Caption Generation

Trent Kindvall
DTSA 5511 Final Presentation

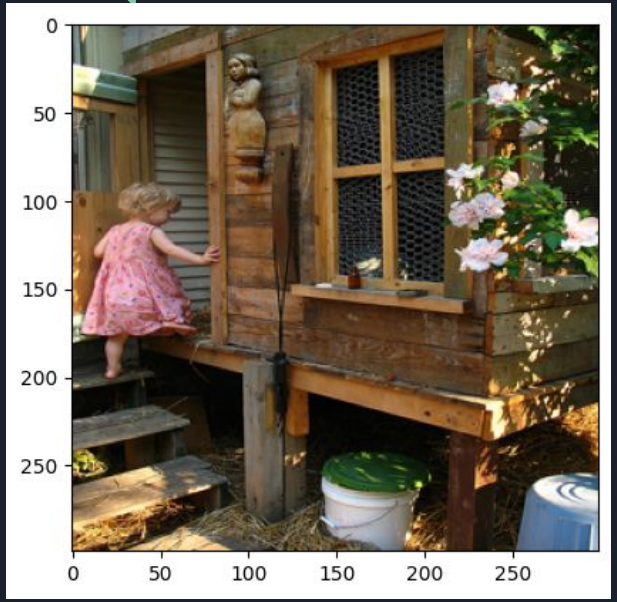
Caption Generation

- Computer vision
- Natural Language Processing (NLP)
- Can assist people with vision impairment

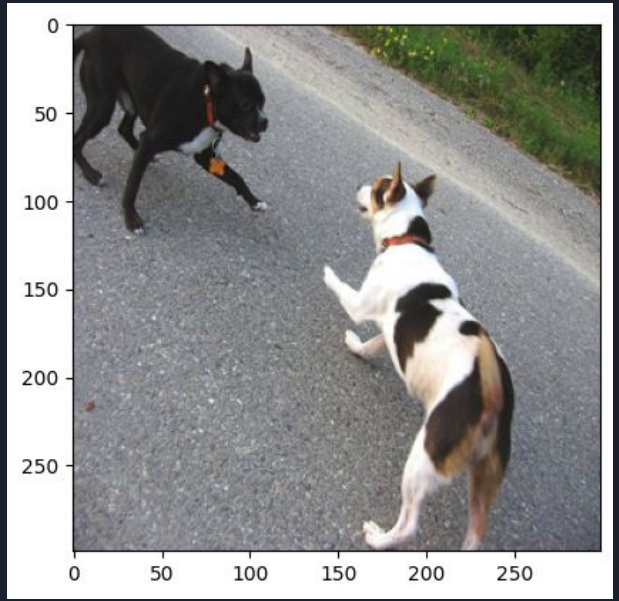
Flickr 8k Dataset

- ~8000 images with 5 captions describing the image

Images

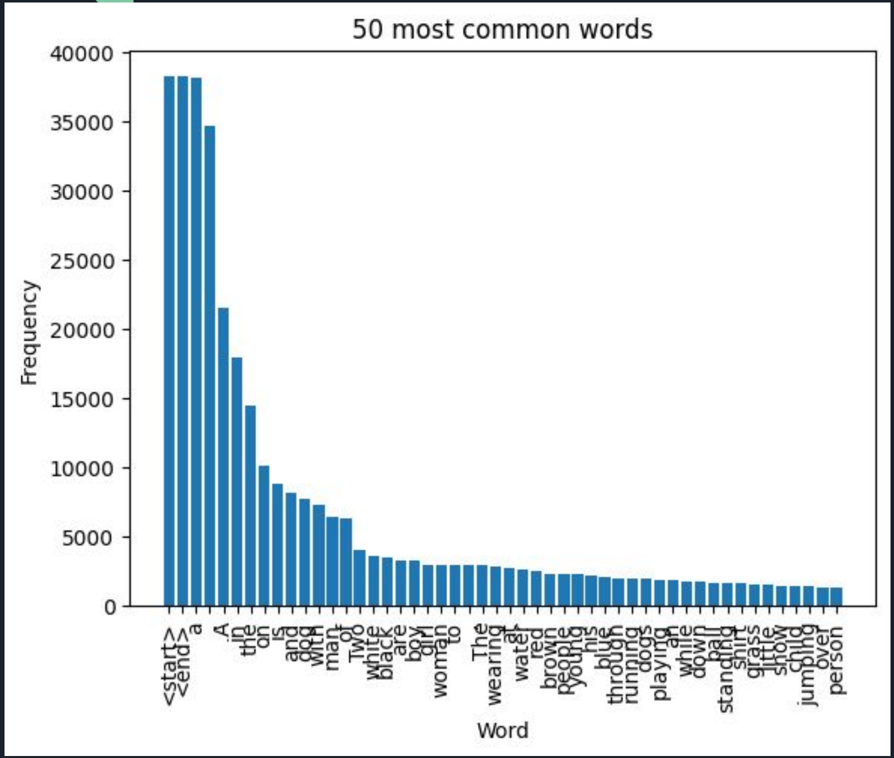


Caption:
A girl going into a wooden building .

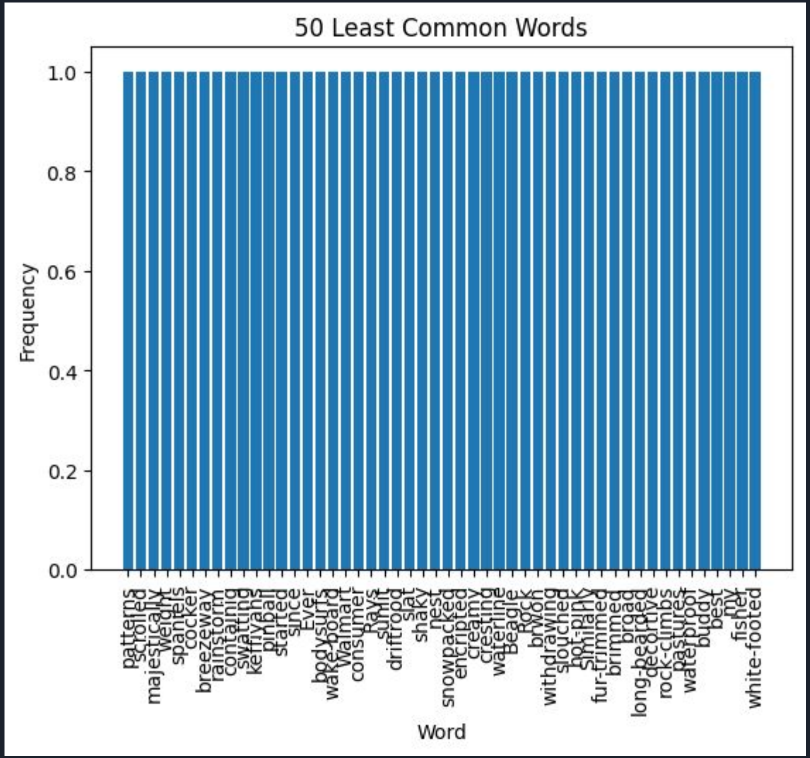


Caption:
A black dog and a tri-colored dog playing with each other on the road .

Captions



50 most common words
Frequency
40000
35000
30000
25000
20000
15000
10000
5000
0
Word
<start>
<end>
a
A
in
the
on
is
and
dog
with
man
of
Two
white
black
are
boy
girl
woman
to
The
wearing
at
water
red
brown
people
young
his
blue
through
running
dogs
playing
an
while
down
ball
standing
shirt
grass
little
snow
child
jumping
over
person

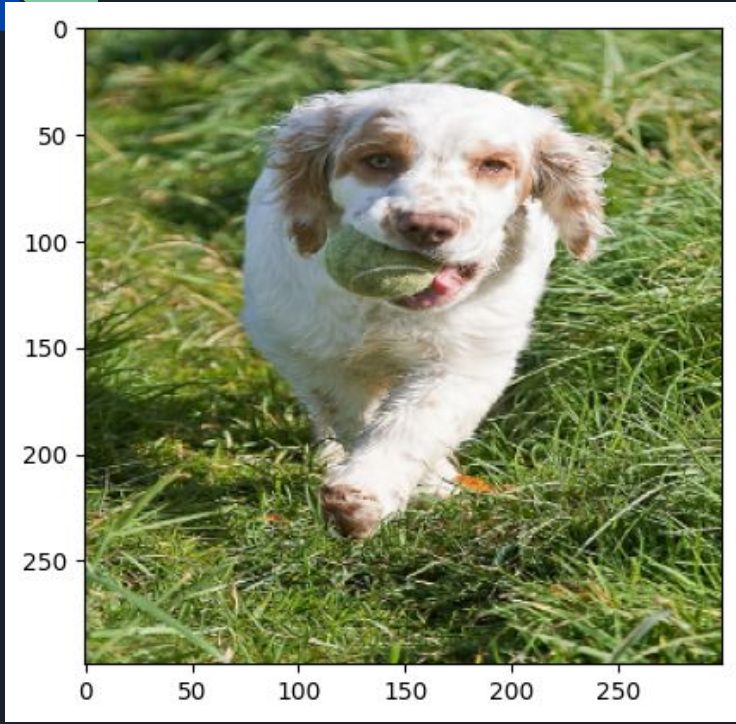


50 Least Common Words
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
Word
patterns
scrolled
majestically
spaniels
cocker
breezeway
rainstorm
swatting
pinball
started
since
Ever
bodysurfs
Walmart
consumer
Rays
sunlit
driftroad
slat
shaky
nest
snowpacked
encrypted
creamy
cresting
waterline
Beagle
Rock
brown
withdrawing
slouched
hot-pink
similarly
fur-trimmed
brimmed
broad
long-bearded
rock-climbs
pastures
waterproof
buddy
best
my
fisher
white-footed

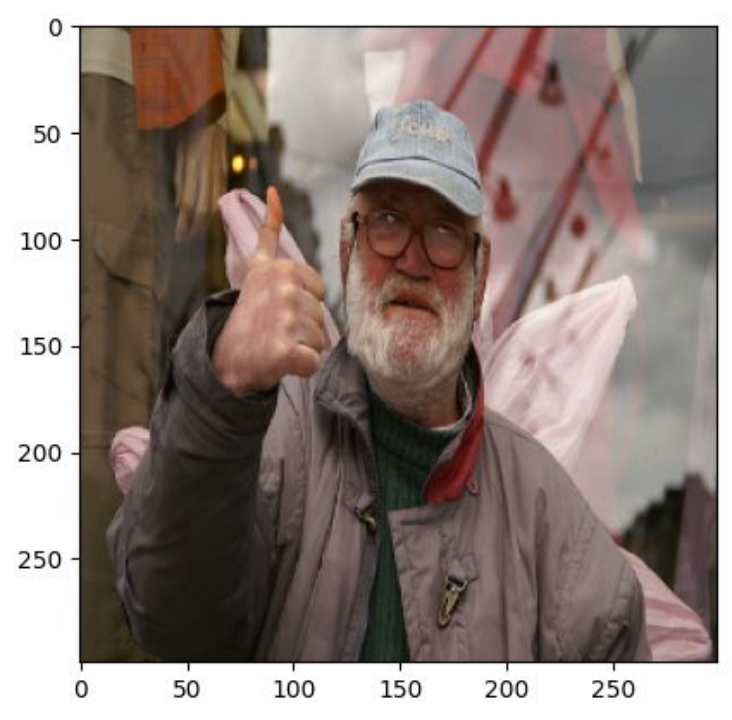
Model Architecture

- CNN
  - Image processing
  - Transfer learning
    - EfficientNet-b0
      - Can categorize images into a 1000 different categories
- Transformer Encoder
  - Translator between the cnn and the decoder
- Transformer Decoder
  - Generates the caption

Results

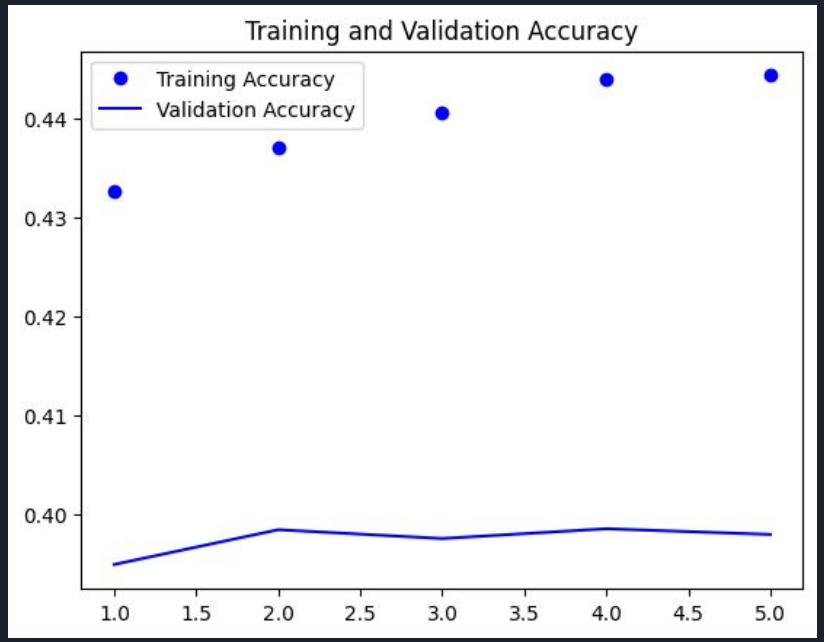


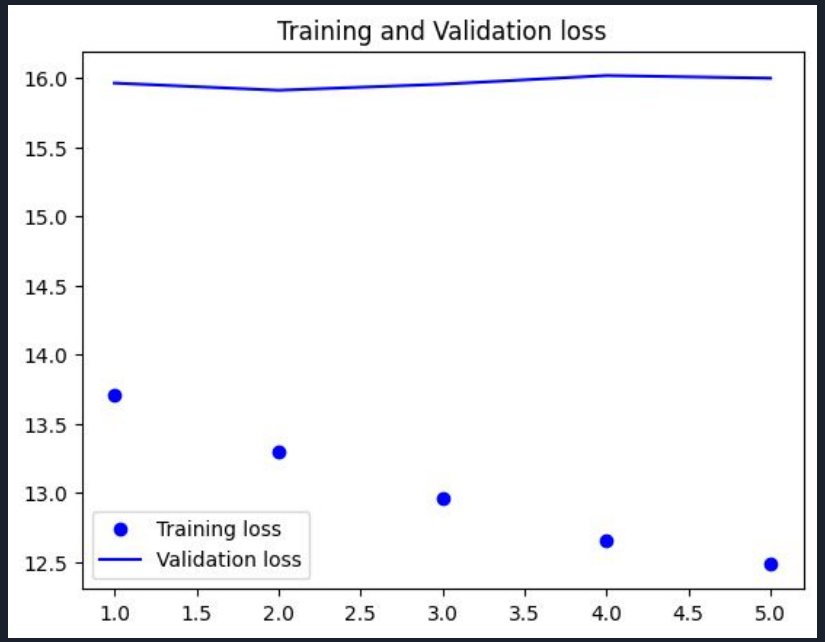
Predicted Caption: a white dog is running through the grass



Predicted Caption: a man wearing a hat and sunglasses is holding a sign

Results


Training and Validation Accuracy
Training Accuracy
Validation Accuracy
0.44
0.43
0.42
0.41
0.40
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Training and Validation loss
16.0
15.5
15.0
14.5
14.0
13.5
13.0
12.5
Training loss
Validation loss
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

BLEU Score

$$Bleu\ (N) = Brevity\ Penalty \cdot Geometric\ Average\ Precision\ Scores\ (N)$$

I achieved a score of .59 which was higher than a lot of the other attempts I saw on Kaggle

Conclusion

- Good results
- Fun use of deep neural networks and generative applications