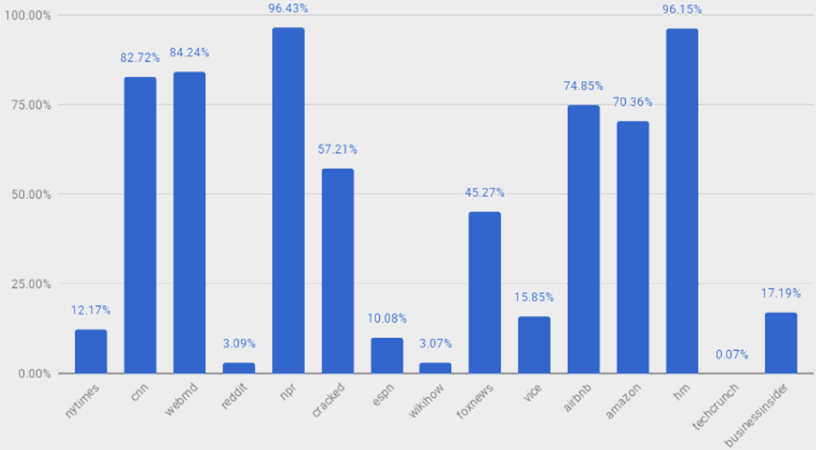
Automated Image Captioning System

B.TECH. IT CAPSTONE PROJECT

Saurabh Mathur


Percentage of Images with Populated Alt Tags per Website

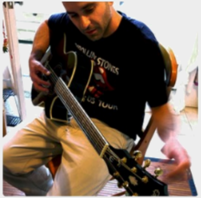
PROBLEM STATEMENT

The web is full of multimedia content.

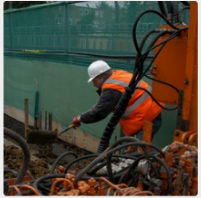
Alternative text is missing from many images.

Such content is inaccessible to screen readers.

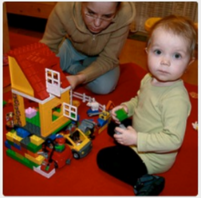
Manually captioning images is expensive.



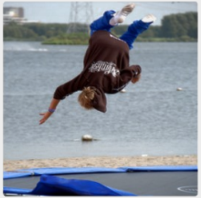
"man in black shirt is playing guitar."



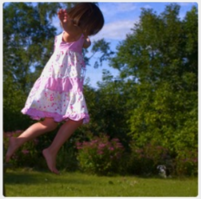
"construction worker in orange safety vest is working on road."



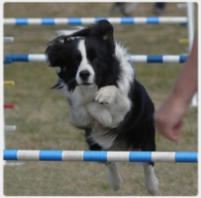
"two young girls are playing with lego toy."



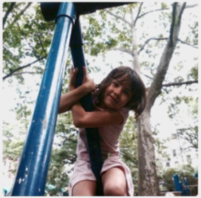
"boy is doing backflip on wakeboard."



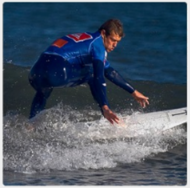
"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."

RELATED WORK

Ranking descriptions for given image.

Co-Embedding image and descriptions in same vector space.

Embedding image crops with annotations.

End to End generation of image descriptions.

KEY IDEA

Resize image and Rescale pixel values.

Extract image features.

Generate word-by-word occurance probabilities.

Decode caption.

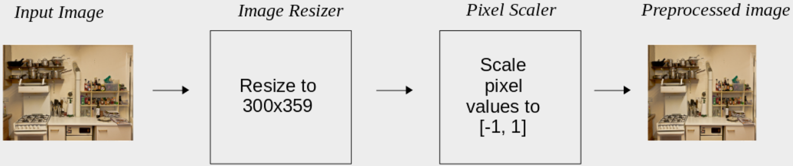
KEY IDEA

Resize image and Rescale pixel values.

Extract image features.

Generate word-by-word occurance probabilities.

Decode caption.



Input Image

Image Resizer

Resize to 300x359

Pixel Scaler

Scale pixel values to [-1, 1]

Preprocessed image

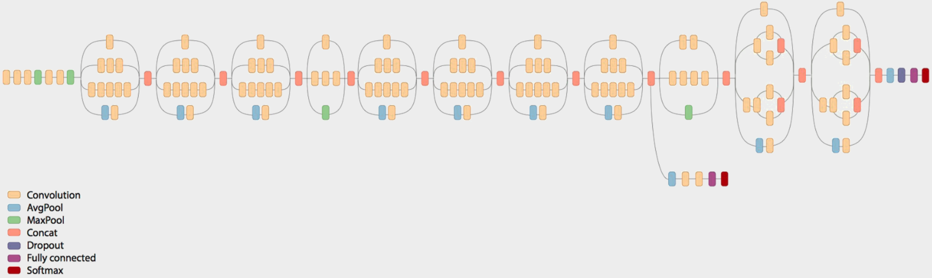
KEY IDEA

Resize image and Rescale pixel values.

Extract image features.

Generate word-by-word occurance probabilities.

Decode caption.


Convolution
AvgPool
MaxPool
Concat
Dropout
Fully connected
Softmax

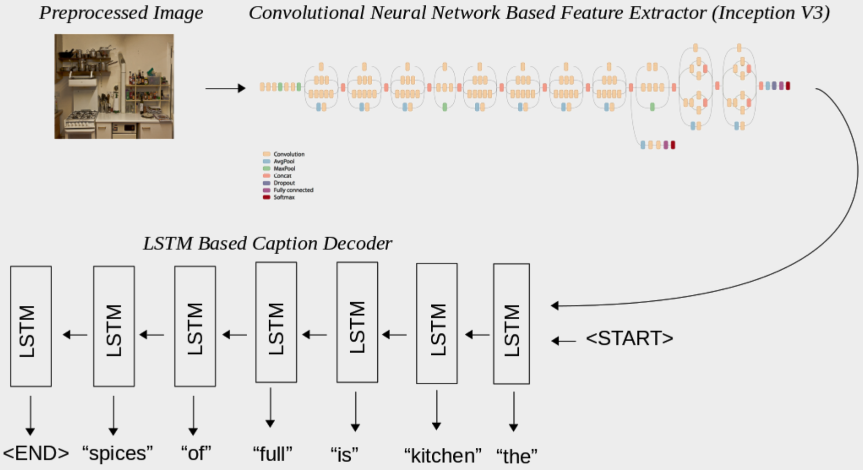
KEY IDEA

Resize image and Rescale pixel values.

Extract image features.

Generate word-by-word occurance probabilities.

Decode caption.



Preprocessed Image

Convolutional Neural Network Based Feature Extractor (Inception V3)

LSTM Based Caption Decoder

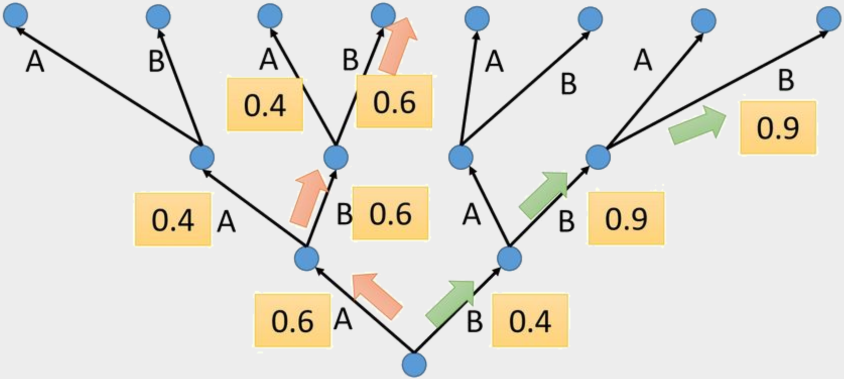
KEY IDEA

Resize image and Rescale pixel values.

Extract image features.

Generate word-by-word occurance probabilities.

Decode caption.


A
B
A
B
0.4
0.6
A
B
A
B
0.9
0.4
A
B
0.6
0.9
A
B
0.6
A
B
0.4

IMPLEMENTATION PHASES

Data Acquisition.

Data Preprocessing.

Training.

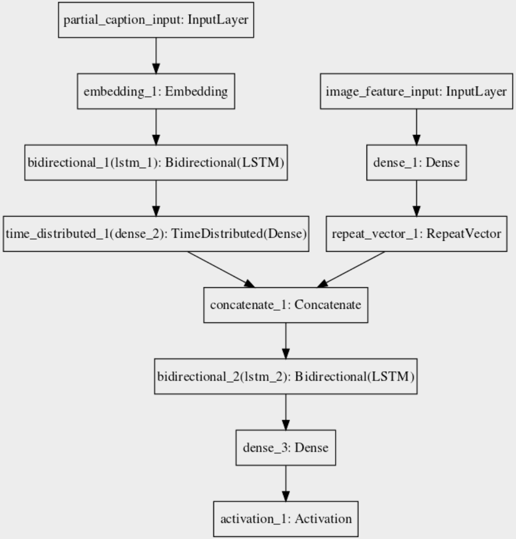
Validation.

MS COCO DATASET

Common Objects in Context.

120,000 images.

Each image has 5 captions.

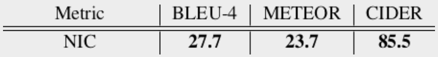

partial_caption_input: InputLayer
embedding_1: Embedding
image_feature_input: InputLayer
bidirectional_1(lstm_1): Bidirectional(LSTM)
dense_1: Dense
time_distributed_1(dense_2): TimeDistributed(Dense)
repeat_vector_1: RepeatVector
concatenate_1: Concatenate
bidirectional_2(lstm_2): Bidirectional(LSTM)
dense_3: Dense
activation_1: Activation

TRAINING

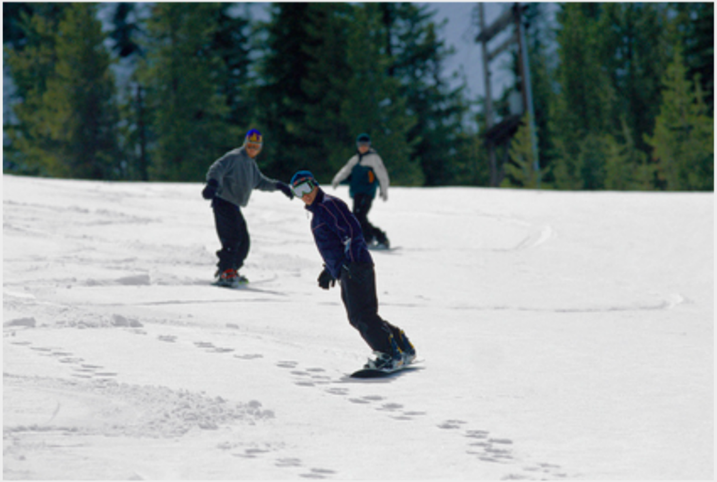
Objective: Maximize the probability of the correct description given the image.

$$\theta^* = argmax_\theta \sum_{(S,I)} logp(S|I;\theta)$$

RESULTS

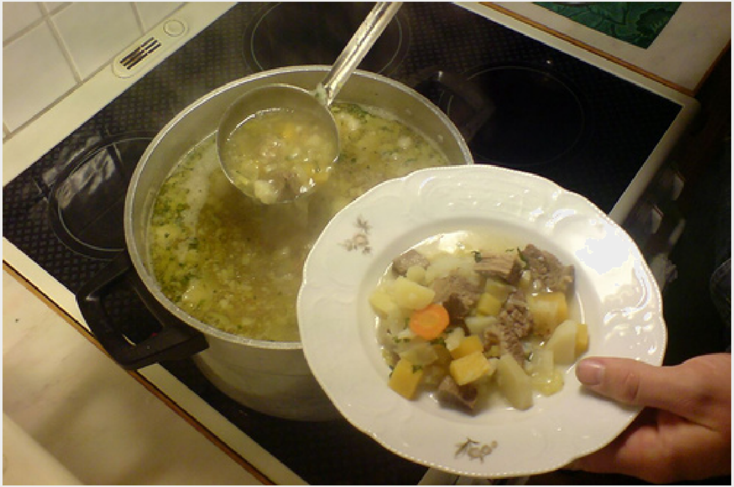


Metric	BLEU-4	METEOR	CIDER
NIC	**27.7**	**23.7**	**85.5**



Greedy: a man is skiing down a snowy slope

Beam: a group of people skiing down a snow covered slope



Greedy: a plate of food with a banana and a spoon
Beam: a close up of a bunch of food on a table