Neural Image Captioning

by Amund Vedal, Martin Hwasser, Wojciech Kryściński

References

Show and Tell: A Neural Image Caption Generator

Oriol Vinyals
Google
vinyals@google.com

Alexander Toshev
Google
toshev@google.com

Samy Bengio
Google
bengio@google.com

Dumitru Erhan
Google
dumitru@google.com

Dataset

Common Object in Context

- Common Object in Context
  - Training set - 2014 Contest Training images [83K images/13GB]
  - Validation set - 2014 Contest Val images [41K/6GB]
  - Test set - 2014 Contest Test images [41K/6GB]
- Each image in the training set had *at least* 5 reference captions

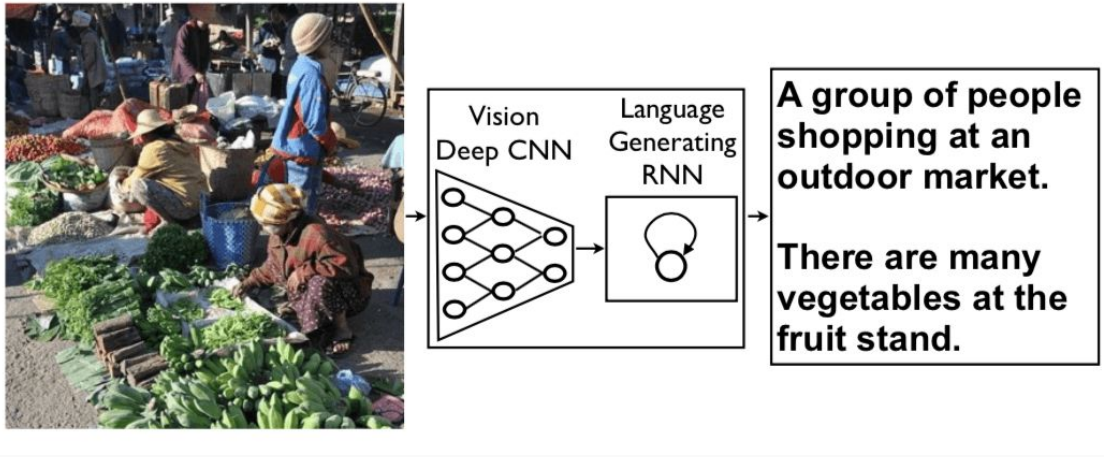
Dataset

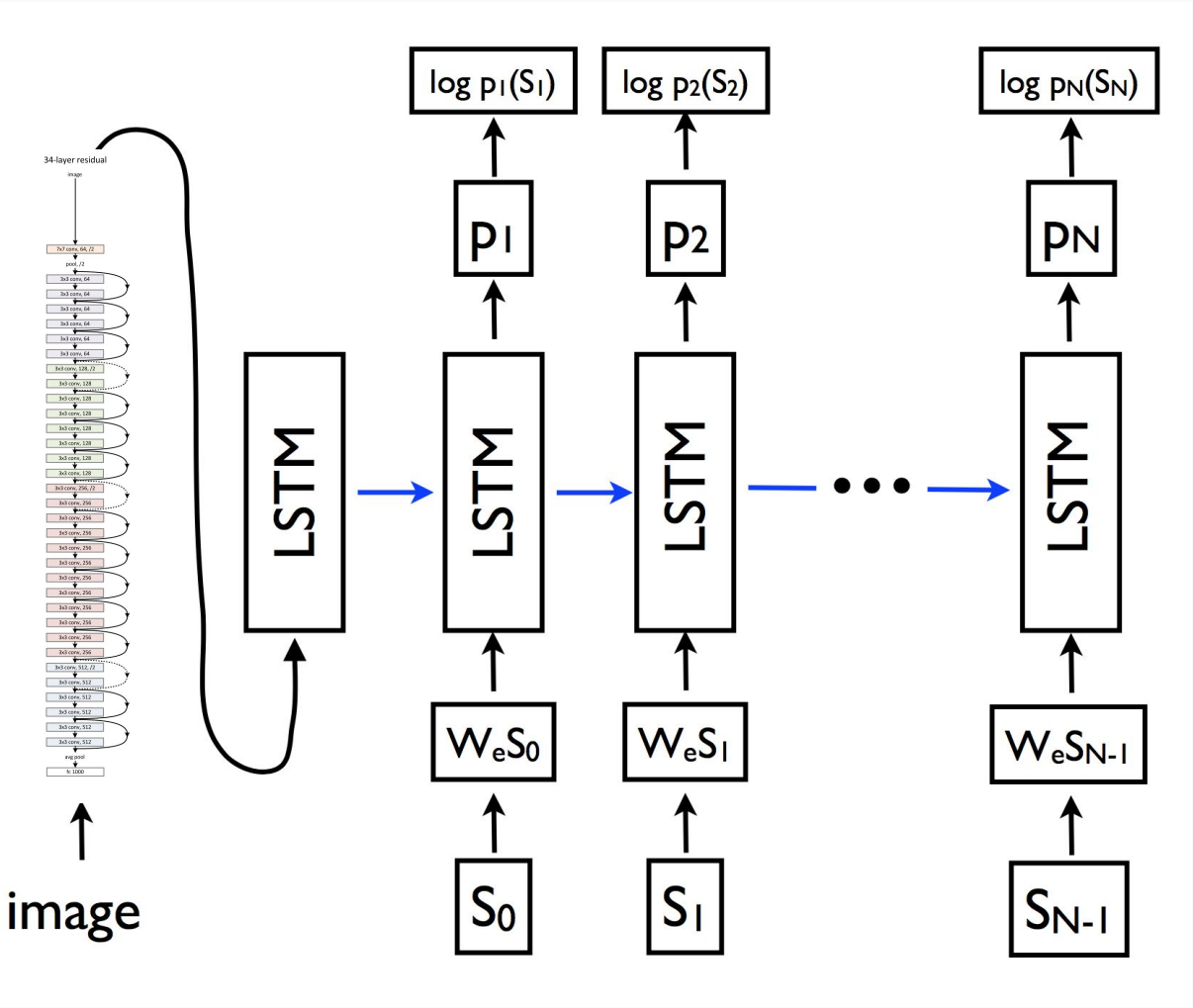
Example

the airplane takes off over the picnic tables in a park.
a large aircraft flying in the air above some benches in a field
an airplane flying low in the sky over picnic tables.
an airplane is flying low over a park with picnic tables.
a plane flying over a park, with the washington monument in the back.

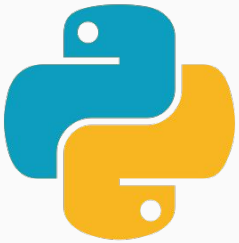
Architecture

Encoder-Decoder Network




34-layer residual
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
avg pool
fc 1000
image
LSTM
LSTM
LSTM
LSTM
log p1(S1)
log p2(S2)
log pN(SN)
p1
p2
pN
WeS0
WeS1
WeSN-1
S0
S1
SN-1

Implementation



Experimental Setup

Training parameters:

- Number of epochs: `3`
- Batch size: `128` (3236 batches per epoch)
- Vocabulary size: `15,000` most popular words
- Embedding size: `512` (image summary vector, word embeddings)
- RNN hidden state size: `512` and `1024`
- Learning rate: `1e-3`, with LR decay every 2000 batches

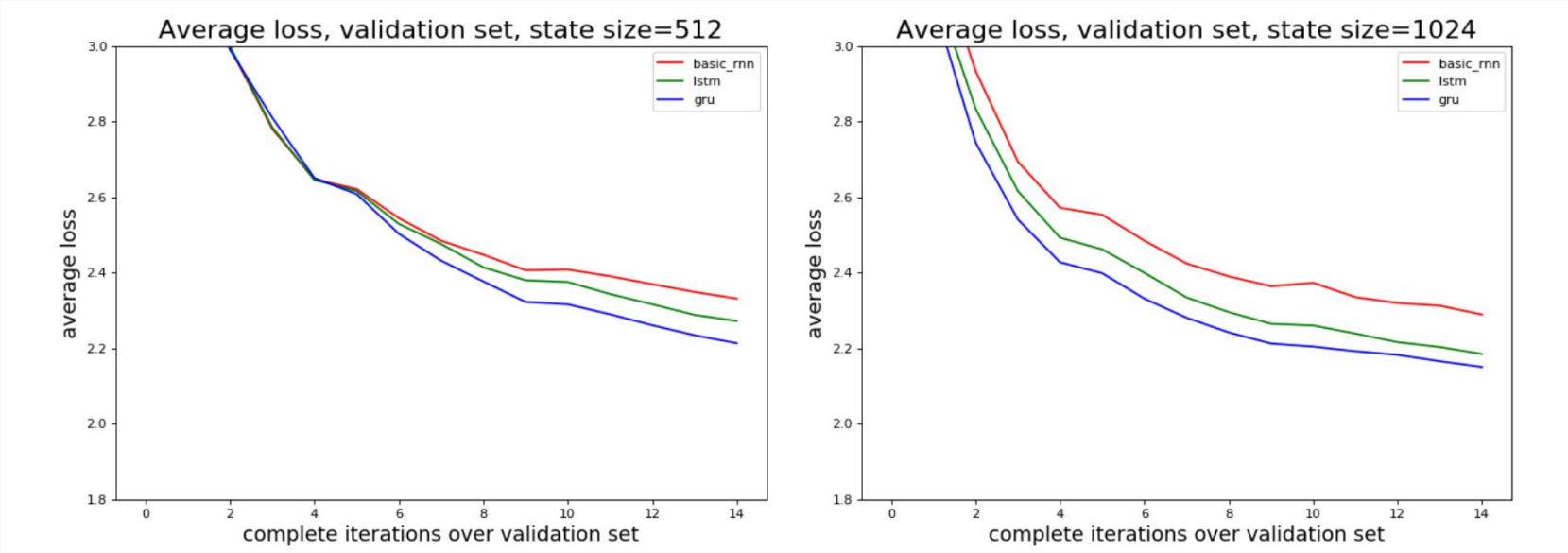
RNN units: Elman, LSTM, GRU

Evaluation
Methodology

- Quantitative evaluation
  - BLEU-1, BLEU-2, BLEU-3, BLEU-4 (precision)
  - ROGUE-L (recall)
  - METEOR (harmonic mean of precision and recall)
  - CIDEr
- Qualitative evaluation
  - Manually analyzing results

Training set loss

Validation set loss



Results

Validation set

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROGUE-L	CIDEr
Vinyals et al. (4k subset)	N/A	N/A	N/A	27.7	23.7	N/A	85.5
elman_512	62.5	43.2	29.1	19.8	19.5	45.6	57.7
elman_1024	61.9	42.9	28.8	19.6	19.9	45.9	58.7
gru_512	63.9	44.9	30.5	20.8	20.4	46.6	62.9
gru_1024	**64.0**	**45.3**	**31.2**	**21.5**	**21.1**	**47.1**	**66.1**
lstm_512	62.9	44.3	29.8	20.3	19.9	46.1	60.2
lstm_1024	63.4	45.0	31.0	21.4	20.8	**47.1**	64.4

Results

Test set

Model	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROGUE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Vinyals et al.	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
elman_1024	61.8	79.9	42.8	66.2	28.7	51.9	19.5	39.8	19.9	26.7	45.7	58.4	58.0	60.0
gru_1024	63.8	81.2	45.0	68.1	30.1	54.4	21.3	42.5	21.0	27.8	47.0	59.5	65.4	66.4
lstm_1024	63.3	81.0	44.8	67.9	30.7	54.0	21.1	42.0	20.7	27.4	46.9	59.2	63.7	64.8

Qualitative evaluation

4 categories:

1. No errors
2. Minor errors
3. Somewhat related
4. Not related

No errors

a group of giraffes standing in a field .

No errors

a man surfing the waves in the ocean

Minor errors

a vase filled with flowers sitting on a table .

Minor errors

a cat sitting on a chair in a living room .

Minor errors

a man flying through the air while riding skis .

a living room with a couch , chair , table and a television .

Somewhat related

a large white vase with a bunch of bananas on it .

Not related

Not related

a large white airplane flying in the sky .

Discussion

- GRU > LSTM > Elman
- Increasing size of hidden state (usually) improves performance
- Embeddings capture semantic similarities (ski / snowboard)
- Shows the power of transfer learning

Follow up work

- Fine-tuning Encoder (Convolutional Network)
- Using pre-trained Word Embeddings (word2vec/GloVe)
- More advanced decoding techniques (Beam Search)

Thank You

Questions?