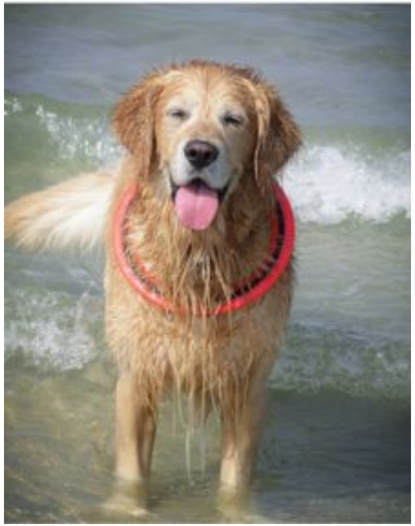
Image Captioning with Attention Mechanism

...and its Robustness to Adversarial Attacks

Abhranil Chakrabarti, Nikos Galanos, Arushi Jain, Xiaoyu (Chloe) Wu

Image Captioning



1. **Image understanding**
2. **Natural Language Generation**

A happy dog is standing in the ocean

Motivation

Broad Business Applications

- Provide image descriptions for visually impaired individuals
- Produce product descriptions at scale for e-commerce
- Create captions for social media posts

Cybersecurity

- Assess impact of adversarial attacks on captioning

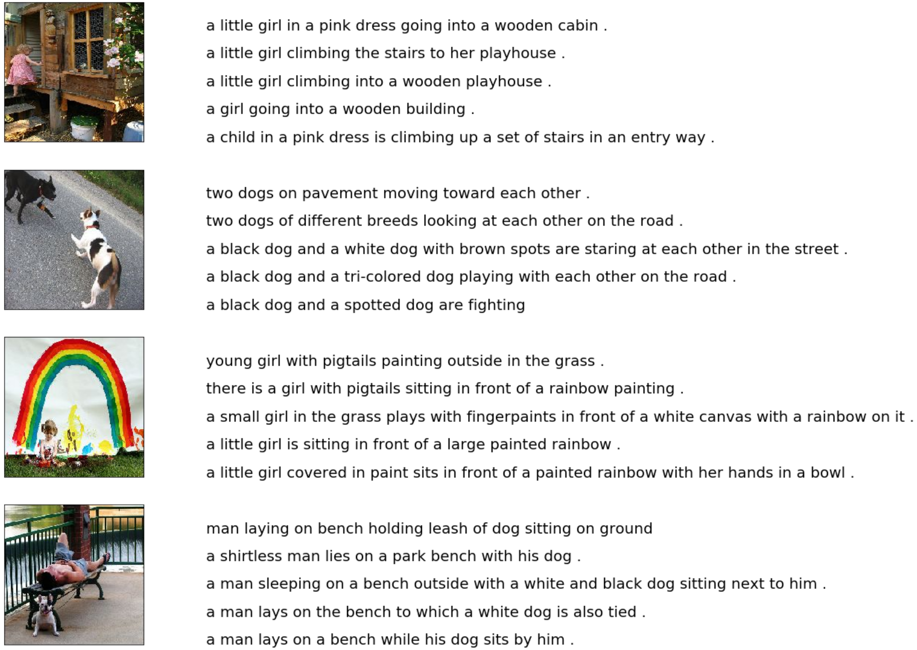
Data

Flickr 8k

- 8000 images from daily scenarios
- 5 captions for each image
- 75-25 train-test split

Preprocessing

- **Images**: Convert to 224 x 224 x 3 and RGB to BGR (required by VGG-16)
- **Captions**: Remove punctuations and stopwords, tokenize, add padding to make them of fixed length



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 .

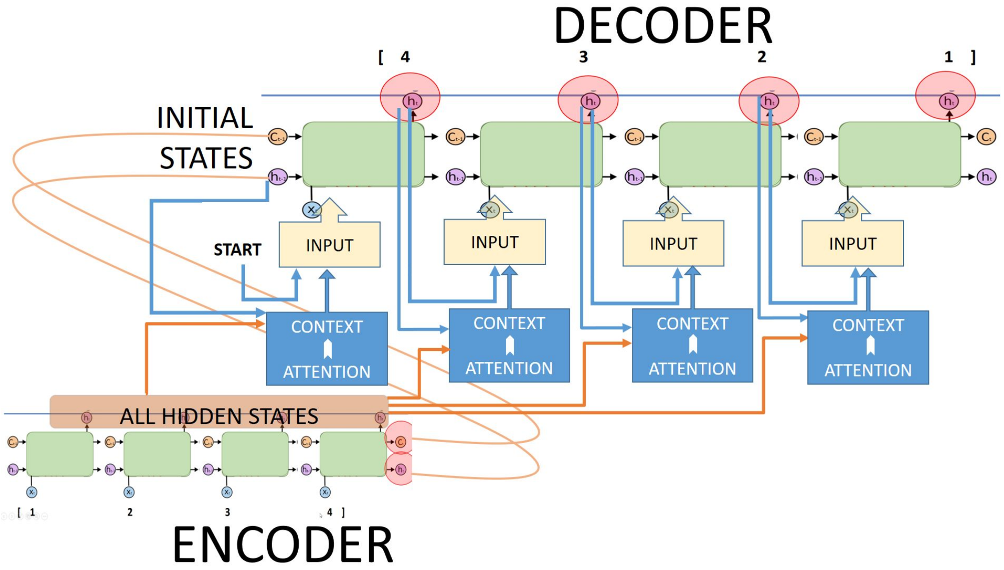
two dogs on pavement moving toward each other .
two dogs of different breeds looking at each other on the road .
a black dog and a white dog with brown spots are staring at each other in the street .
a black dog and a tri-colored dog playing with each other on the road .
a black dog and a spotted dog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

Figure: Examples of labeled Flickr 8k images

Model – Encoder Decoder Architecture with Attention



<u>Figure</u>: An illustration of Attention Mechanism

- **Feature Extraction from VGG-16:** Pass pre-processed images to VGG-16 and extract features from second last layer
- **Encoder:** Pass VGG-16 features through fully connected layers
- **Attention Mechanism:** Compute attention score between encoded image features and hidden state of previous layer of decoder, normalize encoded features by attention scores
- **GRU Decoder:** Pass normalized image features and caption sequence embeddings through fully connected layers

Prediction – Beam Search

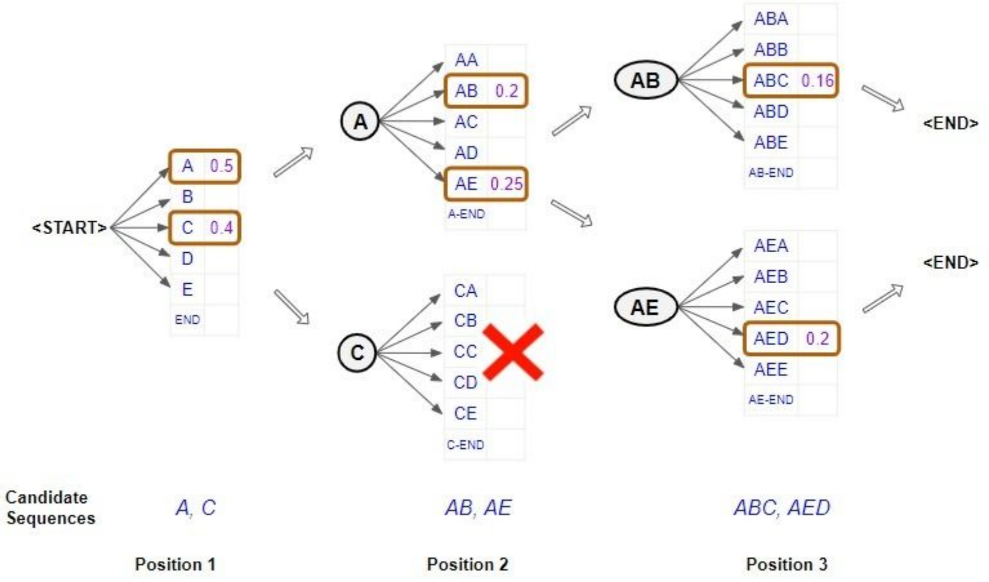


Figure: Caption Prediction with Beam Search

1) A possible list of tokens with corresponding probabilities generated at each time step

2) Select top k candidate tokens and generate the word at next time step

3) Keep repeating until <end> token is predicted or caption max length is reached

Results on Image Captioning

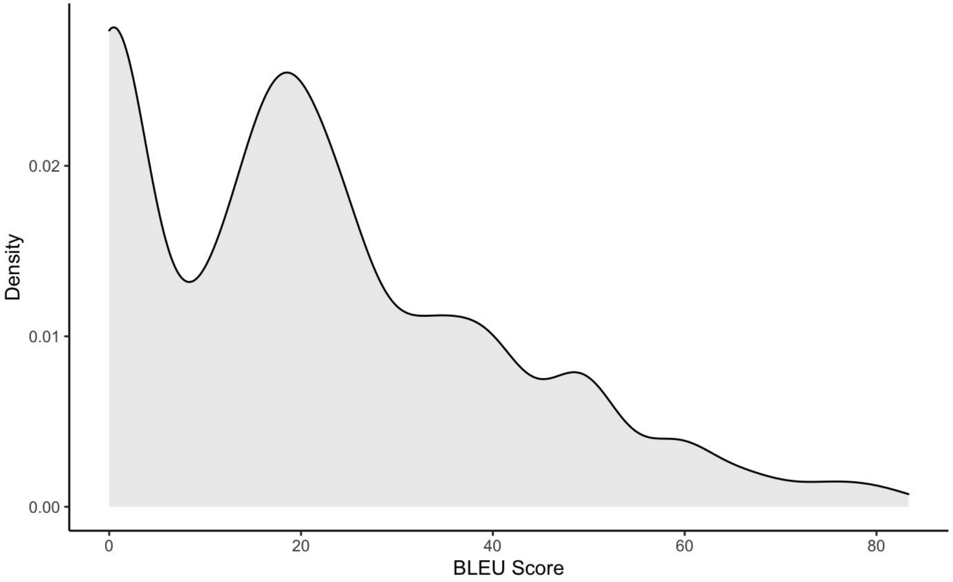


Figure: Distribution of BLEU Score on Test Set

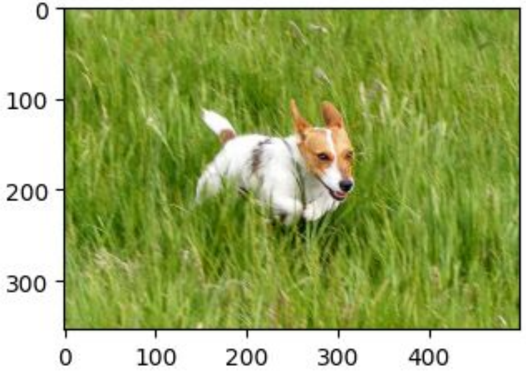
BLEU Score*

- The fraction of n-grams in the predicted sentence that appear in the ground-truth caption
- A third of test images have good-quality captions (score >=30)
- About half of the predicted captions are informative (score >= 20)

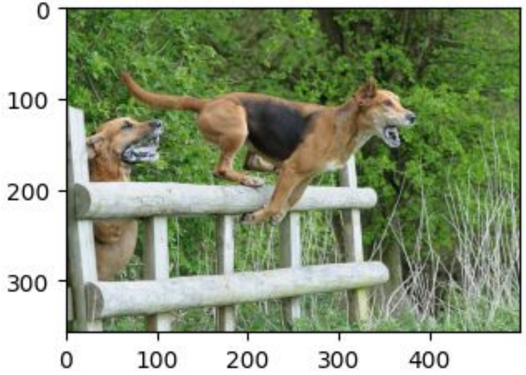
*Score = 10: almost useless | = 20: gist is clear but has grammatical error | > 30: good quality translations

Results on Image Captioning – Good

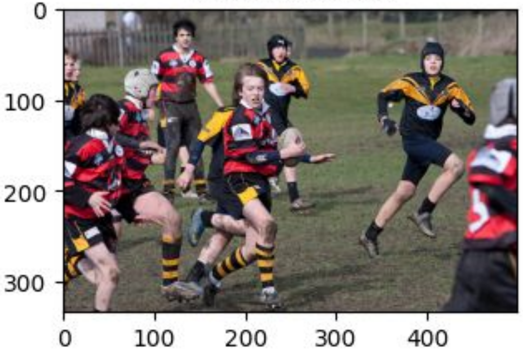
Predicted: A small brown dog green grass
Actual: A medium size brown white streaked dog run tall grass
BLEU Score: 83.3



Predicted: a dog jump wooden fence
Actual: Two brown black dog jump fence
BLEU Score: 80

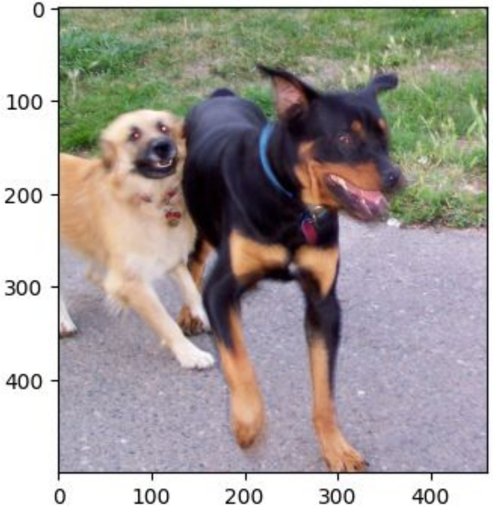


Predicted: group kid play game
Actual: Young person uniform play rugby game
BLEU Score: 75



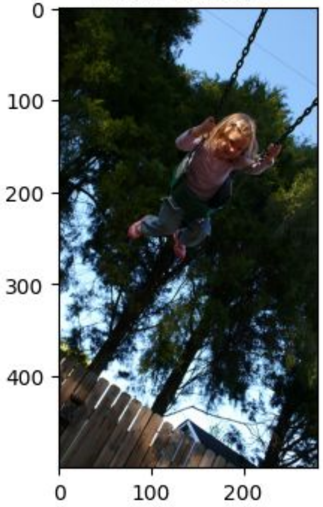
Results on Image Captioning – Bad

Predicted: A dog
Actual: A black brown dog tan dog cement path front grass
BLEU Score: 11



Predicted: A dog
Actual: A black dog run snow
BLEU Score: 11

Predicted: A child jump
Actual: Girl wear pink swing swing
BLEU Score: 0



Methodology – Adversarial Attack

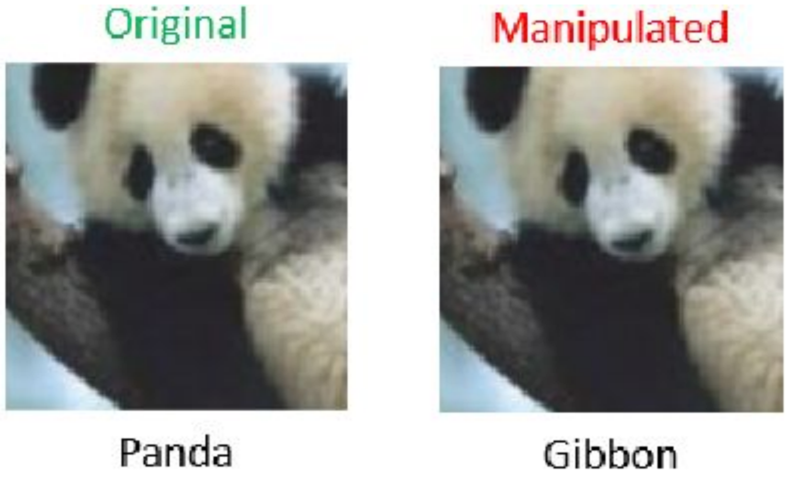
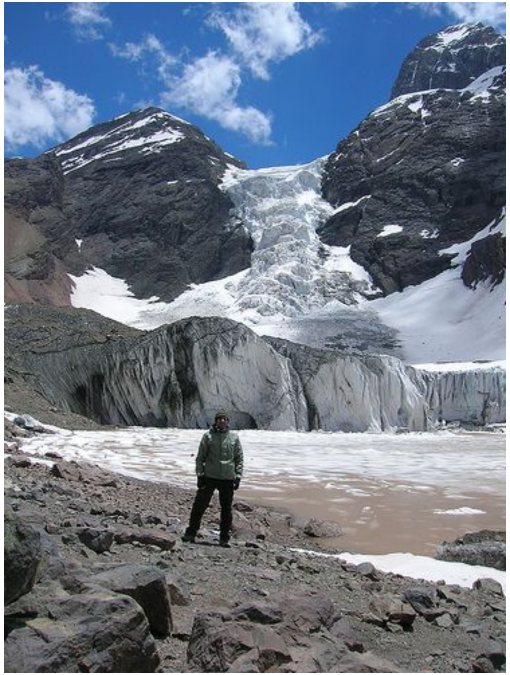


Figure: An Illustration of Adversarial Attack

1) Generate image/features that appear similar to the human eye but deceive the network

2) Use backpropagation to obtain adversarial features by minimizing the loss for incorrect captions

Results for Adversarial Attack



Original Prediction: “A man stand front rise rock formation background” => BLEU Score = 62.5

Incorrect Caption: “A man street racer armor examine tire another racer motorbike”
=> BLEU Score = 10

Caption generated by Adversarial Feature: “A small blue shirt hold large blue shirt hold large”
=> BLEU Score = 4.34

<u>Distance between original and adversarial feature</u>
=> 9.24 units

Lessons Learned

Wins

- Application of Attention mechanism on combination of images and texts
- Reverse engineering to generate adversarial features using gradient descent

Limitations

- Captures main objects but doesn’t understand the macro environment
- Cannot re-generate image from adversarial features

Future Work

- Train the model on a larger dataset such as MS-COCO or Flickr30 for better generalization to improve the quality of generated captions
- Enable training on pre-trained VGG-16 to further backpropagate error and generate images from adversarial features