iMage Captioning

By **Rishabh Jha**
S20190010150

wHat is Image Captioning?

- It refers to the task of generating textual descriptions of a given Image such that It captures the objects and the actions happening in the Image.

- It is an Interdisciplinary Research problem that lies between the Area of Computer Vision and Natural Language Processing

wHy Image Captioning?

- Image related search
- A Boon to Visually Impaired People.
- Social Media

The DATAset

FLICKr 8K

A person is walking along a beach with a big dog
A black and white dog carries a tennis ball in its mouth
A soccer player takes a soccer ball in the grass
A man is doing a trick on a snowboard
A surfer dives into the ocean
A black and white dog leaps to catch a Frisbee

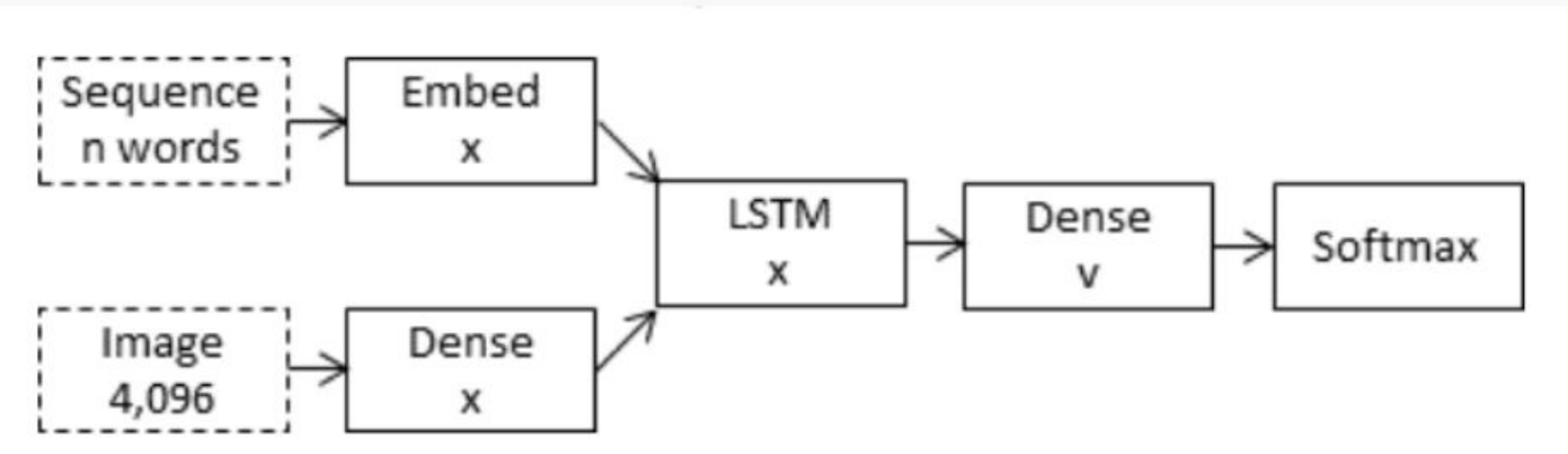
- Total 8k Images
- Good small dataset to carry out the Experiment

6k train Images
1k validation Images
1k test Images

- Images will contain 5 different captions describing them

The archiTECTURE

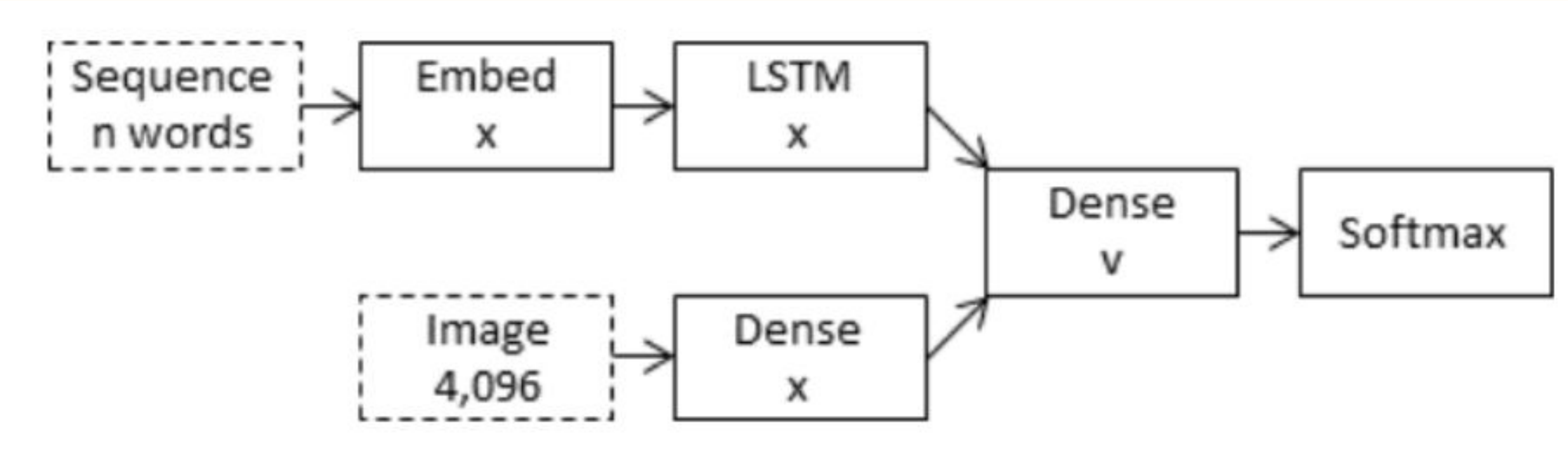
Inject Architecture



Inject Architecture

- Conditioning-by-inject
- In this Architecture/model , the RNN is trained to predict the sequences based on histories containing both linguistic and perceptual features
- In this model, the RNN is primarily responsible for image-conditioned language generation, hence conditioning-by-inject
- This model learns a "**visuo-linguistic**" representation of each word.
- This model can learns how to disambiguate the tokens of same word having different meaning using visual representation, such as crane, can be a bird or construction equipment.
- All these points implies that inject models learns larger vocabularies.

Merge Architecture



Merge Architecture

- Conditioning-by-merge
- In the Merge Architecture/model , the RNN is functioning as an encoder of sequences of word embeddings.
- This model learns only "**linguistic**" representation by RNN and combining it with the visual representation in a later layer.
- Visual features are extracted from image using a CNN model (VGG/InceptionV3) and then merged with linguistic feature vector in a multimodal layer, which in turn is responsible for generating the caption.
- This model generally fails to disambiguate the tokens of same word having different meaning using visual representation, such as Bank, can be a River Bank or a Financial Institution.

reSULts

		% Vocabulary		CIDEr		METEOR		ROUGE-L	
Layer	Vocab.	Merge	Inject	Merge	Inject	Merge	Inject	Merge	Inject
128	2539	**14.730 (0.40)**	10.555 (0.34)	**0.460 (0.01)**	0.431 (0.01)	**0.192 (0.00)**	0.183 (0.00)	**0.445 (0.00)**	0.430 (0.00)
128	2918	**13.719 (0.49)**	8.876 (0.24)	**0.456 (0.00)**	0.431 (0.00)	**0.191 (0.00)**	0.185 (0.00)	**0.437 (0.00)**	0.434 (0.00)
128	3478	**11.223 (0.35)**	8.175 (0.31)	**0.458 (0.01)**	0.433 (0.01)	**0.192 (0.00)**	0.187 (0.00)	**0.442 (0.00)**	0.432 (0.00)
256	2539	**15.439 (0.84)**	11.448 (0.71)	**0.462 (0.01)**	0.456 (0.01)	**0.192 (0.00)**	0.189 (0.00)	**0.439 (0.00)**	0.436 (0.00)
256	2918	**13.697 (0.19)**	10.430 (0.34)	**0.456 (0.01)**	0.451 (0.01)	**0.190 (0.00)**	0.189 (0.00)	0.438 (0.00)	**0.440 (0.00)**
256	3478	**11.252 (0.51)**	8.405 (0.39)	**0.470 (0.01)**	0.449 (0.02)	**0.191 (0.00)**	0.189 (0.00)	**0.439 (0.00)**	0.437 (0.00)
512	2539	**15.741 (0.40)**	12.761 (0.81)	0.452 (0.01)	**0.464 (0.00)**	0.191 (0.00)	**0.192 (0.00)**	0.437 (0.00)	**0.442 (0.00)**
512	2918	**13.114 (0.75)**	10.155 (0.42)	**0.469 (0.01)**	0.457 (0.00)	**0.193 (0.00)**	0.189 (0.00)	**0.440 (0.00)**	0.437 (0.00)
512	3478	**11.501 (0.49)**	8.587 (0.50)	**0.458 (0.01)**	0.439 (0.01)	**0.192 (0.00)**	0.188 (0.00)	**0.439 (0.00)**	0.434 (0.00)

(a) Flickr8k: % of vocabulary used, CIDEr, METEOR and ROUGE-L results.

		BLEU-1		BLEU-2		BLEU-3		BLEU-4	
Layer	Vocab.	Merge	Inject	Merge	Inject	Merge	Inject	Merge	Inject
128	2539	**0.600 (0.00)**	0.592 (0.01)	**0.410 (0.00)**	0.405 (0.01)	**0.272 (0.00)**	0.270 (0.01)	**0.179 (0.00)**	0.177 (0.00)
128	2918	**0.595 (0.01)**	0.590 (0.00)	0.405 (0.01)	**0.406 (0.00)**	0.267 (0.01)	**0.271 (0.00)**	0.175 (0.00)	**0.178 (0.00)**
128	3478	**0.608 (0.01)**	0.586 (0.01)	**0.416 (0.01)**	0.401 (0.01)	**0.276 (0.01)**	0.268 (0.01)	**0.182 (0.01)**	0.178 (0.01)
256	2539	**0.594 (0.00)**	0.591 (0.00)	0.407 (0.01)	**0.408 (0.00)**	0.269 (0.01)	**0.276 (0.00)**	0.176 (0.01)	**0.184 (0.00)**
256	2918	**0.596 (0.01)**	0.596 (0.01)	0.405 (0.01)	**0.413 (0.01)**	0.265 (0.00)	**0.278 (0.01)**	0.172 (0.00)	**0.184 (0.00)**
256	3478	**0.601 (0.00)**	0.596 (0.01)	**0.411 (0.00)**	0.409 (0.01)	0.272 (0.01)	**0.274 (0.01)**	0.179 (0.01)	**0.181 (0.01)**
512	2539	0.597 (0.01)	**0.603 (0.00)**	0.406 (0.01)	**0.419 (0.00)**	0.267 (0.01)	**0.283 (0.00)**	0.176 (0.01)	**0.188 (0.00)**
512	2918	**0.593 (0.01)**	0.589 (0.01)	0.404 (0.01)	**0.409 (0.00)**	0.268 (0.00)	**0.277 (0.00)**	0.177 (0.00)	**0.185 (0.00)**
512	3478	**0.597 (0.01)**	0.587 (0.00)	**0.407 (0.01)**	0.405 (0.00)	0.270 (0.01)	**0.272 (0.00)**	0.178 (0.00)	**0.180 (0.01)**

(b) Flickr8k: BLEU-n scores.

cOnclusionS

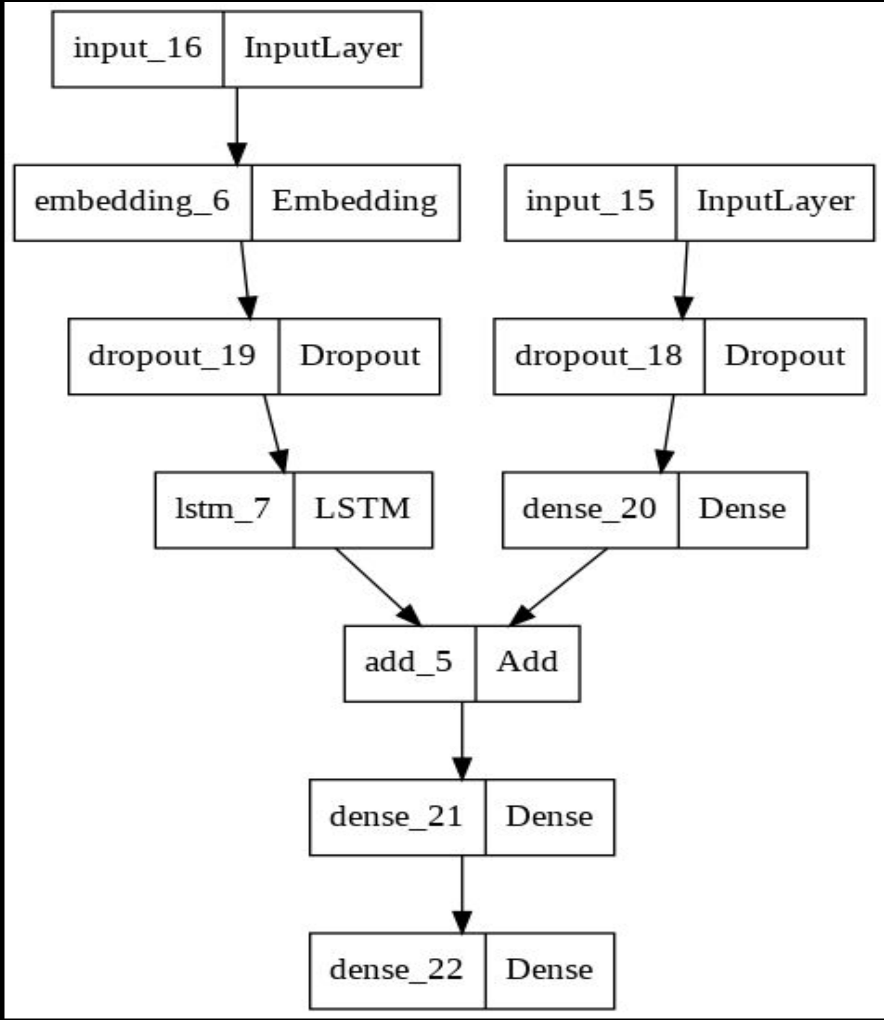
- With lesser number of layer , merge models generate better quality captions.
- As we increase the number of layers, inject models tend to perform better with captions.
- Merge Architecture models uses more of training vocab on test captions.

methoDoLogy

PROcedure

- Storing descriptions into a dictionary from a text file
- Cleaning the descriptions
- Creating the Vocabulary
- Extracting feature vector (2048) of every image using InceptionV3 model
- Storing features into a pickle file
- Converting training captions into sequence of vector (number)
- Padding every captions with zero
- Word Embeddings
- Training the Model
- Generating the Captions using Greedy Search and Beam Search

eXACT mOdel

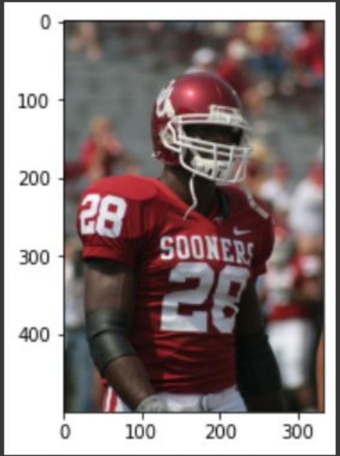


Download my model

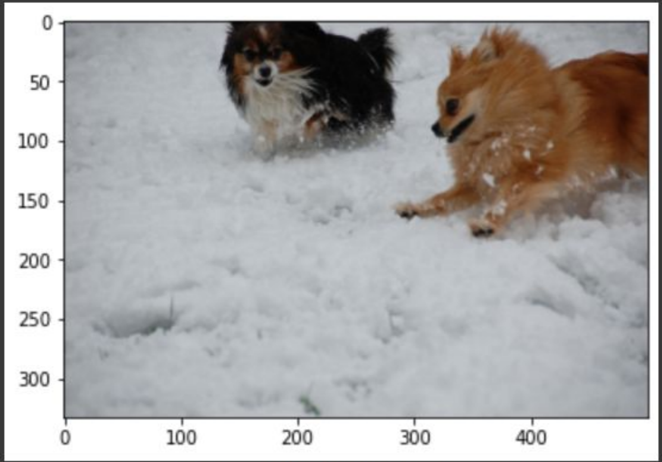
ConclusioN

- Merge model used
- 23% accuracy over 30 epochs
- Beam Search generates better captions than greedy search in general.
- Brute Hyper-parameters analysis can also help in increasing the accuracy of the model
- Accuracy can be improved with more number of epochs or better architecture having attention etc.

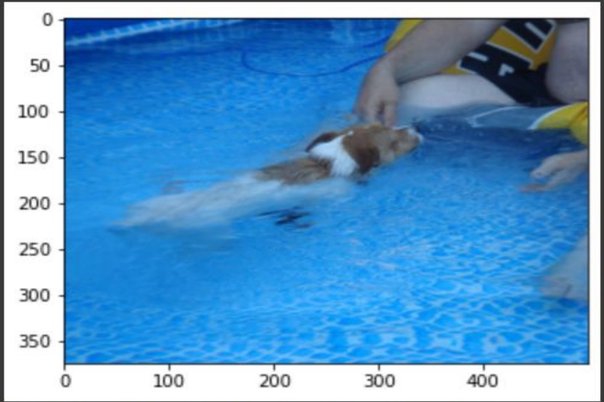
Greedy: man in black on the street of front of front of the background
Beam: group of people sit on front of bus



Greedy: football player in sooners
Beam: two football players in sooners



Greedy: two white and white dog is running in the snow
Beam: white and white dog is running in the snow



Greedy: two girl is pink in the water
Beam: little girl white in the water

fuTure wORk

- Improving the results of the current model
- Trying out more complex architectures such as model with visual attention or Transformers
- Experiment with the larger dataset FLICKR 8k or MS COCO
- Deploy the model online.

REFerences

- Tanti, M., Gatt, A. and Camilleri, K.P., 2017. What is the role of recurrent neural networks (rnns) in an image caption generator?. *arXiv preprint arXiv:1708.02043.*

- Tensorflow Image Captioning Article

- Image Captioning with keras

THANK yoU