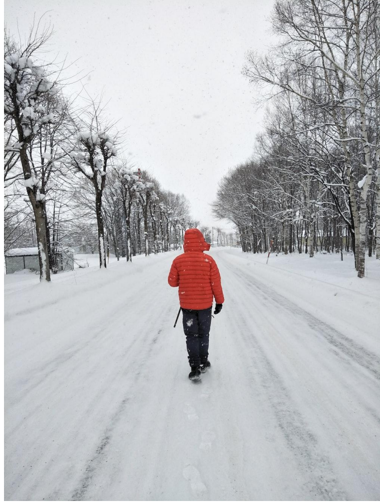
man holding fishing pole and tackle box walking in from the ocean

person is standing in the ocean fully clothed holding fishing pole in one hand and tackle box in the other

Predicted:

Inception v3 Base:	man is carrying fishing pole
Inception v3 Base + Dropout:	man in dark shorts is fishing in the ocean
Inception v3 Base + Dropout + Kernel Regularizer:	two boys in wetsuits surf

Captions for Unseen Images



Predicted:

VGG16 Base: the boy is jumping in the air

VGG16 Base + Dropout: small dog is running through the sand

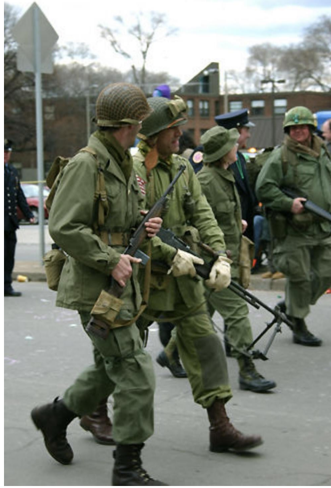
VGG16 Base + Dropout + Kernel Regularizer: the skier is running through the snow

Inception v3 Base: two children and one little boy are riding wheelchair in the snow

Inception v3 Base + Dropout: two people are walking on the side of the road holding purple car

Inception v3 Base + Dropout + Kernel Regularizer: man in red jacket is riding on the side of mountain

Captions for Unseen Images



Predicted:

VGG16 Base: boy in black shirt is running down sandy street

VGG16 Base + Dropout: little boy is running on the grass

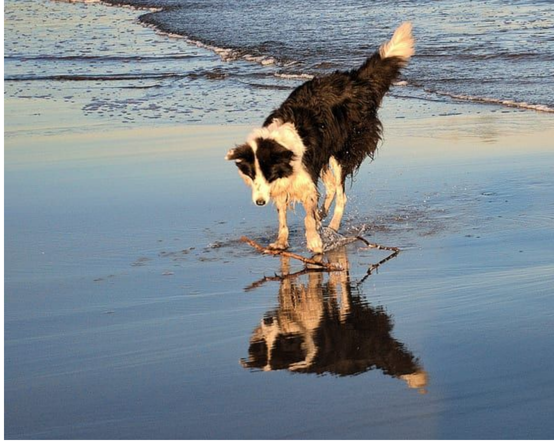
VGG16 Base + Dropout + Kernel Regularizer: the dog is running through the snow

Inception v3 Base: group of men in camouflage pants and hats playing with their hands in the air

Inception v3 Base + Dropout: man in uniform and two officers wearing hats on head head

Inception v3 Base + Dropout + Kernel Regularizer: group of people are standing in front of graffiti angry and white vehicles

Captions for Unseen Images



Predicted:

VGG16 Base: black and white dog is walking through the water

VGG16 Base + Dropout: small white dog is jumping over an orange gate

VGG16 Base + Dropout + Kernel Regularizer: dog is running through the snow

Inception v3 Base: dog is running through the water with large chunk of snow behind him

Inception v3 Base + Dropout: black and white dog is running on the beach

Inception v3 Base + Dropout + Kernel Regularizer: two dogs are running through the water

Summary and Improvements

Summary

- It is highly subjective as to what is the right interpretation of an image
- EDA of captions gave a good understanding of the images that are in the dataset and what to expect in the results
- Some of the captions are not useful in training the model
- Overfitting is greatly reduced after introducing Dropout and Kernel Regularizer
- Improving the generalisation resulted in a decrease in overall accuracy
- BLEU Score is perhaps not a very good indicator of how well the model is performing
- Training of models took a long time and depends a lot on luck

Improvements

- Available applications (for Transfer Learning) on TF Keras are trained on ImageNet dataset. Explore using Yolo pretrained model (trained using MS Coco dataset)
- Extract features using a larger dataset (30K images dataset on Kaggle)
- Test out more values in terms of the Dropout Rate and try out different Kernel Regularizer (Lasso, Elastic Net) and Activation Functions (ReLU vs GeLU/Swish)
- Explore manipulating different batch sizes, learning rate, train/test split values in the training process
- Apply other Evaluation Metrics e.g. METEOR/ROUGE
- Compare results with a Image-to-Speech model

Questions?