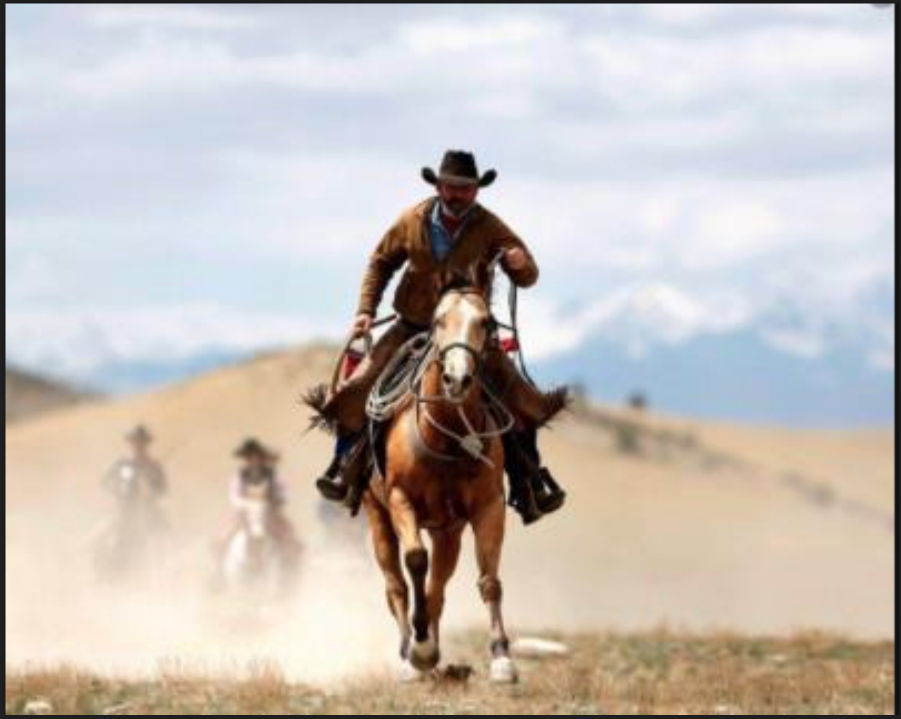
Image Captioning-Using Multimodal Neural Networks

Sampath Vaddadi

Motivation

Ease of Humans in generating captions



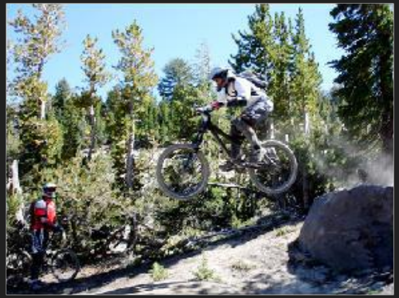
- A man is riding a horse.
- A cowboy is on the horse.
- A horse is being ridden by a man with hat.

How can computers replicate this?

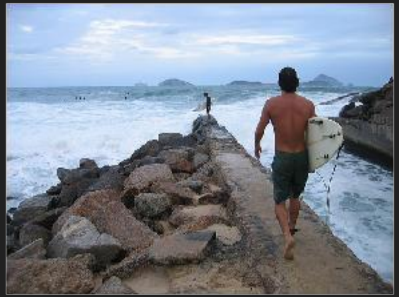
- Greatest challenge once upon a time.
- Wide scale adaptation of deep-learning, it is solved to some extent.
- Requires large variety of images with their captions and huge compute power.
- Alternative: for a small dataset use a pre-trained model (transfer learning) and fine-tune the model for satisfactory results.

Dataset Description & Exploration

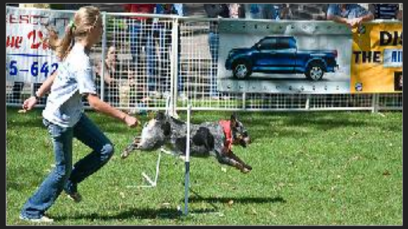
- We use Flicker 8k dataset.
- This dataset contains 8000 images (hence the name 8k) and each image has 5 captions telling us what is happening in the image. These 8000 images are split as follows:
  - 1. 6000 Training images and their descriptions.
  - 2. 1000 Development/Validation images and their descriptions.
  - 3. 1000 Test images and their descriptions
- The dataset comes with a separate text file listing the names of images that go into 3 groups: training, validation/development(dev) and testing. Using this file, the dataset is split into train, dev and test datasets.



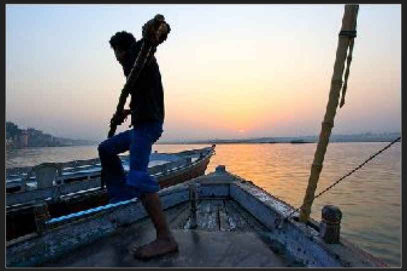
[‘helmeted man jumping off rock on mountain bike’, ‘man jumping on his bmx with another bmxer watching’, ‘mountain biker is jumping his bike over rock as another cyclist stands on the trail watching’, ‘person taking jump off rock on dirt bike’, ‘the bike rider jumps off rock’]



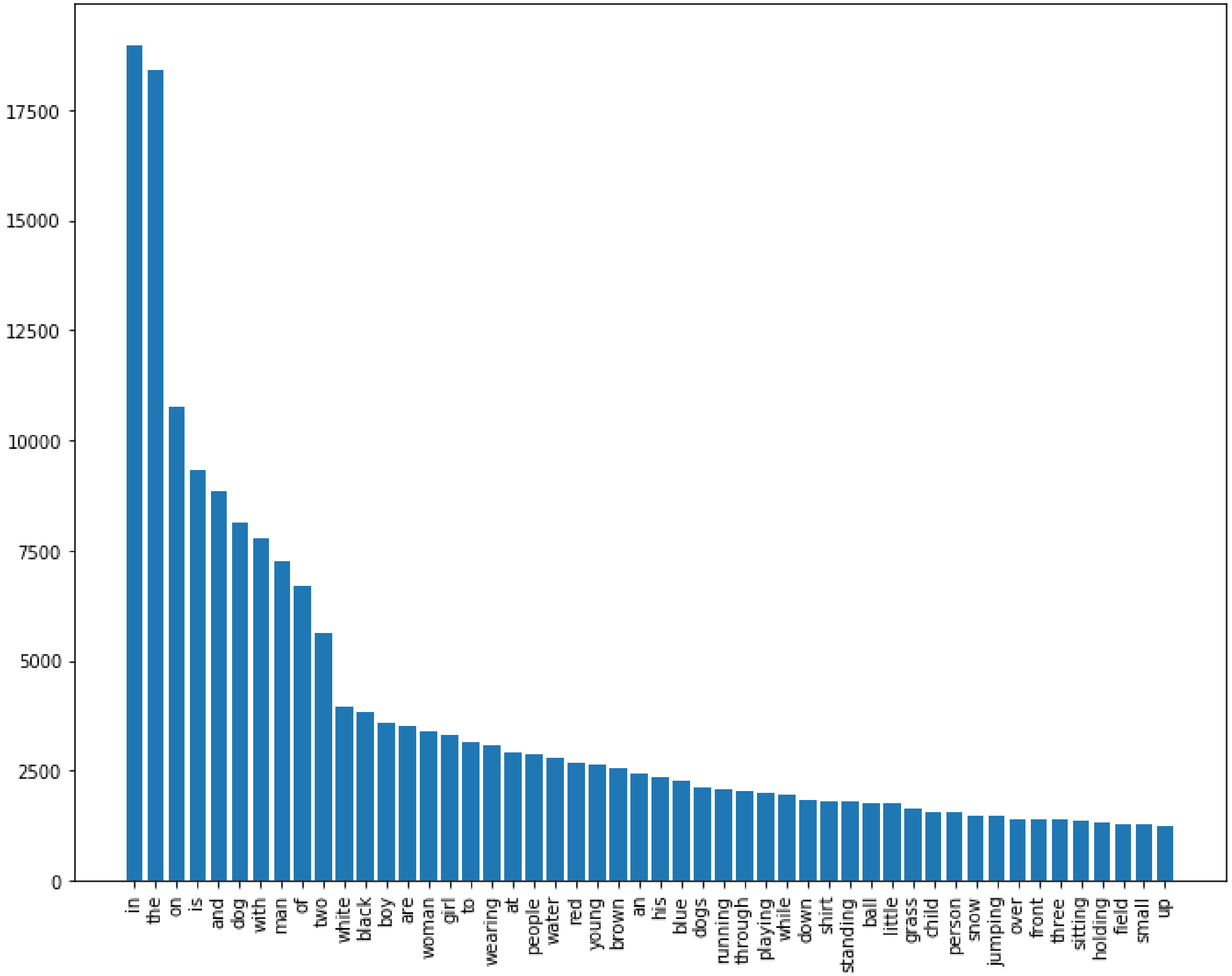
[‘man is carrying surfboard along stone wall that leads to the ocean’, ‘man in green swim trunks carrying surfboard along rock wall to the ocean’, ‘surfers walking along seawall as the ocean churns around them’, ‘two surfers carry their boards to the end of walkway to join the others in the water’, ‘two surfers walk along rock wall to reach the waves’]



['dog and his trainer at challenge', 'dog with bandanna jumps through hoop', 'girl leads her dog over hurdle', 'woman leading grey and black spotted dog in red bandanna through competition', 'woman runs alongside her dog as he jumps hurdle']



['man is docking boat', 'man is using long pole to push two boats apart', 'man on boat is using large stick to push another boat away', 'man on fishing boat pushing wooden pole into the water', 'man standing on boat is pulling another boat close by with bamboo stick']

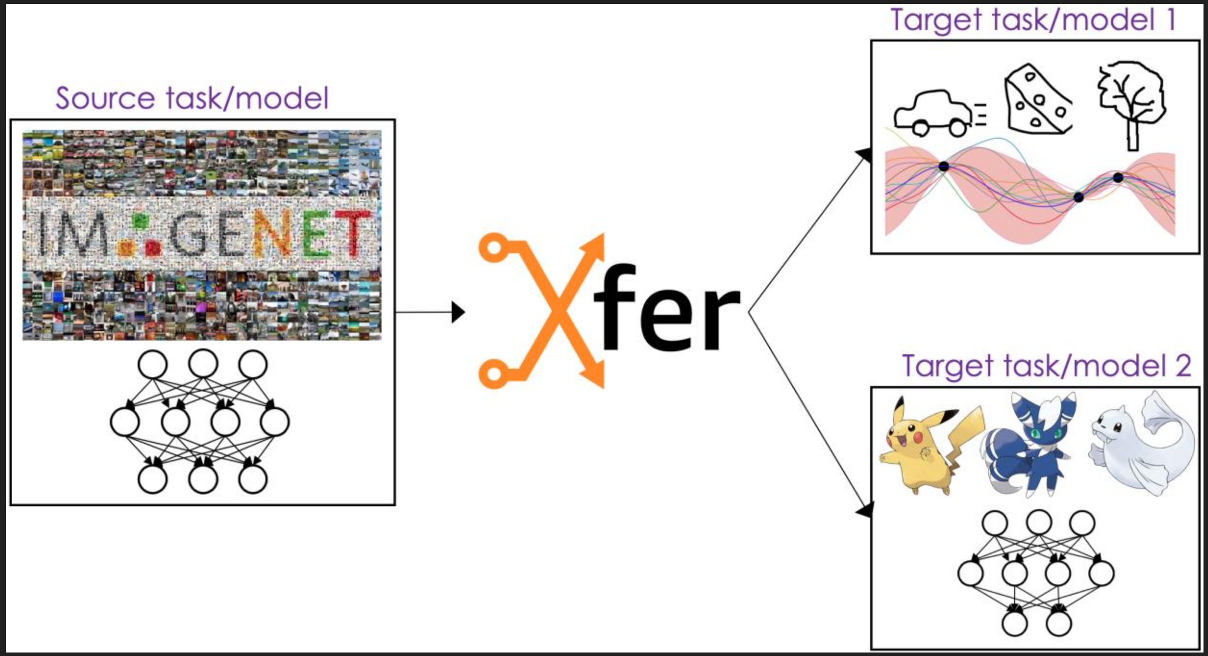

17500
15000
12500
10000
7500
5000
2500
0
in
the
on
is
and
dog
with
man
of
two
white
black
boy
are
woman
girl
to
wearing
at
people
water
red
young
brown
an
his
blue
dogs
running
through
playing
while
down
shirt
standing
ball
little
grass
child
person
snow
jumping
over
front
three
sitting
holding
field
small
up

Data Preprocessing

- Three main steps:
  - Image data preprocessing
  - Text data preprocessing
  - Creating input-output pairs

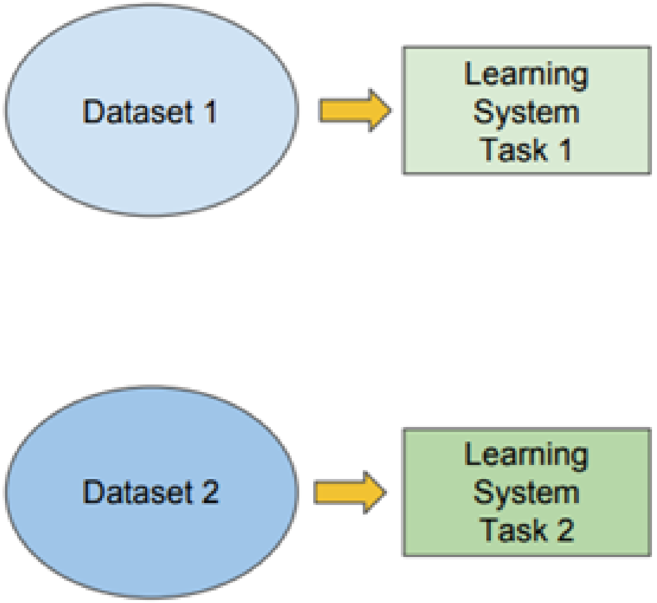
Image data preprocessing

- Used InceptionV3 pre trained network from TF Hub.
- Removed last layer of the network (used in image classification task-Image net challenge)
- Used embeddings coming out of the penultimate layer to transform images into vectors

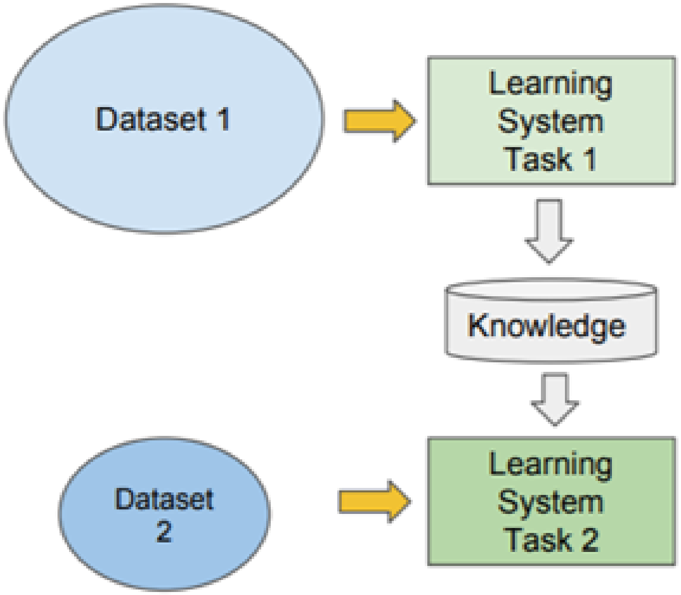


Traditional ML vs Transfer Learning

- Isolated, single task learning:
  - Knowledge is not retained or accumulated. Learning is performed w.o. considering past learned knowledge in other tasks

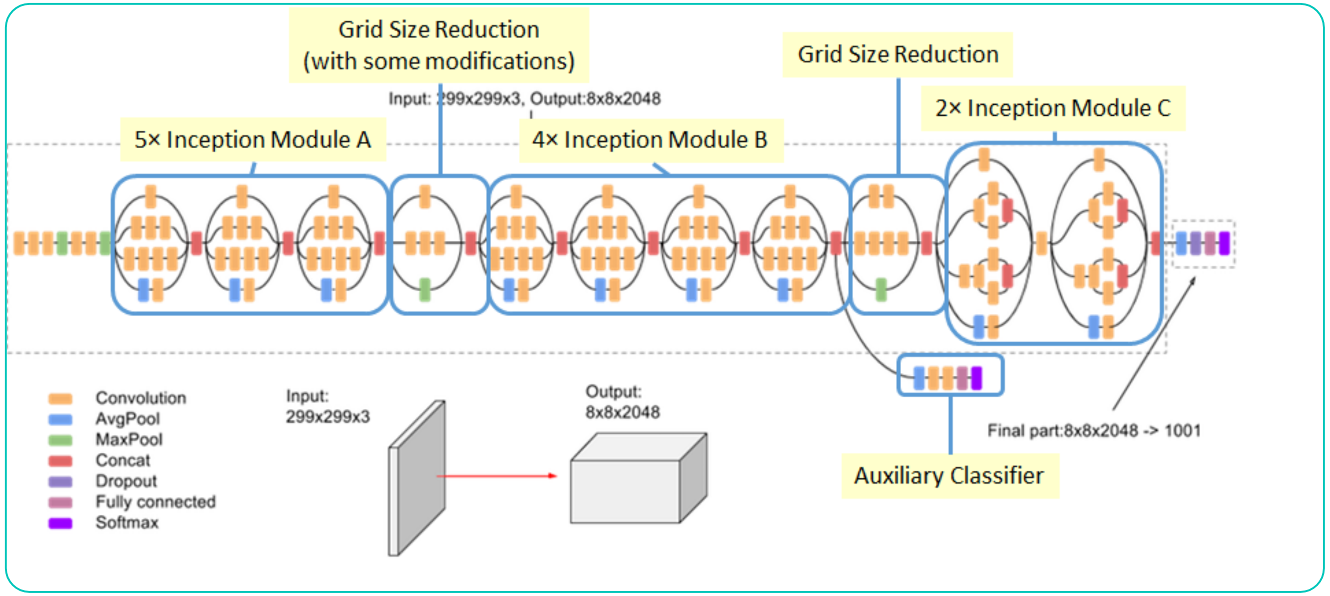


- Learning of a new tasks relies on the previous learned tasks:
  - Learning process can be faster, more accurate and/or need less training data

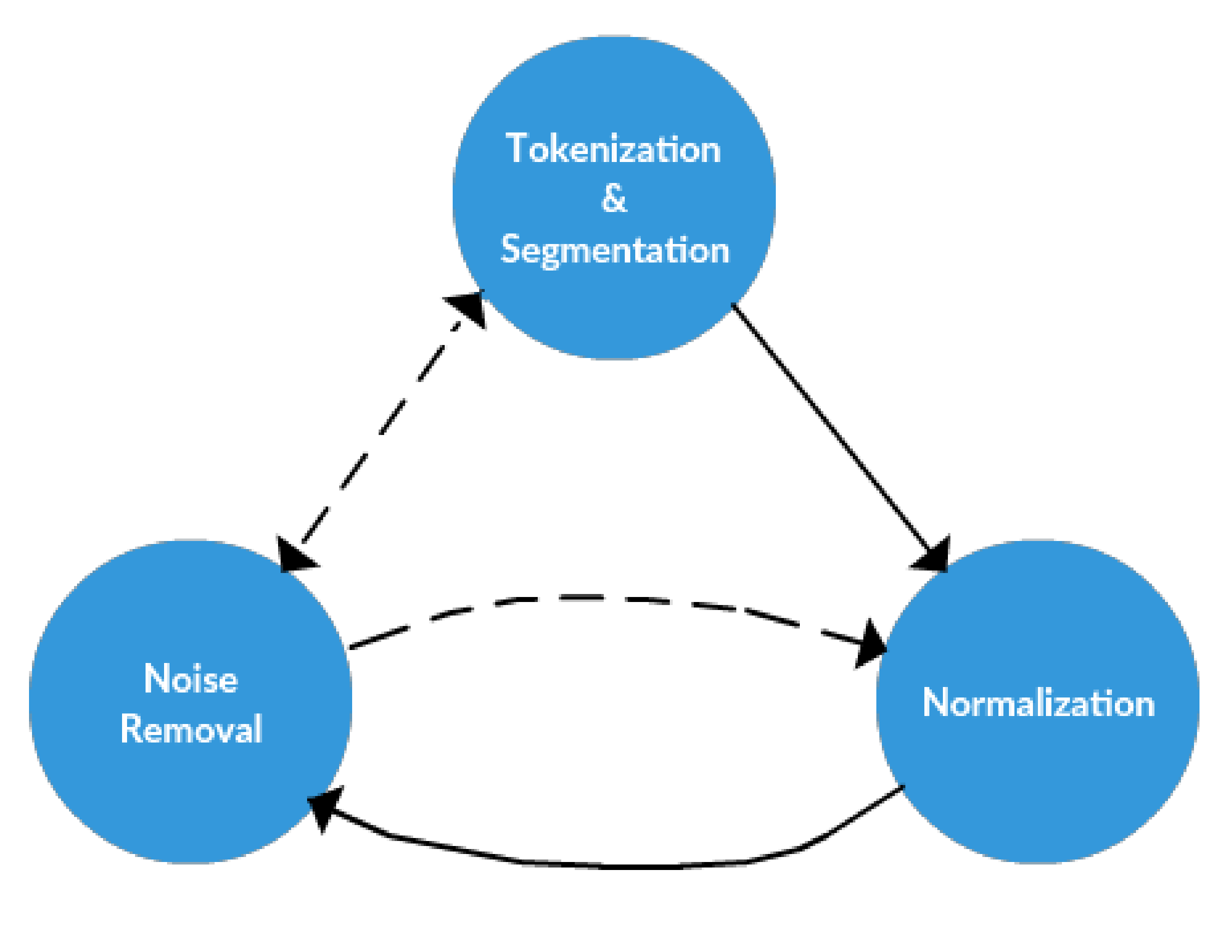


InceptionV3

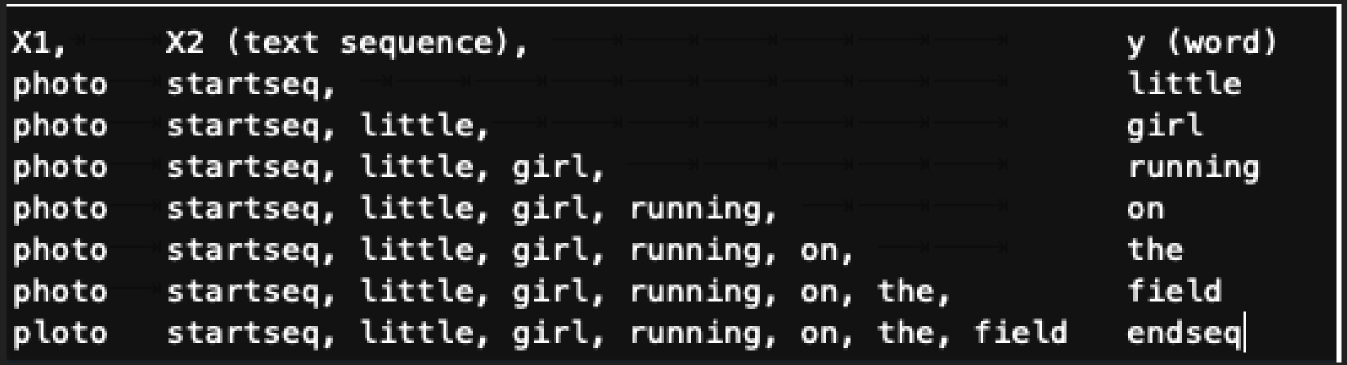
- Developed by google and is the first runner up of image net competition in 2015.
- Has 42 layers. This model has been trained on large dataset of 1.2 million images belonging to 1000 categories.
- If we remove the last layer) and pass in an image, we get a 2048-dimensional image (this is the no. of units in the Dense layer).
- This 2048-dimensional vector can be considered as a very high-level representation of an image which might not make sense to us.



Text data preprocessing



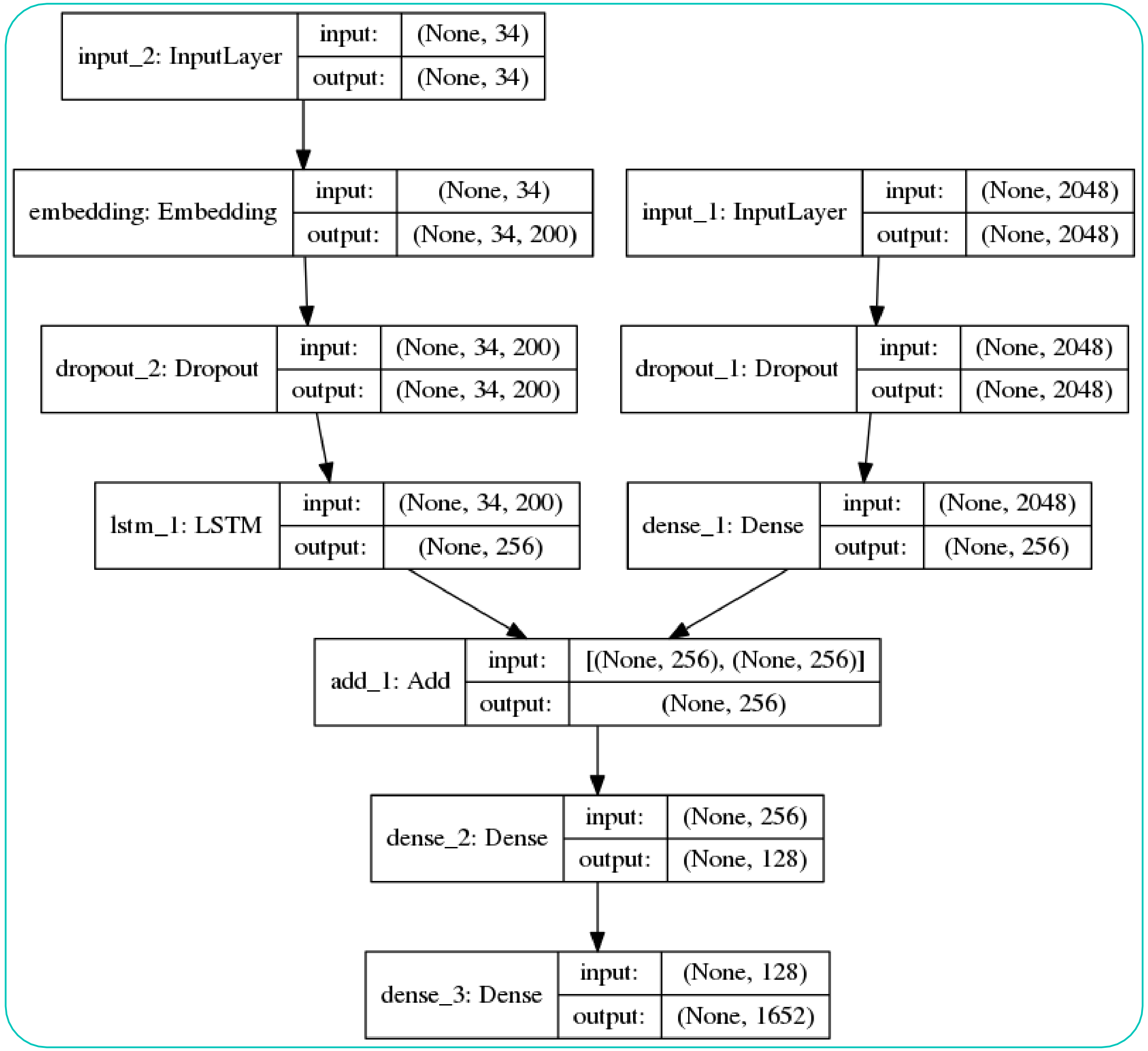
Creating input-output pairs



```
X1,     X2 (text sequence),                                  y (word)
photo   startseq,                                            little
photo   startseq, little,                                    girl
photo   startseq, little, girl,                              running
photo   startseq, little, girl, running,                     on
photo   startseq, little, girl, running, on,                 the
photo   startseq, little, girl, running, on, the,            field
ploto   startseq, little, girl, running, on, the, field      endseq
```

- Each time, the network is supplied with the photo (embedding).
- Along with the embedding, we also send GloVe embeddings of n words in the sentence and predict the (n+1)th word embedding, where n=1 to N.

Multi-modal Neural Network



input_2: InputLayer	input:	(None, 34)
	output:	(None, 34)

embedding: Embedding	input:	(None, 34)
	output:	(None, 34, 200)

input_1: InputLayer	input:	(None, 2048)
	output:	(None, 2048)

dropout_2: Dropout	input:	(None, 34, 200)
	output:	(None, 34, 200)

dropout_1: Dropout	input:	(None, 2048)
	output:	(None, 2048)

lstm_1: LSTM	input:	(None, 34, 200)
	output:	(None, 256)

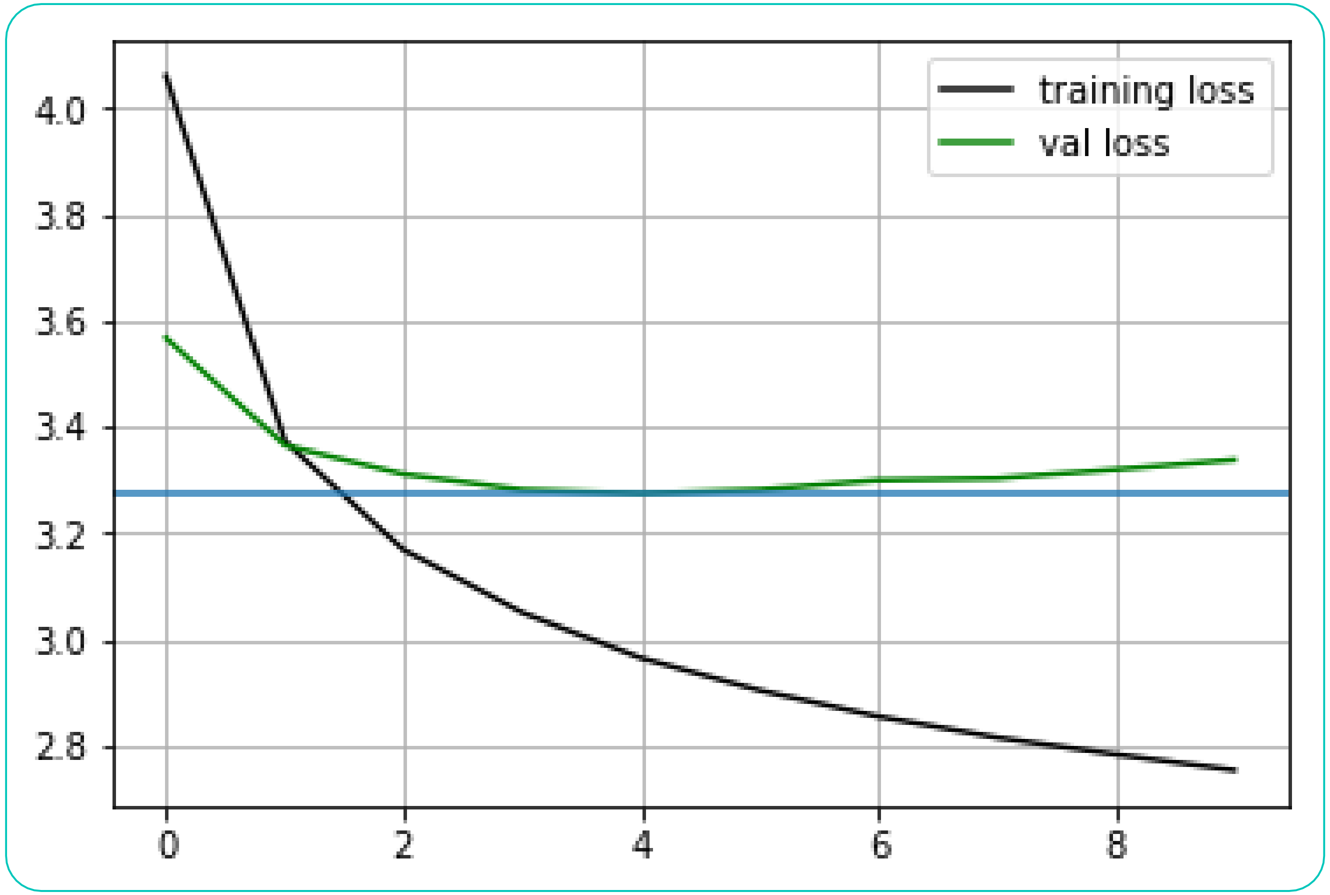
dense_1: Dense	input:	(None, 2048)
	output:	(None, 256)

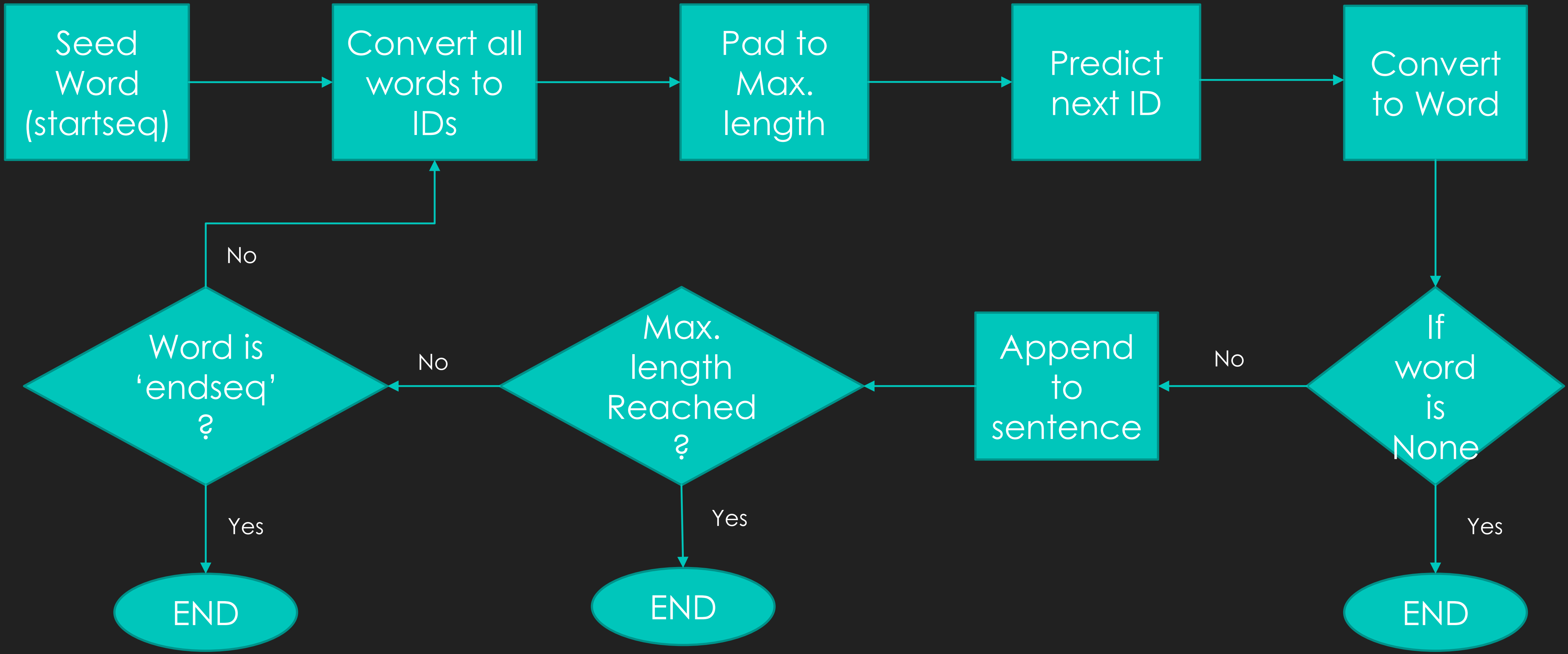
add_1: Add	input:	[(None, 256), (None, 256)]
	output:	(None, 256)

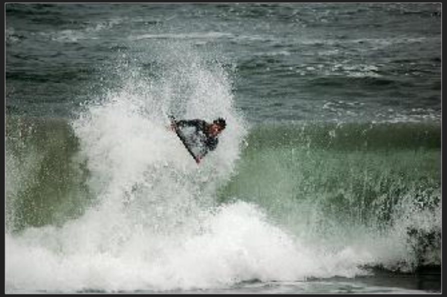
dense_2: Dense	input:	(None, 256)
	output:	(None, 128)

dense_3: Dense	input:	(None, 128)
	output:	(None, 1652)

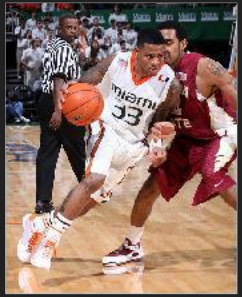
Training and Validation loss



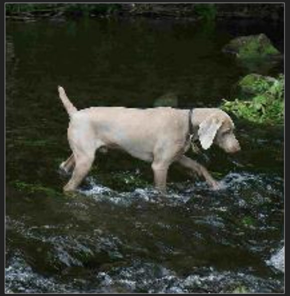

Seed Word (startseq)
Convert all words to IDs
Pad to Max. length
Predict next ID
Convert to Word
If word is None
Yes
END
No
Append to sentence
Max. length Reached ?
Yes
END
No
Word is 'endseq' ?
Yes
END
No



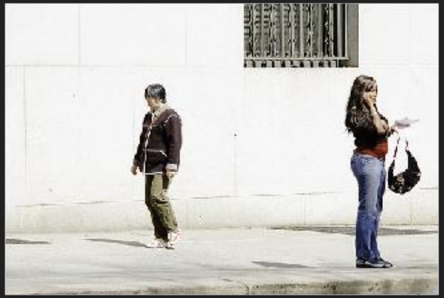
startseq surfer is jumping over wave endseq



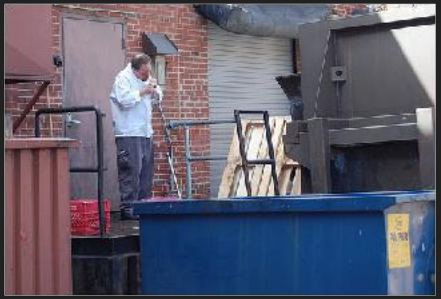
startseq basketball player in red is running endseq



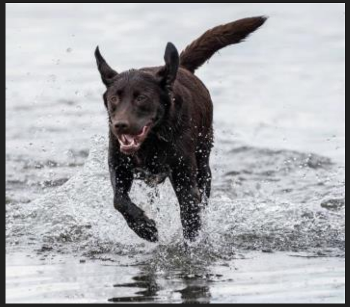
startseq dog is running through water endseq



startseq woman in black shirt and white shirt is standing in front of white building endseq



startseq man in black jacket is standing on sidewalk endseq



startseq black dog is swimming in water endseq

Future Work

Some ideas to improve the performance of the model are:

- Using a larger dataset like the MS-COCO data. This ensures that our model is exposed to a wide variety of images and descriptions and hence perform better during inference.
- Doing more hyper parameter tuning (learning rate, batch size, number of layers, number of units, dropout rate, batch normalization etc.) like a grid search using dev set.
- Using Beam Search instead of Greedy Search during Inference.
- Using attention based and/or Bi-Directional LSTM to improve the network i.e. make it better to process long sequences/sentences.

Applications

- Captions scenes in CCTV footage.
- Medical Image diagnosis (aid to the physician).
- Analysis of many images and cluster similar ones (google image search).
- An aid for the blind.
- Boost to self driving cars (Tesla).

References

- https://arxiv.org/pdf/1810.04020.pdf
- https://www.aclweb.org/anthology/L18-1273.pdf
- https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/
- https://medium.com/mlreview/multi-modal-methods-image-captioning-from-translation-to-attention-895b6444256e
- https://towardsdatascience.com/multimodal-deep-learning-ce7d1d994f4
- https://towardsdatascience.com/image-captioning-with-keras-teaching-computers-to-describe-pictures-c88a46a311b8