Image Caption Generation

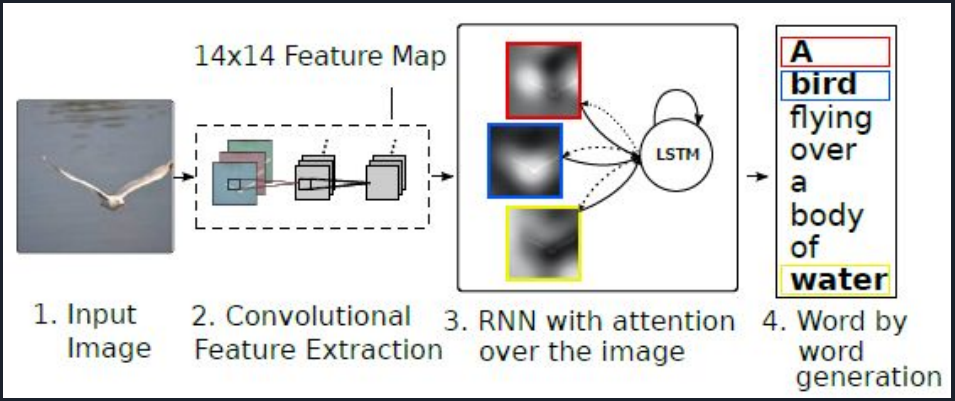
Machine Learning and Neural Networks

Julen Etxaniz and Oihane Cantero

17-12-2020

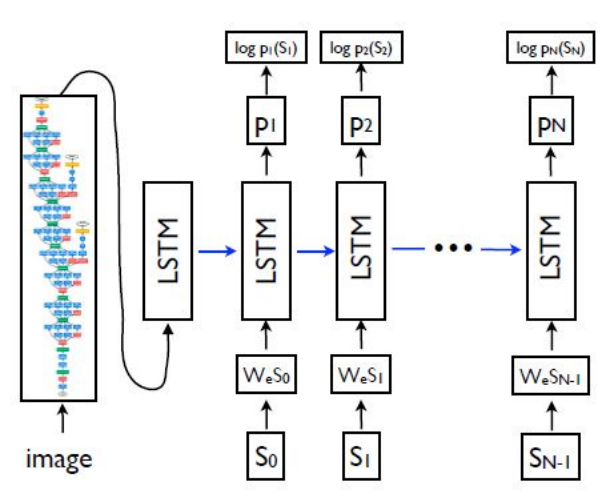
Contents

1. Introduction



2 objectives:

- Implement a caption generation model using a CNN to condition a LSTM language model
- Add attention mechanism to the model



2.1. Get Dataset

Caption1: closeup of white dog that is laying its head on its paws
Caption 2: large white dog lying on the floor
Caption 3: white dog has its head on the ground
Caption 4: white dog is resting its head on tiled floor with its eyes open
Caption 5: white dog rests its head on the patio bricks

Flickr8k Dataset:

- 8000 images
- 5 captions for each image

Larger datasets:

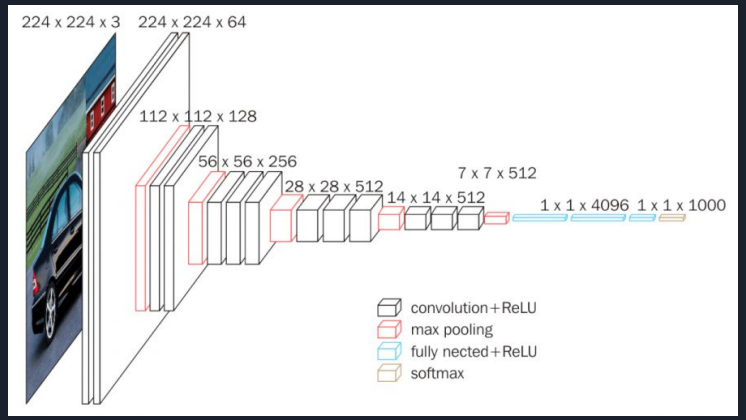
- Flickr30k
- MSCOCO

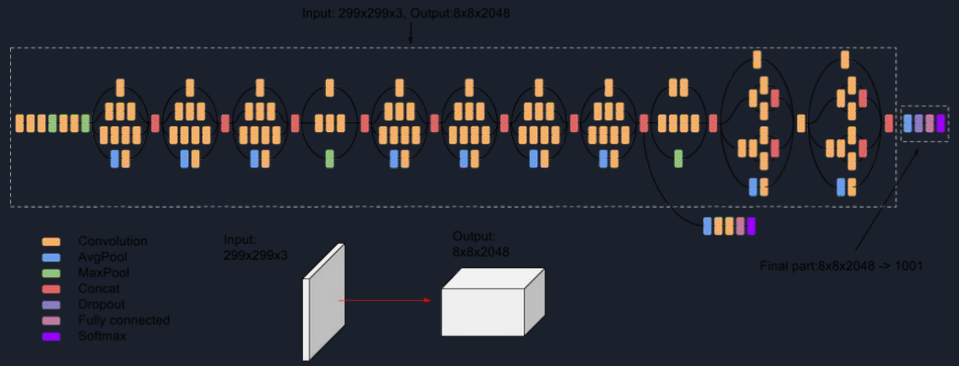
Caption 1: little tan dog with large ears running through the grass
Caption 2: playful dog is running through the grass
Caption 3: small dogs ears stick up as it runs in the grass
Caption 4: the small dog is running across the lawn
Caption 5: this is small beige dog running through grassy field

2.2. Prepare Photo Data

CNN models to extract features:

- Remove last layers used for predicting
- Preprocess image
- Calculate and save features
- VGG16
  - 134,260,544 parameters
  - Input: 224 x 224 image
  - Output: 4096 feature vector
- InceptionV3
  - 21,768,352 parameters
  - Input: 299 x 299 image
  - Output: 2048 feature vector





2.3. Prepare Text Data

- Load all descriptions
  - Create a dictionary mapping descriptions to images
- Clean descriptions:
  - Convert all the words to lowercase
  - Remove all the punctuation
  - Remove words one character long
  - Remove words with numbers
- Create a vocabulary
  - 8,763 unique words
- Save descriptions to a file

```
1000268201_693b08cb0e.jpg#0     A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1     A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2     A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3     A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4     A little girl in a pink dress going into a wooden cabin .
```

```
1000268201_693b08cb0e child in pink dress is climbing up set of stairs in an entry way
1000268201_693b08cb0e girl going into wooden building
1000268201_693b08cb0e little girl climbing into wooden playhouse
1000268201_693b08cb0e little girl climbing the stairs to her playhouse
1000268201_693b08cb0e little girl in pink dress going into wooden cabin
```

2.4. Load Data

- Predefined subsets:
  - Train 6000
  - Validation 1000
  - Test 1000
- Captions
  - Read from saved files
  - Add startseq and endseq
- Image features
  - Read from saved file

Flickr_8k.devImages.txt
Flickr_8k.testImages.txt
Flickr_8k.trainImages.txt

```
2513260012_03d33305cf.jpg
2903617548_d3e38d7f88.jpg
3338291921_fe7ae0c8f8.jpg
488416045_1c6d903fe0.jpg
2644326817_8f45080b87.jpg
218342358_1755a9cce1.jpg
```

features.pkl
features_inceptionv3.pkl

2.5. Encode Text Data

- Convert descriptions to lists of words
- Use a tokenizer
- Map from words of the vocabulary to integers

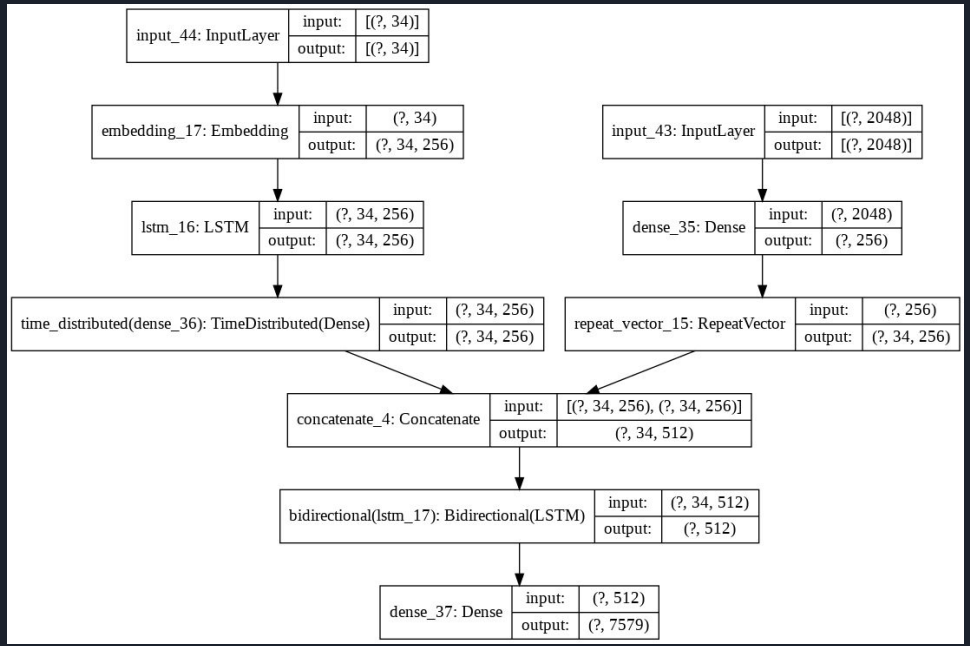
```
[('startseq', 1),
 ('endseq', 2),
 ('in', 3),
 ('the', 4),
 ('on', 5),
 ('is', 6),
 ('and', 7),
 ('dog', 8),
 ('with', 9),
 ('man', 10)]
```

- Calculate vocabulary size
- Compute maximum caption length

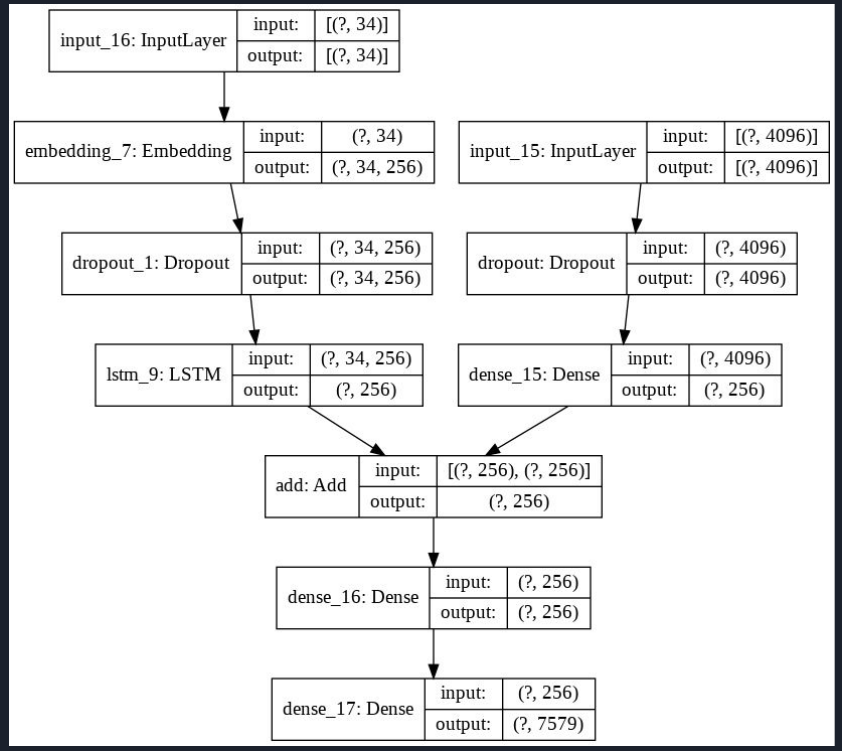
```
Vocabulary Size: 7579
Description Length: 34
```

2.6. Define Model

Model 2

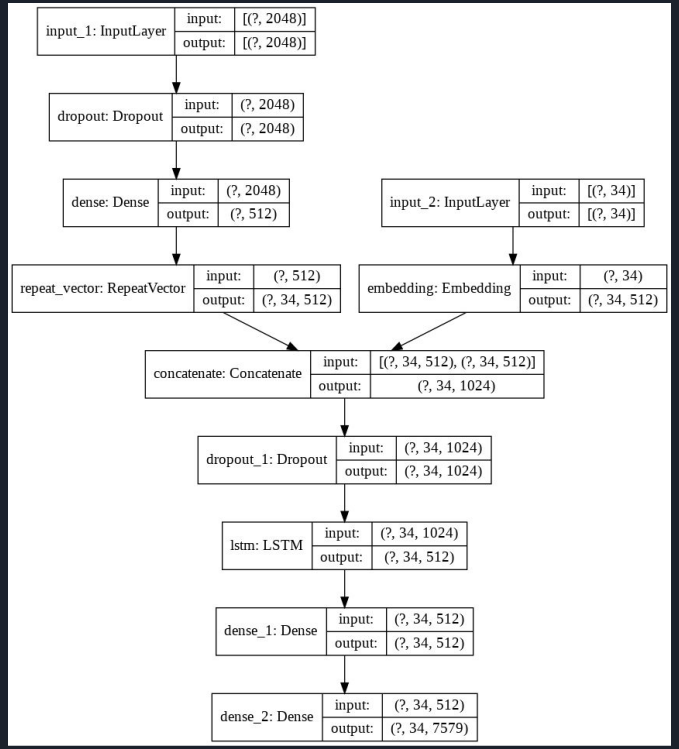


Model 1

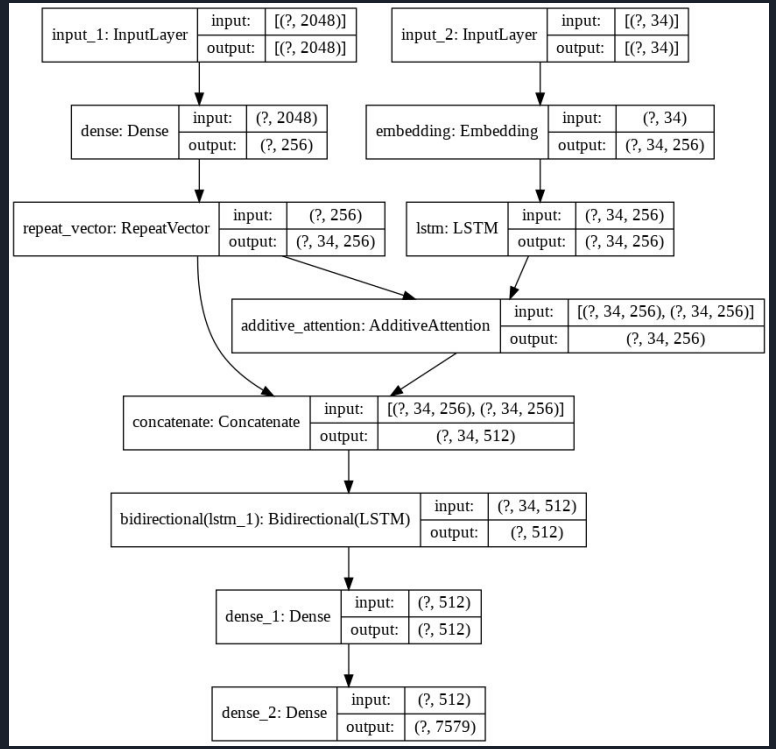


2.6. Define Model

Model 3

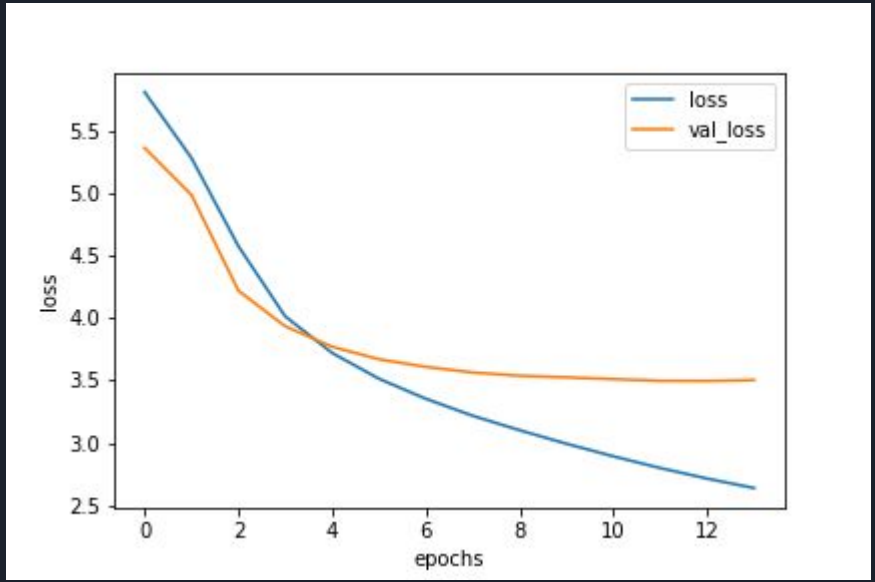


Model 2 Attention



2.7. Fit Model

- 20 epochs max (early stop)
- 32 batch size
- Save only lowest loss model



- Create sequences X1, X2, y
- Data generator
- Avoid RAM and GPU limits
- Shuffle train data
- Work with image ids

```
X1,     X2 (text sequence),                          y (word)
photo   startseq,                                    little
photo   startseq, little,                            girl
photo   startseq, little, girl,                      running
photo   startseq, little, girl, running,             in
photo   startseq, little, girl, running, in,         field
photo   startseq, little, girl, running, in, field,  endseq
```

2.8. Evaluate Model

Sampling:

- Take the best word at each time step

Beam Search

- Considers the k best sentences at each time step
- Increases the chance of getting a better description

Sampling BLEU scores:

- BLEU-1: 0.595796
- BLEU-2: 0.369997
- BLEU-3: 0.272431
- BLEU-4: 0.144684

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text [24]				11
TreeTalk [18]				19
BabyTalk [16]	25			
Tri5Sem [11]			48	
m-RNN [21]		55	58	
MNLM [14][5]		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

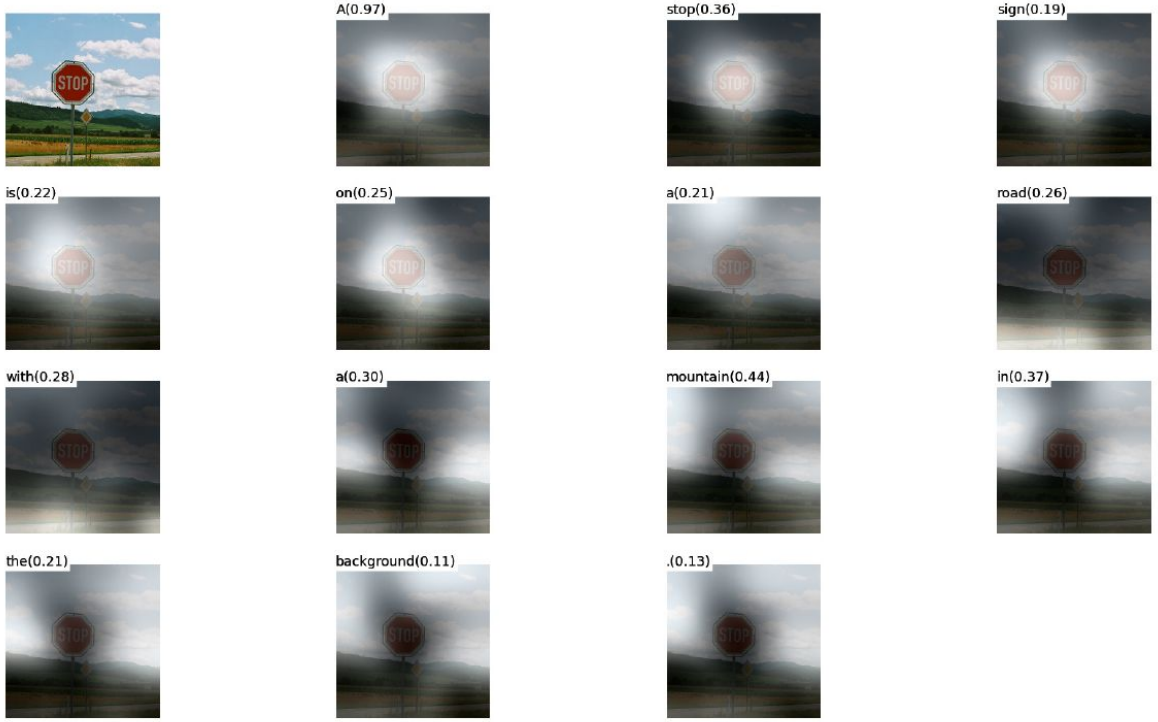
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7
	Soft-Attention	**67**	44.8	29.9	19.5
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**

2.9. Generate Captions

Original 1: closeup of white dog that is laying its head on its paws
Original 2: large white dog lying on the floor
Original 3: white dog has its head on the ground
Original 4: white dog is resting its head on tiled floor with its eyes open
Original 5: white dog rests its head on the patio bricks
Sampling (BLEU-1: 0.500000): dog is jumping over log in the air
Beam Search k=3 (BLEU-1: 0.600000): the white dog is running through the snow
Beam Search k=5 (BLEU-1: 0.583333): the white dog is in the middle of the snow

Original 1: little tan dog with large ears running through the grass
Original 2: playful dog is running through the grass
Original 3: small dogs ears stick up as it runs in the grass
Original 4: the small dog is running across the lawn
Original 5: this is small beige dog running through grassy field
Sampling (BLEU-1: 0.644123): dog runs on the grass
Beam Search k=3 (BLEU-1: 0.900000): the brown dog is running through the grass
Beam Search k=5 (BLEU-1: 0.900000): the brown dog is running through the grass

2.9. Generate Captions



(b) A stop sign is on a road with a mountain in the background.

3. Conclusions

- Caption generation is challenging
- We obtained quite good results
- Still there are many mistakes

- Improvements:
  - Change model architecture
  - Remove some words from vocabulary
  - Increase the size of dataset (Flickr30k)
  - Implement attention correctly
  - Evaluate model with beam search