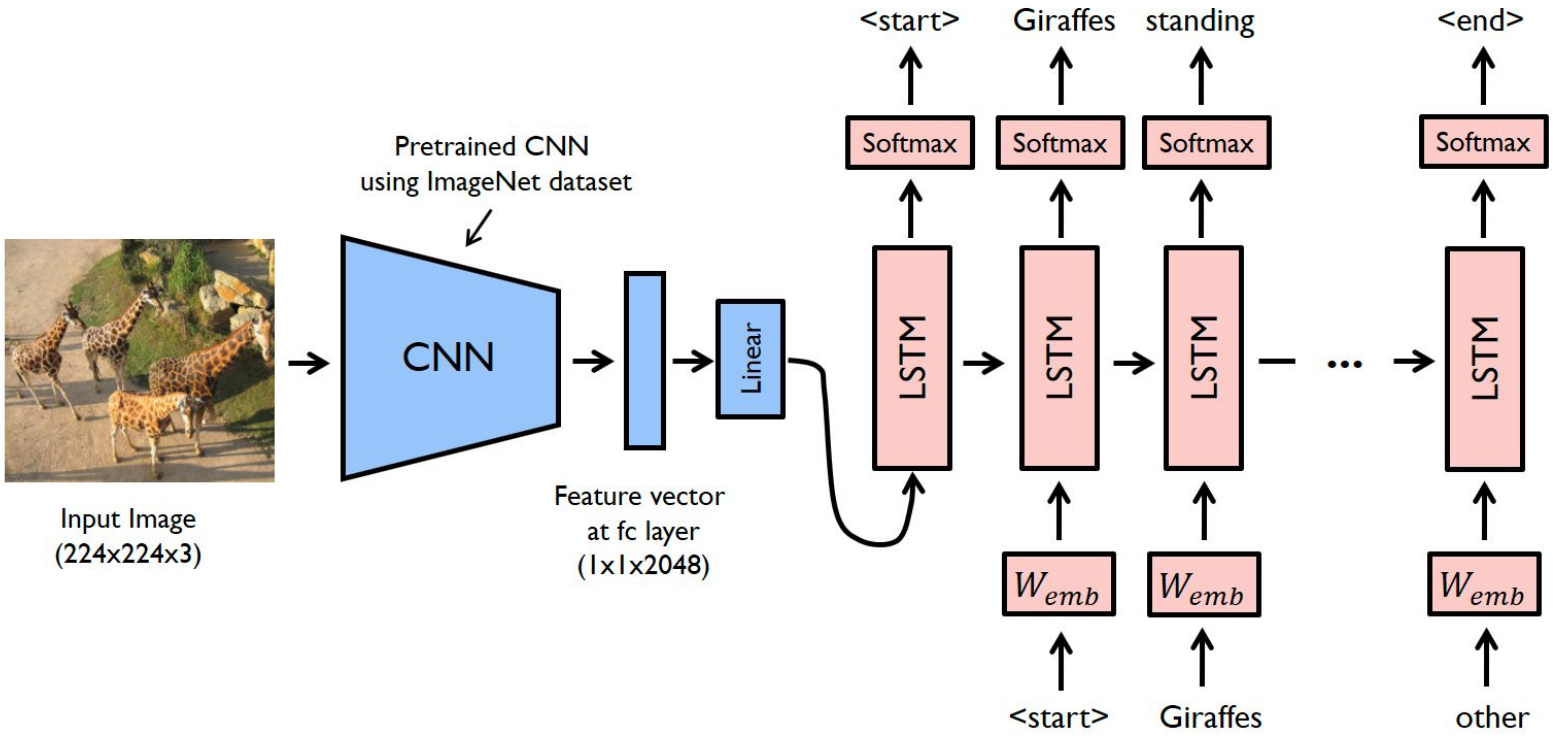
Where to put the Image in an Image Caption Generator

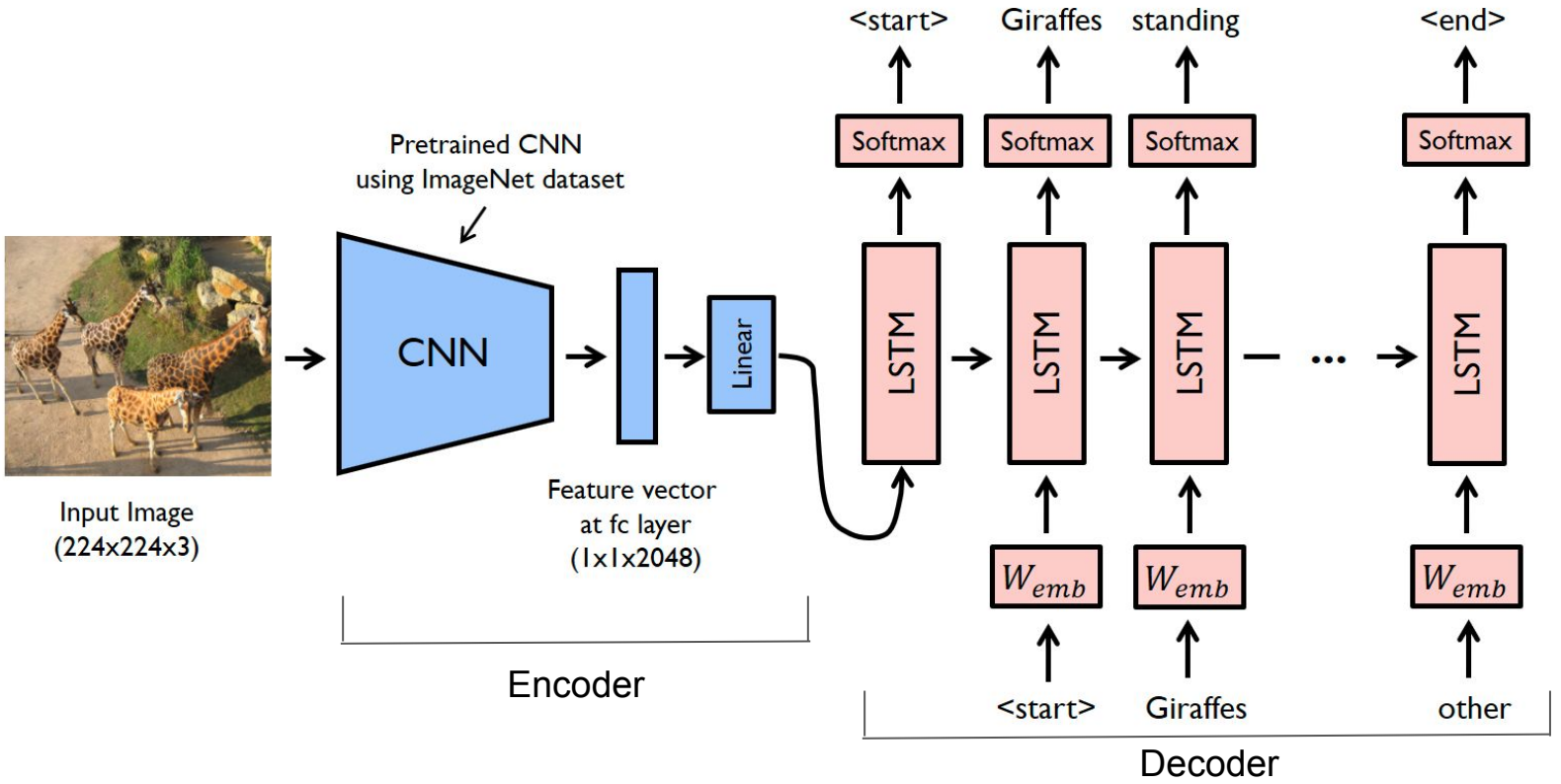
Language and Image Processing

Gunay Abdullayeva

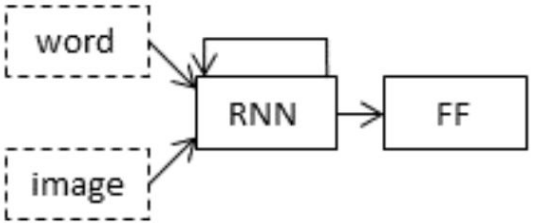
How image captioning works?



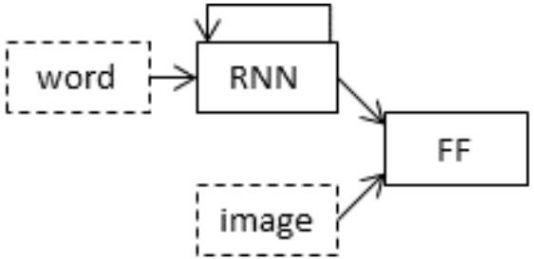
How image captioning works?



Problem Statement

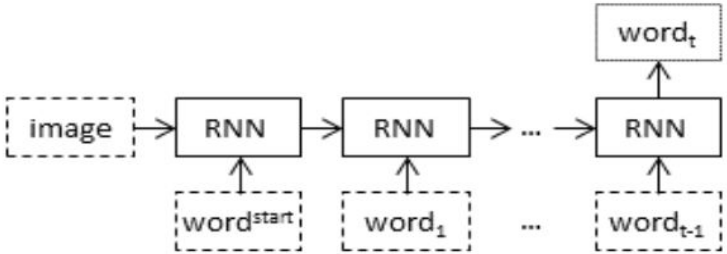


(a) Conditioning by injecting the image means injecting the image into the same RNN that processes the words.

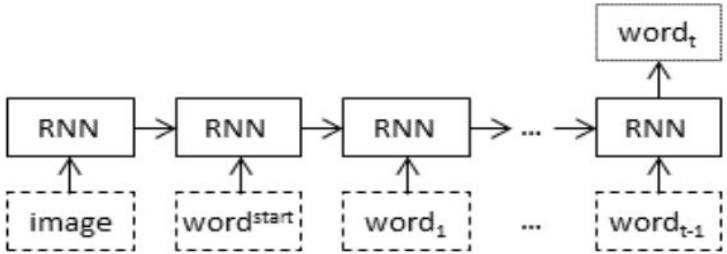


(b) Conditioning by merging the image means merging the image with the output of the RNN after processing the words.

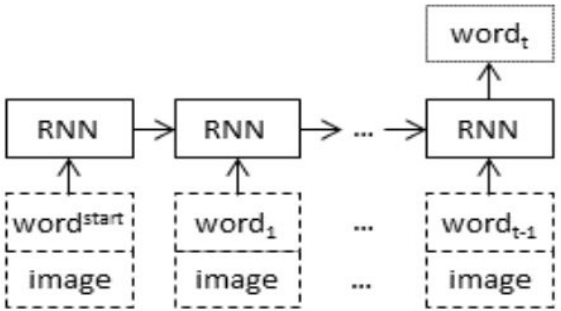
Figure 2: The inject and merge architectures for caption generation. Legend: RNN - recurrent neural network; FF - feed forward layer.



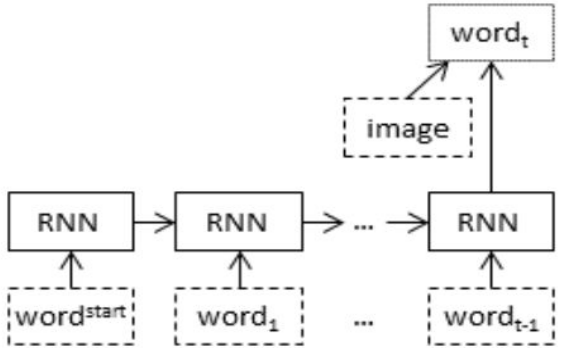
(a) **Init-inject:** The image vector is used as an initial hidden state vector for the RNN.



(b) **Pre-inject:** The image vector is used as a first word in the prefix.



(c) **Par-inject:** The RNN accepts two inputs at once in every time step: a word and an image.



(d) **Merge:** The image vector is merged with the prefix outside of the RNN.

Figure 3: Different ways of conditioning a neural language model with an image. The feedforward layer was left out to save space.

General Configurations

Dataset: Flickr 8K: 6000 train, 1000 dev, 1000 test splits

Pretrained network: VGG

- Remove the last layer from the model
- Get the image features which contains 4096 element vector
- Normalize the image

Text preprocessing:

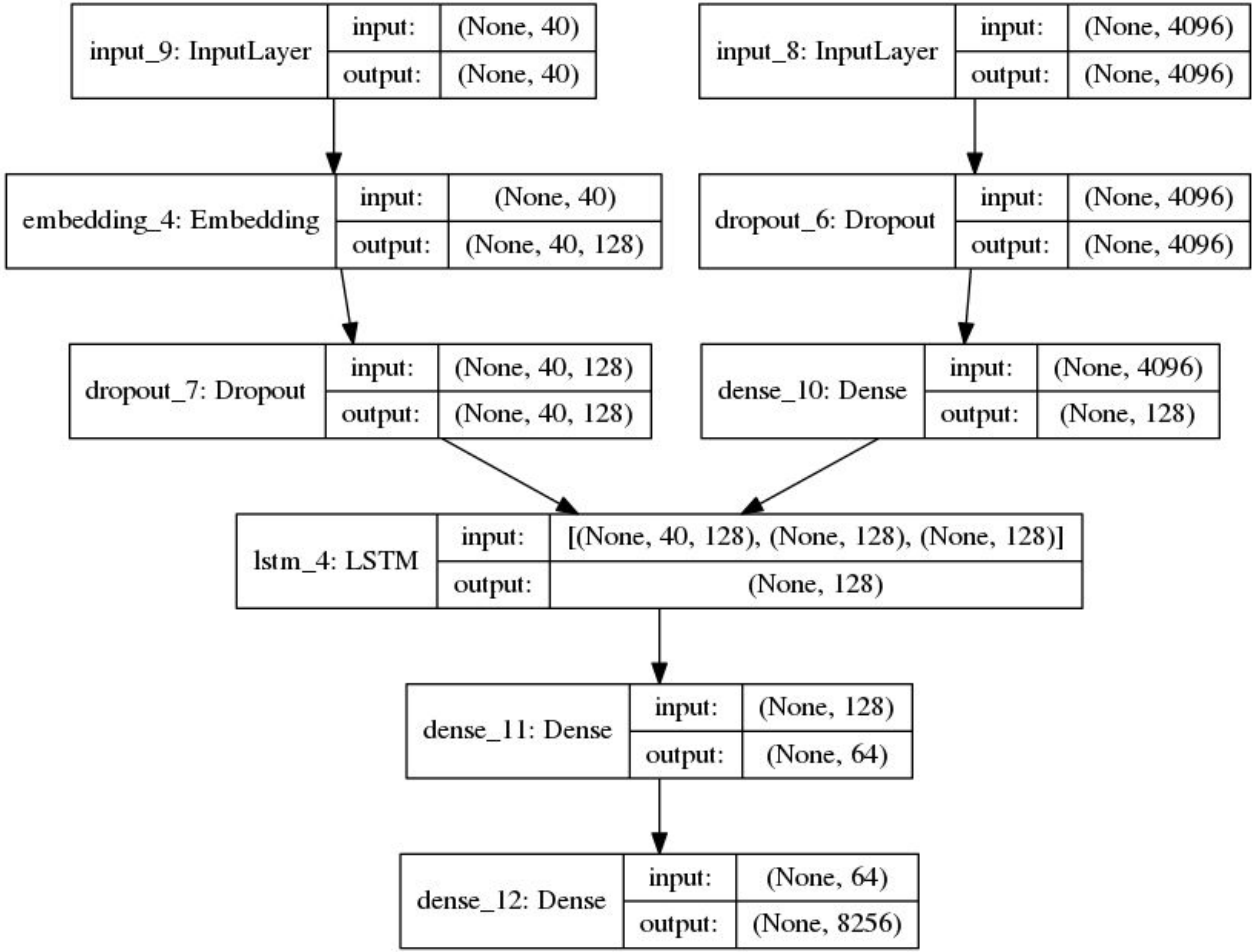
- Convert to lower case
- Remove punctuation from each token
- Remove tokens with numbers in them

Embedding size: vocabulary size = 8256

Padding the sequence with maximum length sentence = 40

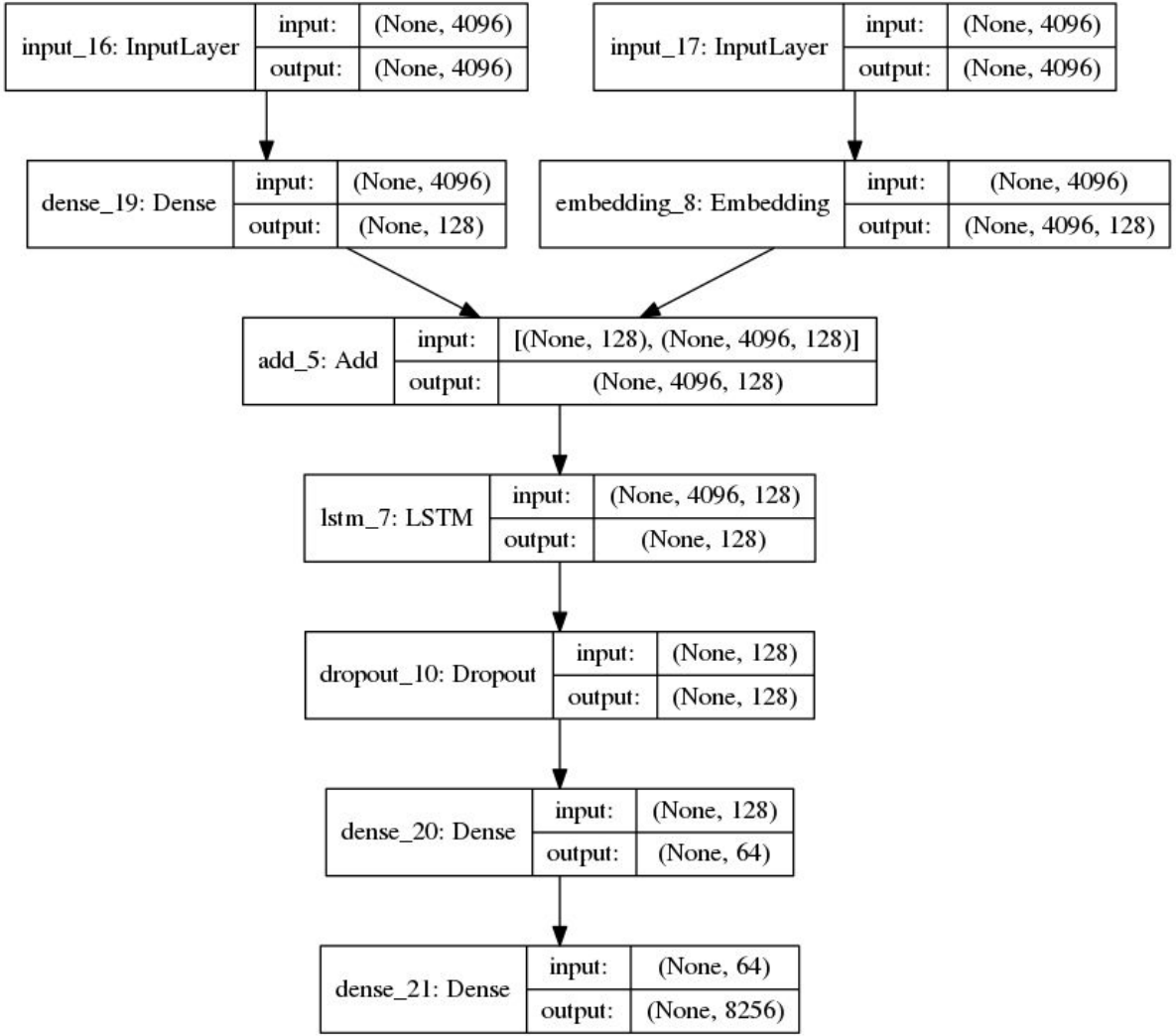
Init-inject Model

The image vector is used initial hidden state



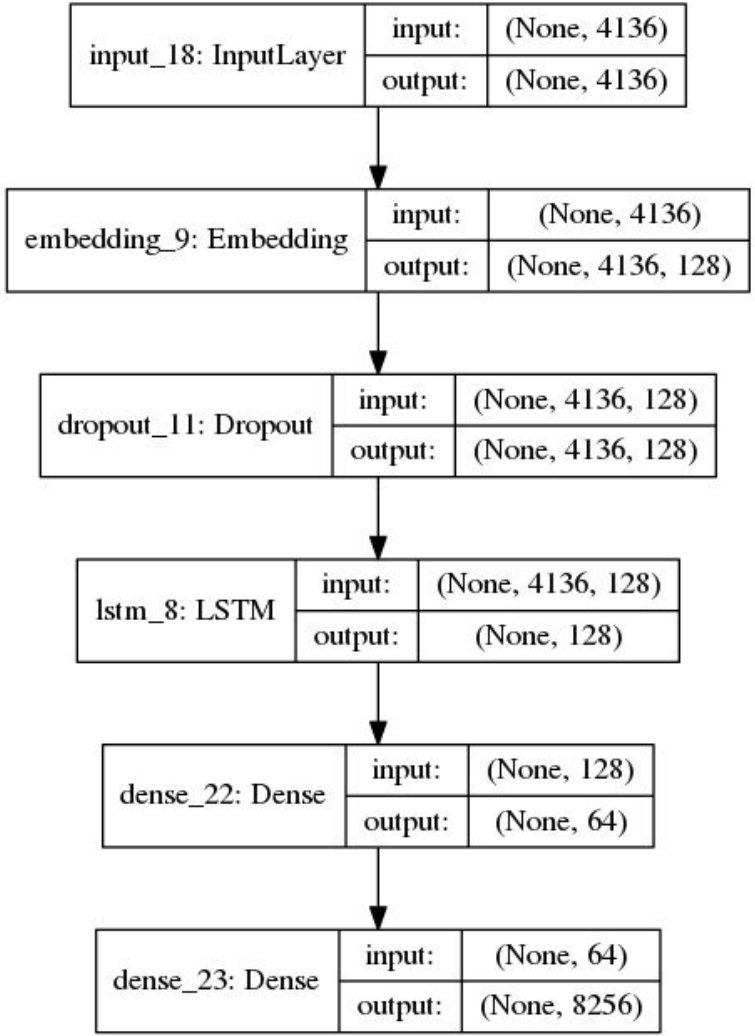
Pre-inject Model

The image vector is used as first word prefix



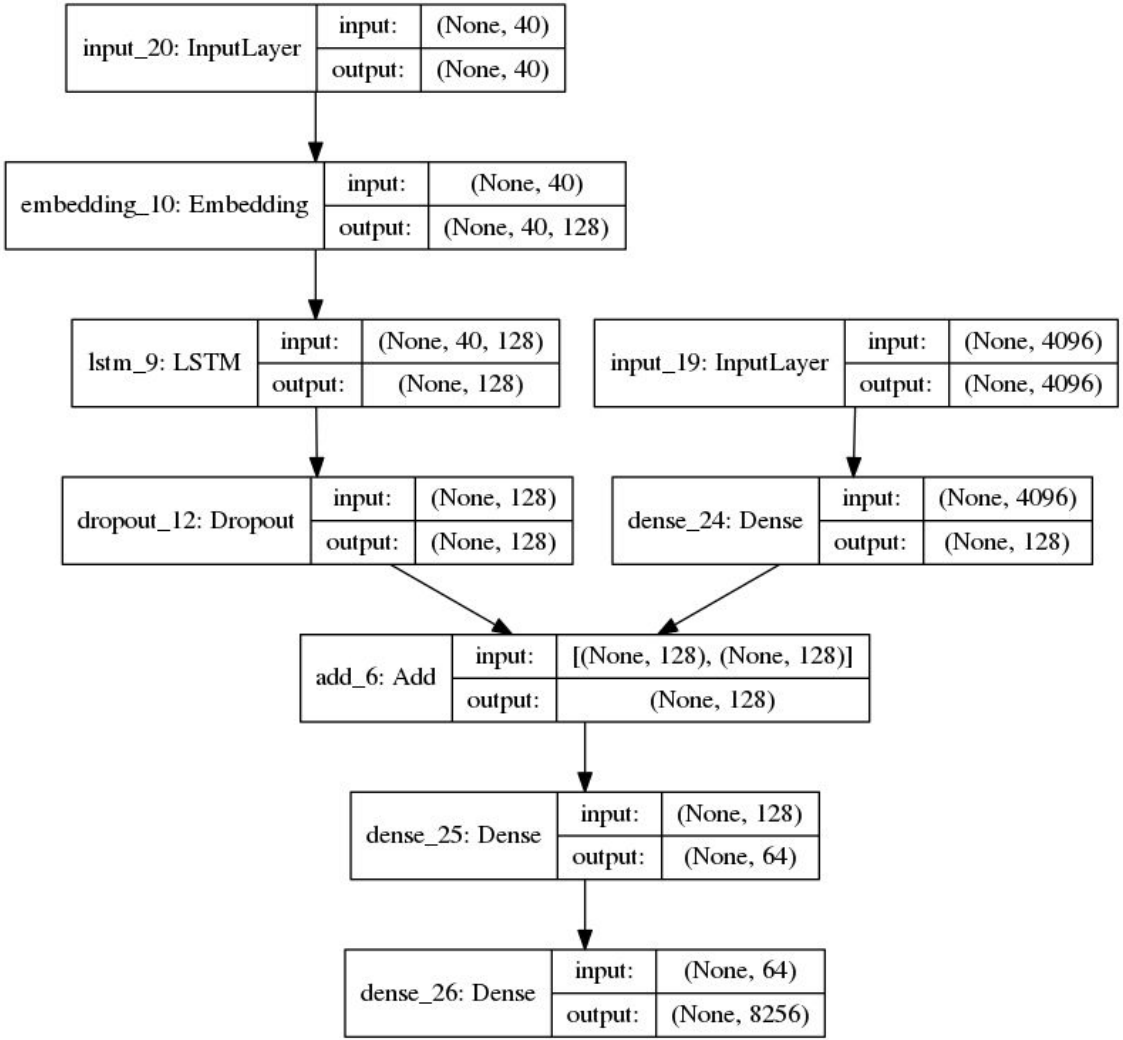
Par-inject Model

RNN accepts two inputs at once in every time steps: image and word



Merge Model

The image vector is merged with the prefix outside of RNN



Evaluation Results

Models were trained for 1 epoch

Methods	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**
init-inject	**0.303877**	0.118964	0.063030	0.022465
pre-inject	0.283756	0.083527	0.031183	0.011743
par-inject	0.238469	0.092813	0.030492	0.014862
merge	0.294013	**0.123476**	**0.075956**	**0.028844**

Prediction Results

Actual:
A dog runs across a grassy lawn near some flowers .
A brown dog running over grass .
A yellow dog is playing in a grassy area near flowers .
A brown dog with its front paws off the ground on a grassy surface near red and purple flowers .
A brown dog running

Predicted:
A black and white dog is running through a field .

Prediction Results

Actual:
The guard drives towards the basket during a college basketball game .
A basketball player in a whit outfit is dribbling the ball around a player in a red outfit as the referee looks on .
A referee watches two basketball players playing basketball .
A basketball player dribbles the ball while another blocks him and an official looks on .
A Miami basketball player dribbles by an Arizona State player .

Predicted:
A man in a red uniform is playing soccer .

Prediction Results

Actual:
A woman next to a dog which is running an obstacle course .
The dog is jumping over the hurdles beside a woman .
A woman walking with a Sheltie through a competition obstacle course .
A woman in a blue shirt guides her dog over an obstacle .
a woman is running beside a dog that is jumping over a red and white obedience training fence .

Predicted:
A little boy jumps over a hurdle .

Prediction Results

Actual:
A surfer in all black is riding a wave .
A woman in a black wetsuit surfs in bad weather .
A surfer rides the waves .
A person is surfing .
A person in a black wetsuit is surfing in the ocean with a wave coming down .

Predicted:
A surfer is riding a wave .