AUTOMATIC IMAGE CAPTION GENERATION

ZAHIDUZZAMAN BISWAS

SPRINGBOARD DATA SCIENCE TRACK

PROBLEM DEFINITION

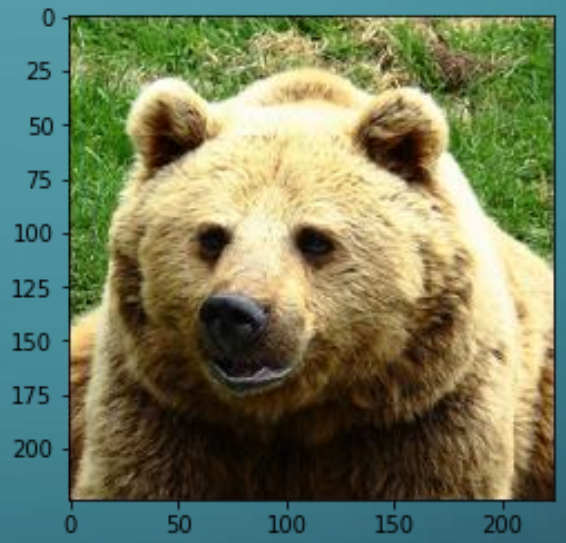
What is the problem?	• Generate descriptive captions of images
Who is interested in this problem?	• Google, Microsoft, Facebook etc.
Why should this problem be solved?	• Computer vision, Visual Impairment Aide, Social Media efficiency etc.
How will I solve this problem?	• Using deep learning and NLP tools

BACKGROUND INFORMATION

- Hard-coded sentences and visual concept of Farhadi and Kulkarni [3]
- VGG16, RESNET, and other ImageNet works on Image Classification
- Scene and object labeling of et al Li, et al Gould, and et al Fidler
- Compare Images and combine captions of Jia, Kuznetsova, and Li [5]
- Multi-model recurrent neural network model of Karphathy [2]

DATA SET

- MS Coco 2017 dataset
  - Collected from Flickr
    - 5000 images
    - 20,000 captions
    - 80 different categories
    - 12 different sub-categories



A big burly grizzly bear is show with grass in the background.

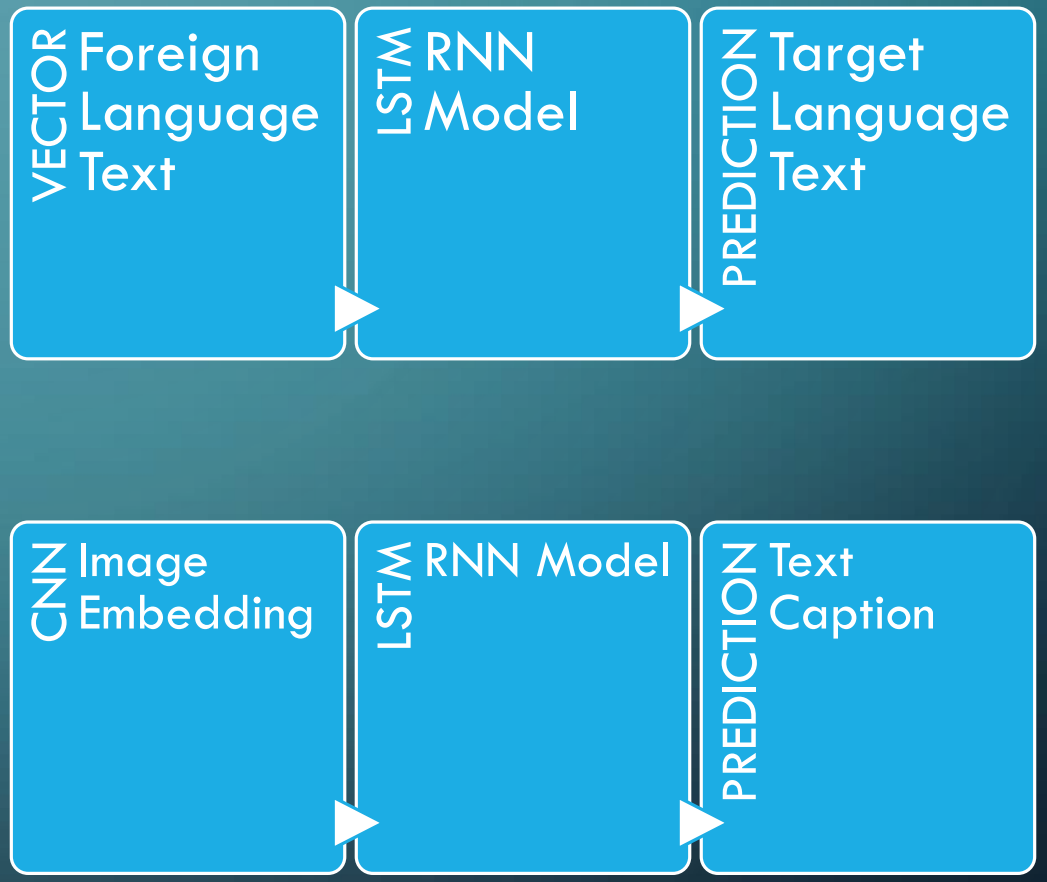
The large brown bear has a black nose.

Closeup of a brown bear sitting in a grassy area.

A large bear that is sitting on grass.
A close up picture of a brown bear's face

THE MAIN IDEA

- Text translation approach from one language to another language
- Input Image Embedding instead of Vectorized Text



OUR APPROACH

Text Processing

- Only the first caption of each image
- Remove the non-alphanumeric characters and change to lower case
- Indexed the words from all the caption and vectorize each caption accordingly

Image Embedding

- Scaled each image to 224 x 224
- Embedded images using VGG16 and ResNet50 pretrained CNN models
- Used the dense layer output for training

Training and Testing

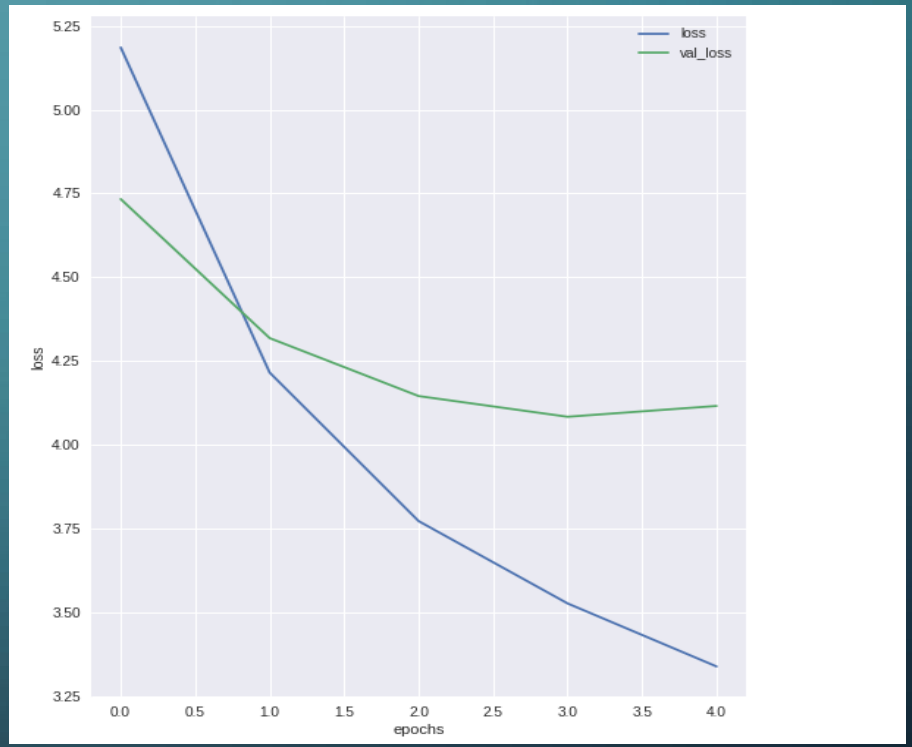
- Split data: 3000 training, 1000 validation, 1000 testing
- Input the image embedding and text vector to a LSTM network model
- Validated using cross entropy and evaluated with BLEU metrics

FINAL RESULT

CNN Model	CNN Output	LSTM Layers	First Dropout	Second Droput	Embedding Space	BLEU Test	BLEU Train
VGG16	fc1 layer	1 w/(256)	0.6	0.3	256	0.188	0.189
VGG16	fc1 layer	1 w/(256)	0.5	0.2	128	0.235	0.233
VGG16	fc1 layer	2 w/(256)	0.5	0.5	128	0.252	.0.238
RESNET50	fc100 layer	1 w/(256)	0.2	0.5	128	0.402	0.380
RESNET50	fc100 layer	1 w/(256)	0.2	0.5	256	0.416	0.392

OVER-FITTING AND ERRORS

- ‘Curse of Dimensionality’
  - 3000 instances
  - 4096 features
- Smoothing function missing
- Cross-validation missing



FUTURE EXPERIMENTS

- Use bigger dataset e.g. flikr 8k, flikr 30k, COCO 2014
- Cross validation on the train-test-splits
- Try different evaluation metrics
- Principal Component Analysis (dimensionality Reduction)

WORK CITED

[1] “Common Object in Context” COCO. www.cocodataset.org (2017). Web. 5 Apr. 2019

[2] Karpathy, Andrej, and Li Fei-Fei. ”Deep Visual-semantic Alignments for Generating Image Descriptions.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015). Web. 5 April 2016

[3] Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. ”Every Picture Tells a Story: Generating Sentences from Images.” Computer Vision ECCV 2010 Lecture Notes in Computer Science (2010): 15-29. Web. 5 April 2019

[4] Gould, Stephen, Richard Fulton, and Daphne Koller. ”Decomposing a Scene into Geometric and Semantically Consistent Regions.” 2009 IEEE 12th International Conference on Computer Vision (2009). Web. 5 April 2019

[5] Fidler, Sanja, Abhishek Sharma, and Raquel Urtasun. ”A Sentence Is Worth a Thousand Pixels.” 2013 IEEE Conference on Computer Vision and Pattern Recognition (2013). Web. 5 April 2019

[6] Li, Li-Jia, and Li Fei-Fei. ”What, Where and Who? Classifying Events by Scene and Object Recognition.” 2007 IEEE 11th International Conference on Computer Vision (2007). Web. 5 Apr. 2019