LET AIS KNOW IMAGES

IMAGE CAPTION GENERATION

Libin Liang,Yao Song, Ruofan Yu
Department of Statistics
Rutgers University

INTRODUCTION

Objections in the figure:

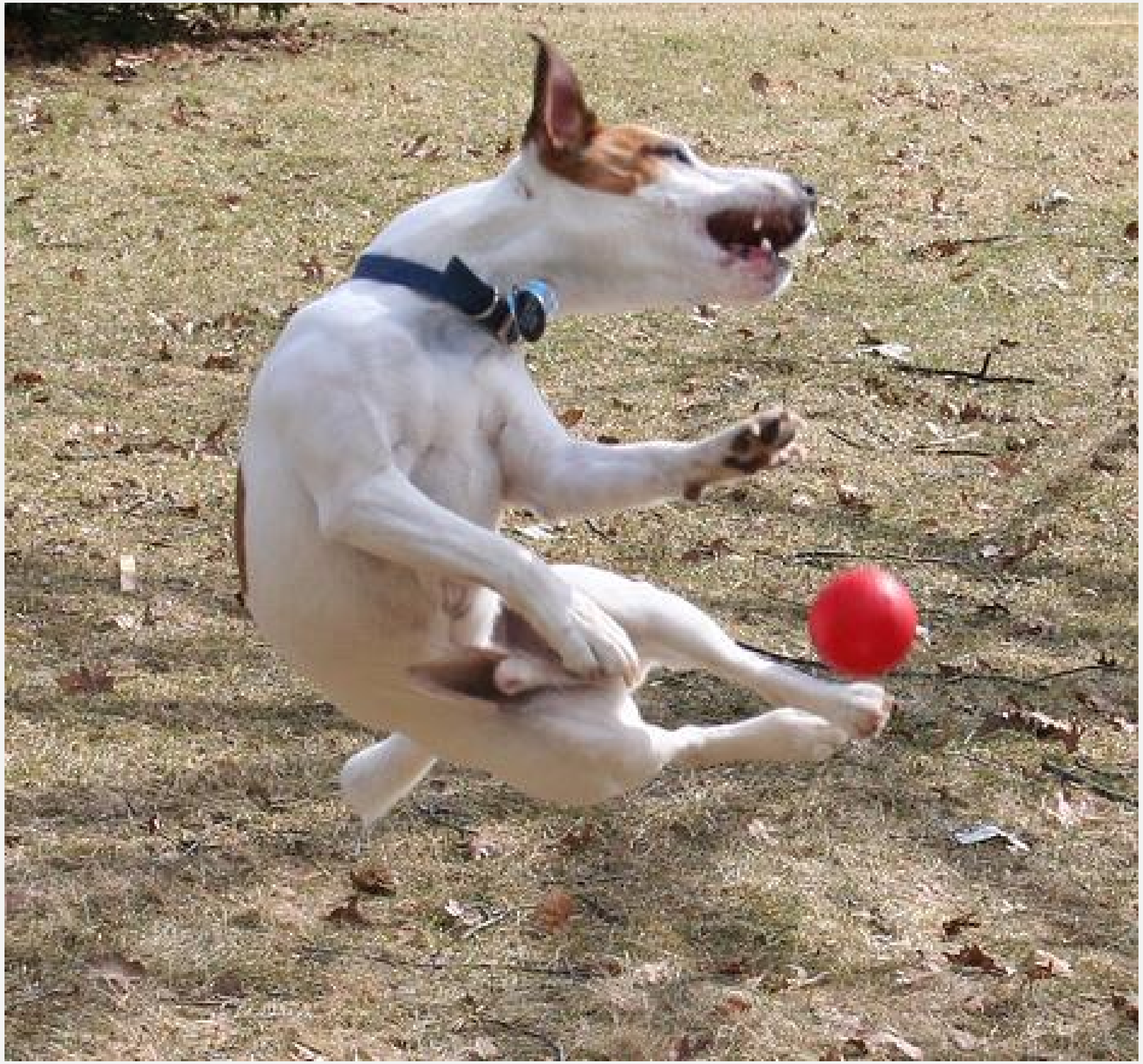
- A white dog
- A red ball

Relationship in the figure:

- A white dog
- Jumping in the air
- Catching
- A red ball

Caption:

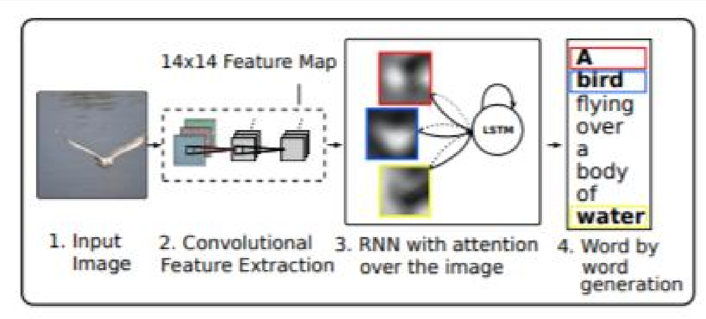
- A white dog is jumping in the air attempting to catch a red ball.



RELATED WORK

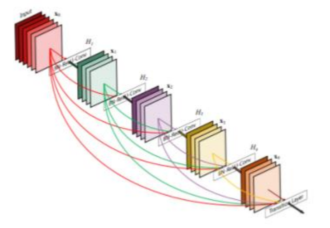
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention(2016)

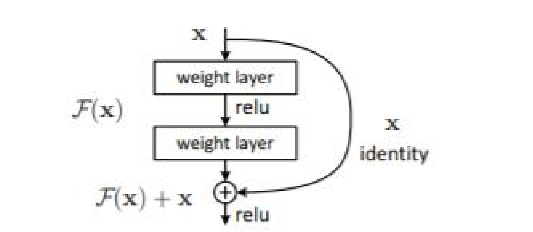
- *Using Attention scheme: Attention scheme allows salient features to the forefront as needed*
- *Using CNN to distill information from image: Generate features vectors for attention scheme*
- *Using LSTM to generate natural language*


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

CNN Architectures

- DenseNet: connects each layer to every other layer in a feed-forward fashion.
- Resnet: layers fit a residual mapping





CHALLENGES

- How to distill information from the image efficiently ?
  - Lower features of CNN. Informative and efficient representation of the image and also keep the 2-D structure.
  - Which kinds of CNN architecture is good?

- How to allow salient features to the forefront as needed?
  - Attention Scheme.
  - How to incoporate it in the NLP model? LSTM or Transformer

- How to evaluate?
  - Automatic Metric: BLEU and METEOR: Considering the words precision, synonyms and recalls

- Difficulties from Coding:
  - Complicated Structure of Transformer
    - ✓ Develop Layer classes and then integrate

  - How to cooperate for coding?
    - ✓ Full communication beforehand to clarify the function of each component.
    - ✓ Clarify the shape and data type of input and output in each part.

CONTRIBUTIONS

- **Combine CNN and LSTM with the MLP attention mechanism**
- Image Attention + Caption Attention → Attentional Context

- **Different CNN models**
- VGG19, Resnet152 and Densenet161
- Compare training-time and performance on Flicr8k dataset
- BLEU metric (BLEU-1, 2, 3, 4) and METEOR metric

- **Transformer Blocks in Decoder**
- Transformer replaces LSTM with added attention mechanisms
- Compare training-time and performance

EXPERIMENT

- Data : Flickr8k , 8000 images, each with 5 reference sentences
- Evaluation procedures : BLEU and METEOR
- Quantitative analysis

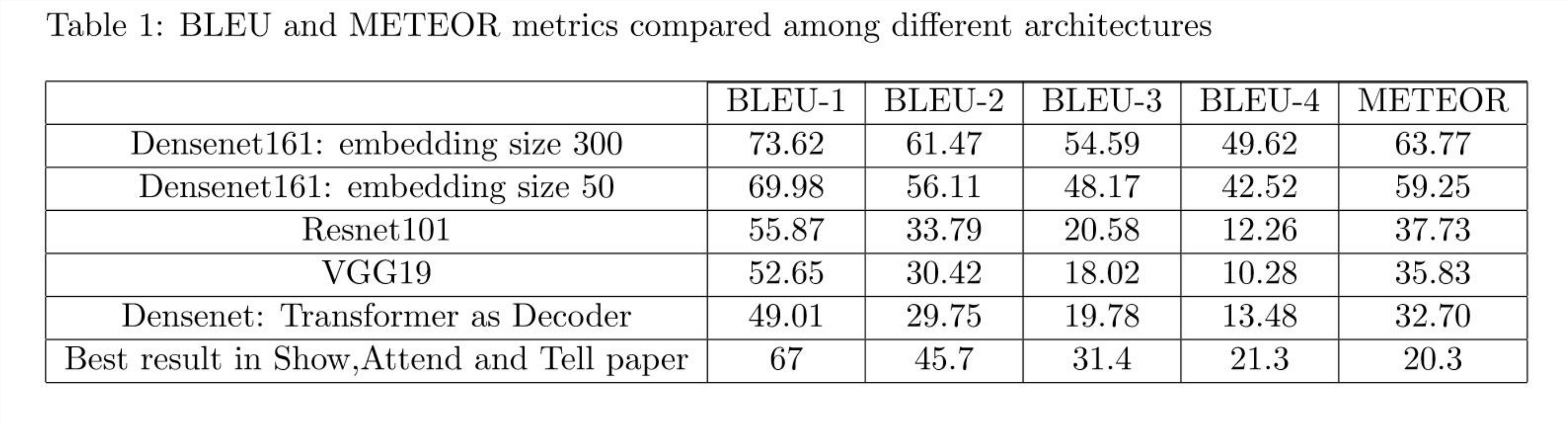


Table 1: BLEU and METEOR metrics compared among different architectures

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Densenet161: embedding size 300	73.62	61.47	54.59	49.62	63.77
Densenet161: embedding size 50	69.98	56.11	48.17	42.52	59.25
Resnet101	55.87	33.79	20.58	12.26	37.73
VGG19	52.65	30.42	18.02	10.28	35.83
Densenet: Transformer as Decoder	49.01	29.75	19.78	13.48	32.70
Best result in Show,Attend and Tell paper	67	45.7	31.4	21.3	20.3

EXPERIMENT

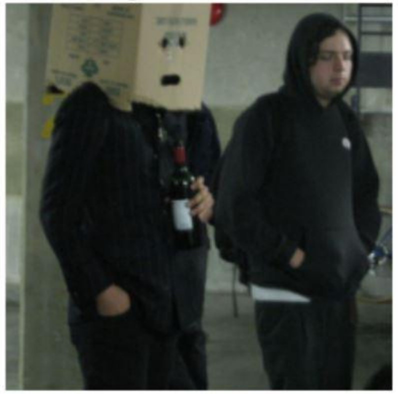
- Caption examples

a man wearing a jean jacket nd a pair of jeans walking next to a red brick building .

three people stand outside a fast food joint with their drinks .

a group of people play soccer on a court .

a man and woman are standing near a brick wall holding a video camera .



CONCLUSIONS

- Proposed an attention-based combination of different CNN architectures and LSTM or Transformer to generate captions on the image efficiently
- DenseNet has the best performance among three CNN architectures
- LSTM outperforms Transformer blocks in decoding
- Learnt captions correspond very well to human intuition