A

ResNet-152
(extract features)

Feature maps
(1, 2048)

B

ResNet-152
(extract features)

Feature maps
(1, 2048)

Concatenation
(1, 4096)

4096 neurons

512 neurons

Feature vector
(1, 512)

Encoding

LSTM
(produce captions)

Output
(1, 512)

512 neurons

130 neurons

Probability of token 1

Probability of token 2

Probability of token 3

Probability of token *n*

Repeat until:

1) the maximum sequence length is reached

OR

2) the sequence ending token is reached

e.g., "<START>"

Highest predicted token
(1)

Word embedder
(130, 512)

Embedding
(1, 512)

Decoding