Show and Tell:
A Neural Image Caption Generator

By
Hanwen Hu, Chunlei Liu, Renjie Wei, Xinyan Yang

Dataset --- MSCOCO

- **80K train images; 40K test images**
- **1.5 million object instances**

Example:

two giraffes standing by trees look at the camera.

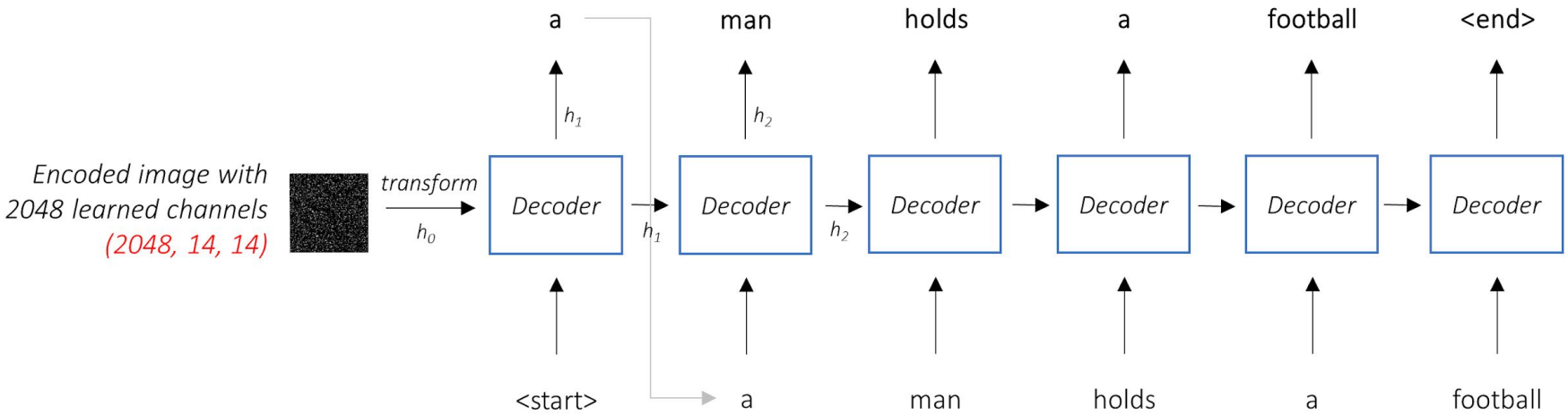
there are two giraffes standing in a wooded area.

two giraffes standing in grass and bushes looking at the camera.

a couple of giraffe standing on top of a grass covered field.

two giraffes are in a field with trees on a sunny day.

Implementation



Models and Hyperparameters

Encoder: ResNet50, ResNet101, ResNet152

Decoder: LSTM

Evaluation Metrics: BLEU - 4

Models and Hyperparameters

Batch Size: 32

Embedding Dimension: 2048.

Decoder Dimension: 512.

Dropout Rate of Decoder: 0.3, 0.5, 0.8

Encoder Learning Rate: 1e^-4

Decoder Learning Rate:1e^-4

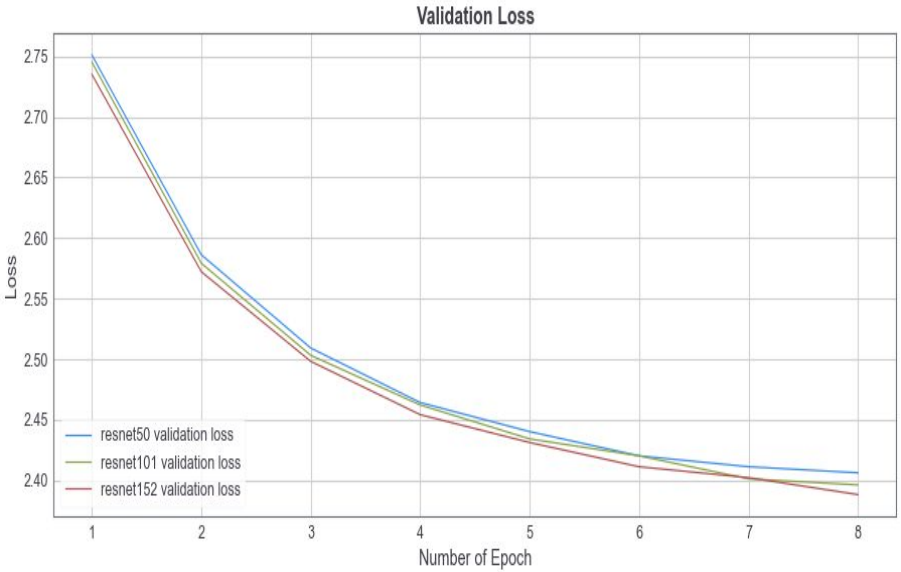
Encoder and Decoder Optimizer: Adam

Criterion: Cross Entropy Loss

Data Preprocessing

- Unify the number of captions per image (default as 5) to avoid imbalance.
- Add “<start>” to the heads of all captions, and “<end>” to their tails. Substitute the words not showing in training as “<unk>”.
- Unify the length of captions (default as 50). Record the original caption lengths for later use.
- Encode all captions with the *wordmap*.
- For images with only 2 channels (i.e. grey scale images), transform them into ‘RGB’ images.
- Resize all images to 224*224, and do the specified transformation, such as normalization.

Results -- Loss & BlEU-4 Comparison

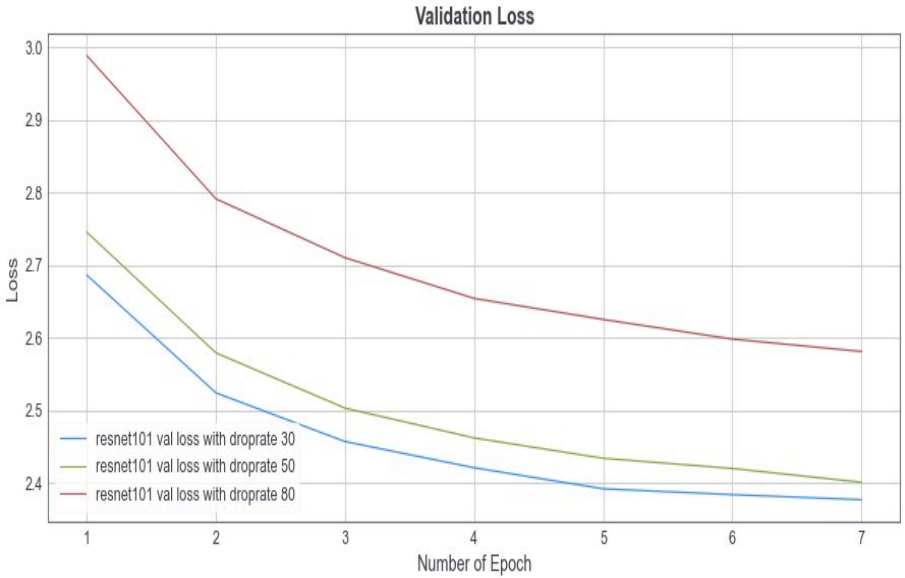


BLEU - 4 Score

Dropout 0.5, Beam Size 5

- Resnet50 : 0.2167
- Resnet101 : 0.2178
- Resnet152 : 0.2178

Results -- Loss Comparison



BLEU - 4 Score

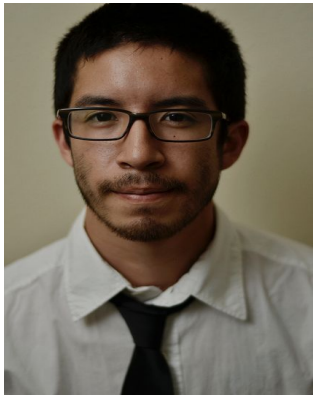
Resnet101, Beam Size 5

- Dropout 0.3: 0.2175
- Dropout 0.5: 0.2178
- Dropout 0.8 : 0.2179

Example Exhibition

Good Examples:

a man holding a tennis racquet on a tennis court



a man wearing a tie and glasses

a bunch of kites that are in the air

Example Exhibition

Neutral Examples:

a dog that is standing on a beach

a man is doing a trick on a skateboard

a black and white photo of a person on a motorcycle

Example Exhibition

Bad Examples:

a bathroom with a toilet and a sink

a couple of people that are in the back of a truck

a view of a lake with a mountain in the background

Thank You

Any Questions?