IMAGE CAPTIONING USING NEURAL NETWORKS

Presented by Linda Cho

LINDA CHO

Data Scientist

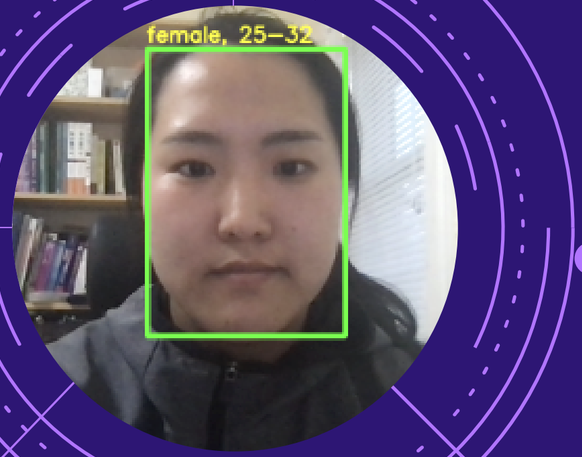
VETERAN

Spent countless hours in search of the perfect picture and caption for storyboards

MOTHER

Struggling to come up with good captions for picture books

MOTIVATION



COMPUTER VISION

NATURAL LANGUAGE PROCESSING (NLP)

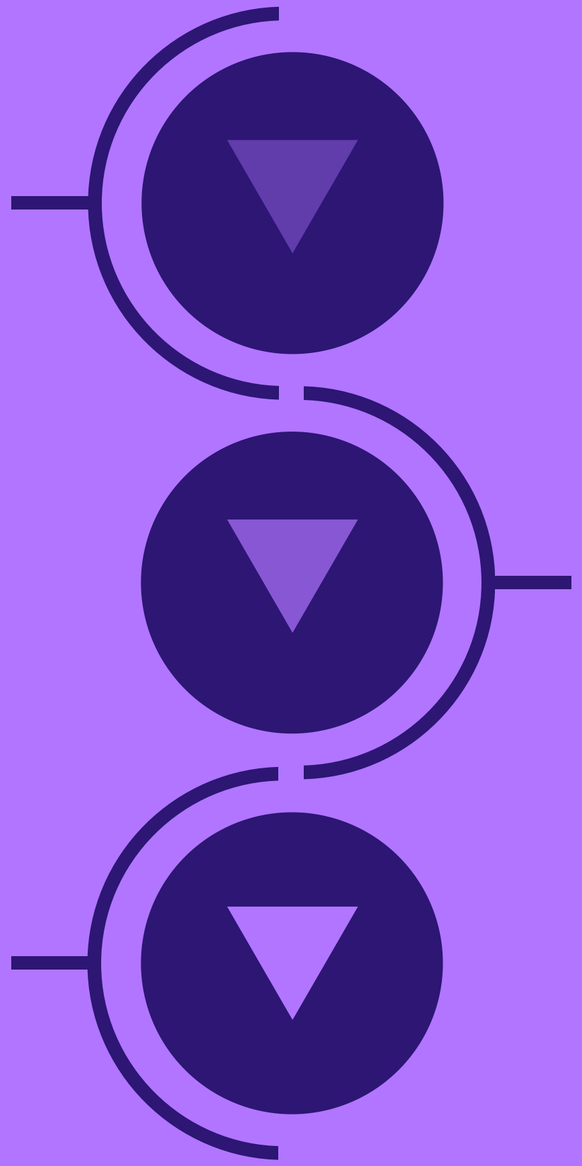
- VISUALIZE WRITTEN CUSTOMER DESIRES
- ASSISTANCE TO VISUALLY AND HEARING IMPAIRED
- DESCRIBE MEDICAL IMAGES

..AND SO MUCH MORE

GOAL: GENERATE IMAGE CAPTIONS

THE DATA

MS COCO Data Set (2014)
50,000 Images and Captions



THE TOOLS

Python
Tensorflow
Google Colab

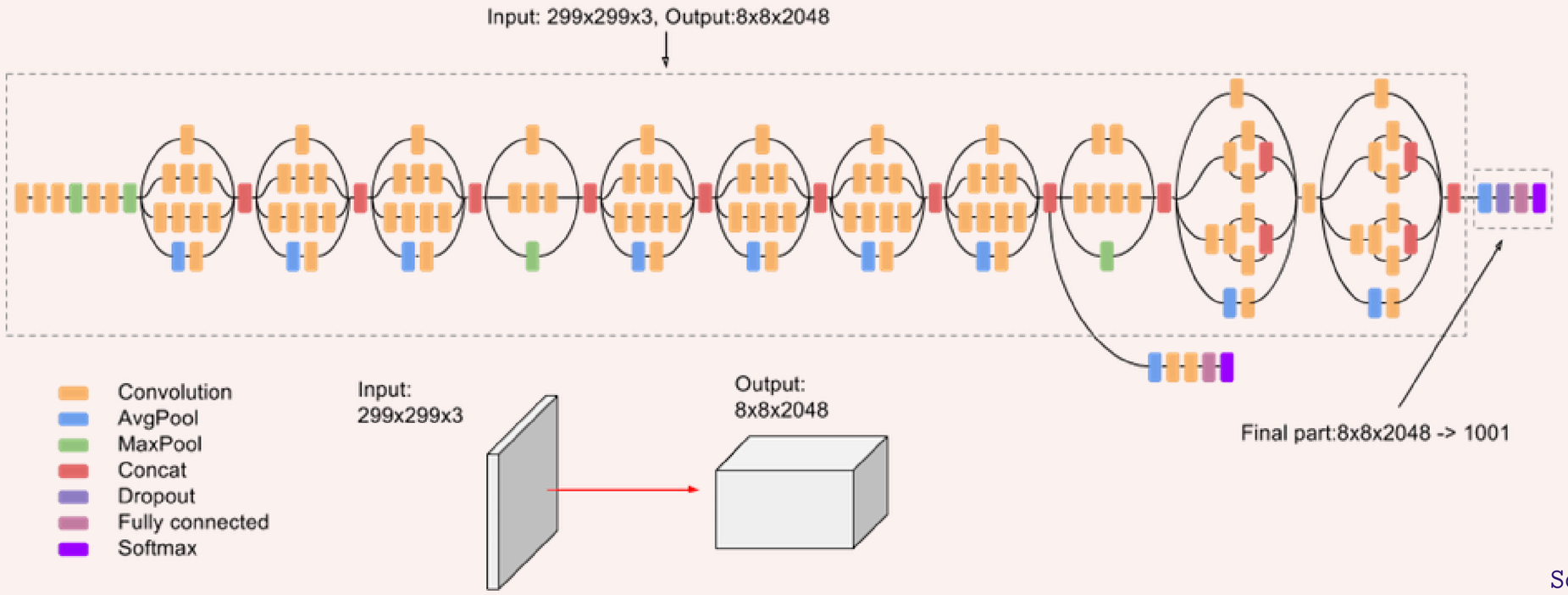
THE MODELS

Inception V3
CNN Encoder
Bahdanau Attention
Gated Recurrent Unit RNN Decoder

IMAGE PRE-PROCESSING

- Inception V3 is a pretrained convolutional neural network (CNN) for object classification
- Trained on millions of images from ImageNet dataset
- 48 layers, 1000 categories
- Feature representation vector at last convolutional layer is used as input to the CNN encoder

Inception V3 Architecture



Source: Google Cloud

CAPTION PRE-PROCESSING

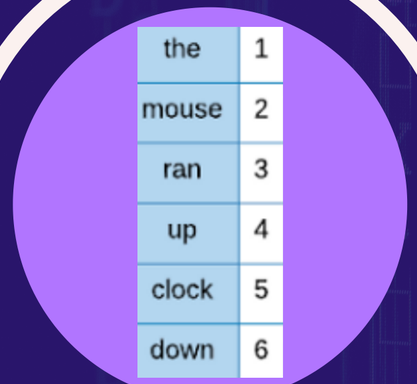
CAPTION

THE MOUSE RAN DOWN

TOKEN SEQUENCE

['THE', 'MOUSE', 'RAN', 'DOWN']

VOCABULARY



the	1
mouse	2
ran	3
up	4
clock	5
down	6

NUMBER SEQUENCE

[1, 2, 3, 6]

PADDING

[1, 2, 3, 6, <pad>,<pad>]

Note: All sequences must be same length for decoder. Example max length = 6

Source: Google Developers

CNN ENCODER

Input:
Image feature vectors from Inception V3

Process:
Pass through a single fully connected layer with ReLU activation (embed_dim = 256)

Output:
Richly encoded image

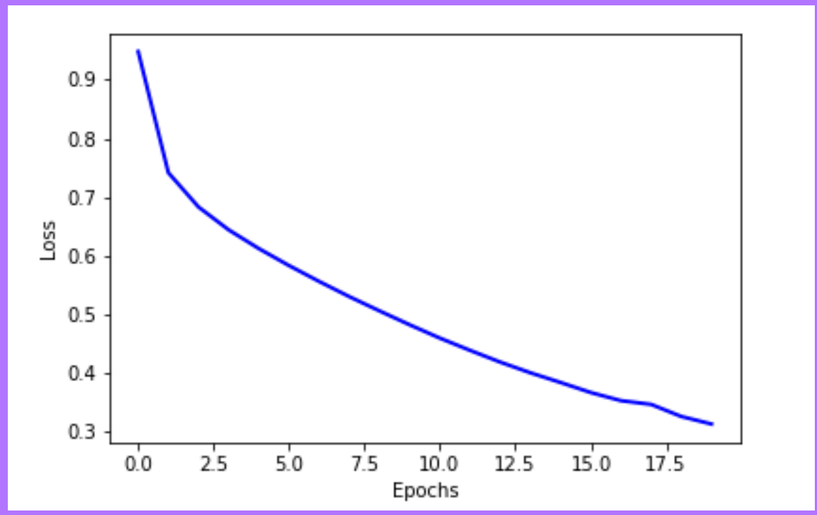
RNN DECODER

Bahdanau Attention uses encoded features vectors to output the context vector and attention weights.

Gated Recurrent Unit (GRU) uses Attention output to sequentially predict words.

Output:
Caption Prediction

LOSS PLOT



- 20 EPOCHS
- ADAM Optimizer
- Loss >30 %

GOOD PREDICTIONS

Actual Caption:
a baseball player is ready to hit the incoming ball

Predicted Caption:
a baseball game with a baseball bat on a field

Actual Caption:
a man on a snow board making a high jump

Predicted Caption:
a young skiers skiing down rocks in a mountains

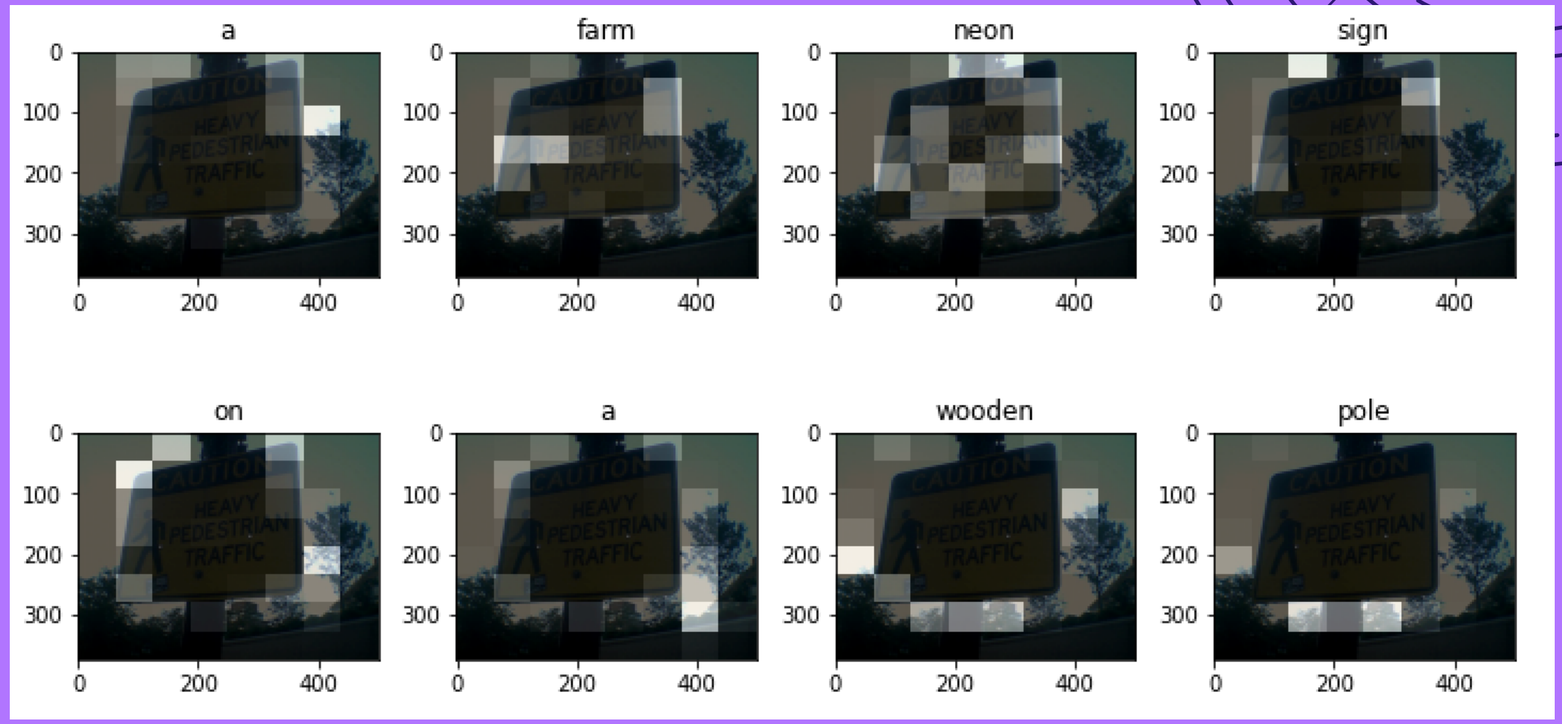
Actual Caption:
a stop sign at the corner of the road

Predicted Caption:
a traffic sign at a street corner

ATTENTION PLOT EXAMPLE

Actual Caption:
a sign that warns of heavy pedestrian traffic

Predicted Caption:
a farm neon sign on a wooden pole



- Understandable error for "neon"
- Picked up on "sign"
- Correct identification of "pole"

REPEATED WORDS

Actual Caption:
a baseball player swinging a bat at a ball

Predicted Caption:
a young man swings at a ball on a ball

Actual Caption:
a group of zebra's and a giraffe eating hay from a trough

Predicted Caption:
a giraffe and zebras standing tall giraffe

RANDOM IMAGES

Predicted Caption:
image of a ball in a ball in a ball

Predicted Caption:
two girls sitting at screen of an intersection

Predicted Caption:
a man reads a nice giraffe with glasses if a tie at his face

Observations:

- Repetition issues in caption prediction
- Training dataset was from 2014 and likely did not include images of video chatrooms
- Giraffes seemed to be a prevalent object detection

PERSONAL PHOTOS

Predicted Caption:
a woman laying down on a book and a book in a wall

Predicted Caption:
a soldier watching the army aircraft is standing at the field looking on

Predicted Caption:
a woman with a red hair and a toothbrush with a teddy bear

CONCLUSION

- Image captioning model is not ready for deployment
- Potential Model Improvements:
  - Different tokenizers
  - Transformers
  - More epochs
- Excited to see what future problems deep learning with neural networks can solve
  -

Little by little, we're giving
sight to the machines.
First, we teach them to see.
Then, they help us to see better.

- FEI FEI LI
TED2015

GET IN TOUCH


github.com/choski23


linkedin.com/in/linda-cho-2009


linda.m.cho.mil@gmail.com