Deep Recurrent Architecture based Scene Description Generator for Visually Impaired

THE 12TH INTERNATIONAL CONGRESS ON ULTRA MODERN TELECOMMUNICATIONS AND CONTROL SYSTEMS (AIDL-HCSY, 2020)

Aviral Chharia[1], Rahul Upadhyay[2]

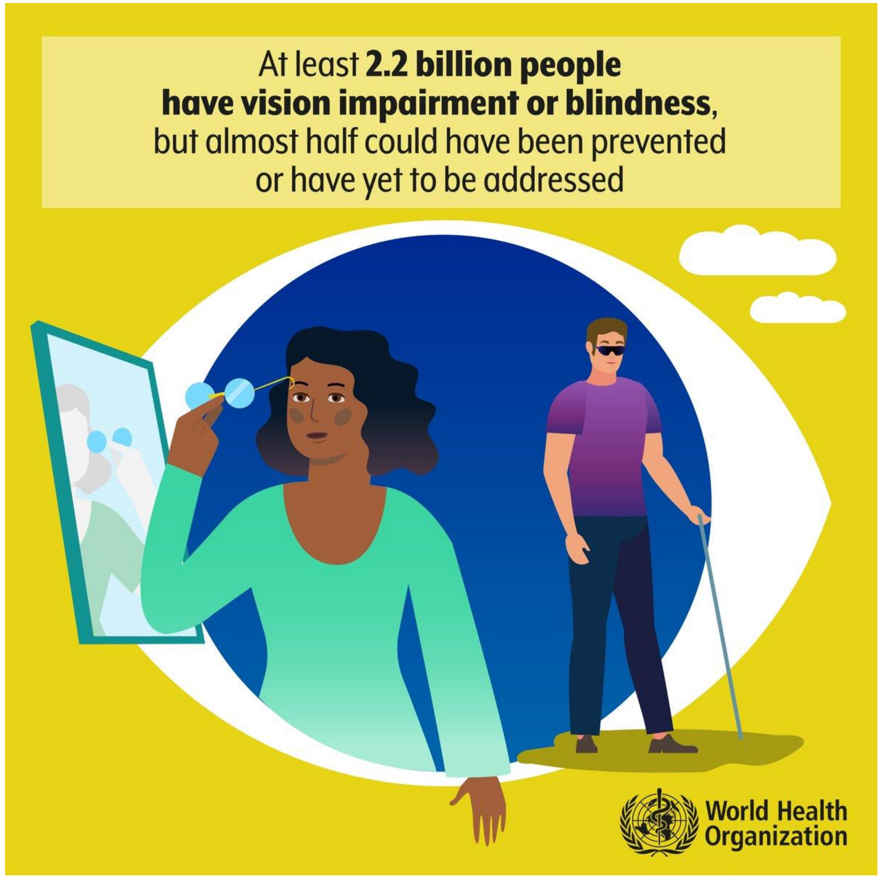
[1]Department of Mechanical Engineering, Thapar Institute of Engineering & Technology, India
[1]Department of Electronics & Communication Engineering, Thapar Institute of Engineering & Technology, India

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

12 ICUMT

Visual Impairment – Present Challenges

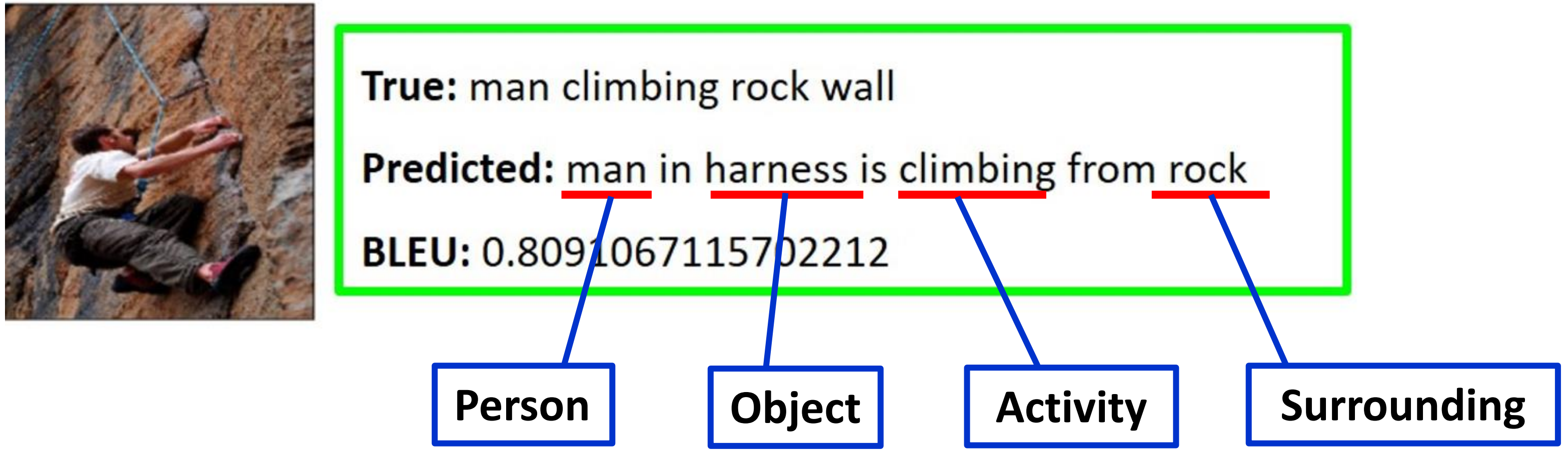
- According to WHO, 2.2 billion people are visually impaired today.
- Still far behind a permanent medical cure for visual impairment.
- Day to Day challenges faced by a Visually Impaired person.
- Urgent need to improve the quality of life of those visually impaired & towards this endeavor, assistive technology plays an essential role.



Reference: World Health Organization, 2020

Assistive Technology- Image Captioning

- Fundamental & challenging problem in artificial intelligence.
- Involves automatically describing contents of an image with proper linguistic properties.
- Involves capturing objects, people, surroundings, etc. & their relationship to each other & activities they are involved in.
- Semantic knowledge is expressed in natural language which requires language model in addition to visual understanding.
- Combines advanced level of computer vision with natural language processing (NLP) methods [2].



Related Works

1. **Template Matching Techniques** [4-6]
   Manually designed & hardcoded templates. Descriptions generated were neither comparable nor expressive.

2. **Retrieval Based Approach** by Hodosh et al. [7]
   Selects a set of visually similar images from a database of training images & fits the nearest captions of these to the test image. Limits output variety & fails to generate new captions if similar images aren't present in training set.

3. **Neural Image Captioning model** by Vinyals et al. [8]
   An encoder-decoder based model is used, in which the output of the encoder (final convolutional layer) is used as the input to the decoder.

4. **Add-on Mechanisms and improvements** by Xu et al. [9]
   Attention mechanisms, GloVe & word2vec algorithm to obtain low-dimensional vector representations of words. RNNs that combine image features with language modelling have been used to generate captions.

5. **Multi-model RNN based architecture** by Karpathy and Fei [10]
   Made use of inferred alignments while training to generate rich descriptive captions.

Proposed End-to-End Methodology

System captures video through vision enabled eye wear & after processing, visually impaired person hears image description in real-time

- The vision enabled eye wear of the visually impaired person captures scenes as real-time video.

Proposed End-to-End Methodology

System captures video through vision enabled eye wear & after processing, visually impaired person hears image description in real-time

Extract Image Frames at every $T = t_1 + t_2$ sec →

Image Frame extracted from the Captured Video

An Image Frame (extracted from Video)
224 x 224 x 3

- The vision enabled eye wear of the visually impaired person captures scenes as real-time video.
- The image frame from the video is extracted & send to the deep recurrent architecture.

Proposed End-to-End Methodology

System captures video through vision enabled eye wear & after processing, visually impaired person hears image description in real-time

Extract Image Frames at every $\boldsymbol{T} = \boldsymbol{t_1} + \boldsymbol{t_2}$ sec →

[Image Frame extracted from the Captured Video]

An Image Frame (extracted from Video)
224 x 224 x 3

VGG-16 Pretrained on ImageNet (last layer removed) →

[4096-Dimensional Image Feature Vector]

at fc2 layer of the VGG16 Net
1 x 1 x 4096

- The vision enabled eye wear of the visually impaired person captures scenes as real-time video.
- The image frame from the video is extracted & send to the deep recurrent architecture.
- The CNN (VGG-16 net pre-trained on ImageNet) obtains the 4096-dimensional image feature vector.

Proposed End-to-End Methodology

System captures video through vision enabled eye wear & after processing, visually impaired person hears image description in real-time

Extract Image Frames at every $T = t_1 + t_2$ sec →

Image Frame extracted from the Captured Video
An Image Frame (extracted from Video)
224 x 224 x 3

VGG-16 Pretrained on ImageNet (last layer removed) →

4096-Dimensional Image Feature Vector
at fc2 layer of the VGG16 Net
1 x 1 x 4096

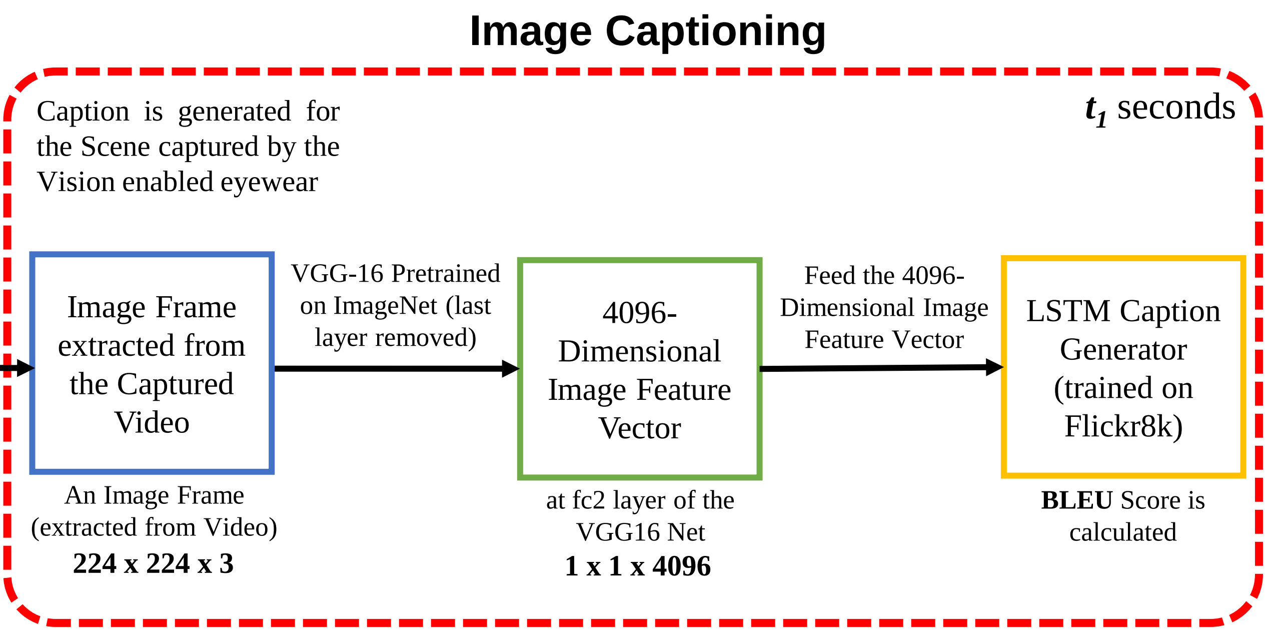
Feed the 4096-Dimensional Image Feature Vector →

LSTM Caption Generator (trained on Flickr8k)
BLEU Score is calculated

- The vision enabled eye wear of the visually impaired person captures scenes as real-time video.
- The image frame from the video is extracted & send to the deep recurrent architecture.
- The CNN (VGG-16 net pre-trained on ImageNet) obtains the 4096-dimensional image feature vector.
- These feature vectors are then fed into an LSTM to generate captions and BLEU score is calculated.

Proposed End-to-End Methodology

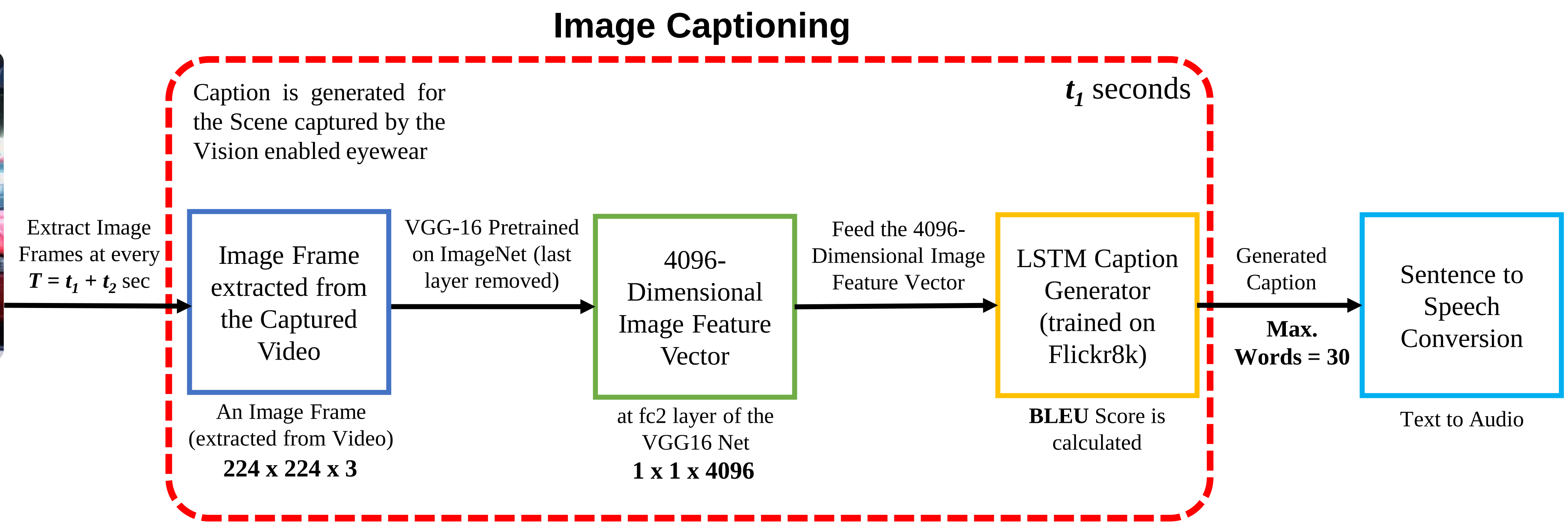
System captures video through vision enabled eye wear & after processing, visually impaired person hears image description in real-time



- The vision enabled eye wear of the visually impaired person captures scenes as real-time video.
- The image frame from the video is extracted & send to the deep recurrent architecture.
- The CNN (VGG-16 net pre-trained on ImageNet) obtains the 4096-dimensional image feature vector.
- These feature vectors are then fed into an LSTM to generate captions and BLEU score is calculated.
- The Image captioning part takes approx. ~ t_1 seconds.

Proposed End-to-End Methodology

System captures video through vision enabled eye wear & after processing, visually impaired person hears image description in real-time



- The vision enabled eye wear of the visually impaired person captures scenes as real-time video.
- The image frame from the video is extracted & send to the deep recurrent architecture.
- The CNN (VGG-16 net pre-trained on ImageNet) obtains the 4096-dimensional image feature vector.
- These feature vectors are then fed into an LSTM to generate captions and BLEU score is calculated.
- The Image captioning part takes approx. ~ t_1 seconds.
- The generated captions are converted to audio for the person to hear (this part takes approx. ~ t_2 seconds.)

Proposed End-to-End Methodology

System captures video through vision enabled eye wear & after processing, visually impaired person hears image description in real-time

Extract Image Frames at every $T = t_1 + t_2$ sec

Image Captioning — t_1 seconds

Caption is generated for the Scene captured by the Vision enabled eyewear

Image Frame extracted from the Captured Video

An Image Frame (extracted from Video)

224 x 224 x 3

VGG-16 Pretrained on ImageNet (last layer removed)

4096-Dimensional Image Feature Vector

at fc2 layer of the VGG16 Net

1 x 1 x 4096

Feed the 4096-Dimensional Image Feature Vector

LSTM Caption Generator (trained on Flickr8k)

BLEU Score is calculated

Text to Audio — t_2 seconds

Generated Caption is read out to Visually Impaired Person

Generated Caption

Max. Words = 30

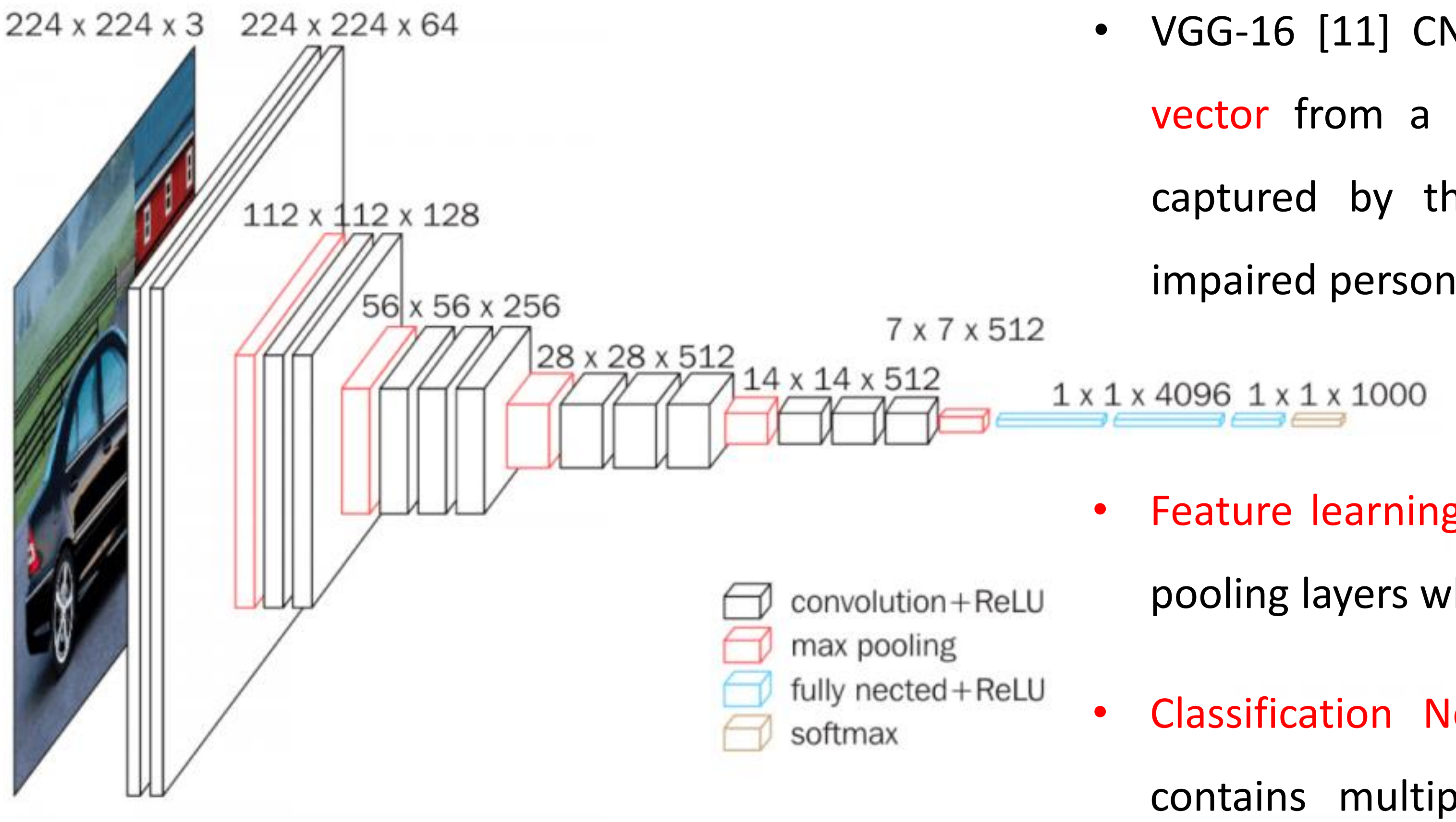
Sentence to Speech Conversion

Text to Audio

$T = t_1$ sec (Time taken to perform Image Captioning) + t_2 sec (Time taken to read out a caption to visually impaired person)

- The vision enabled eye wear of the visually impaired person captures scenes as real-time video.
- The image frame from the video is extracted & send to the deep recurrent architecture.
- The CNN (VGG-16 net pre-trained on ImageNet) obtains the 4096-dimensional image feature vector.
- These feature vectors are then fed into an LSTM to generate captions and BLEU score is calculated.
- The Image captioning part takes approx. ~ t_1 seconds.
- The generated captions are converted to audio for the person to hear (this part takes approx. ~ t_2 seconds.)
- The visually impaired person gets greater assistance through continuous feedback.

VGG-16 Net Convolutional Neural Network

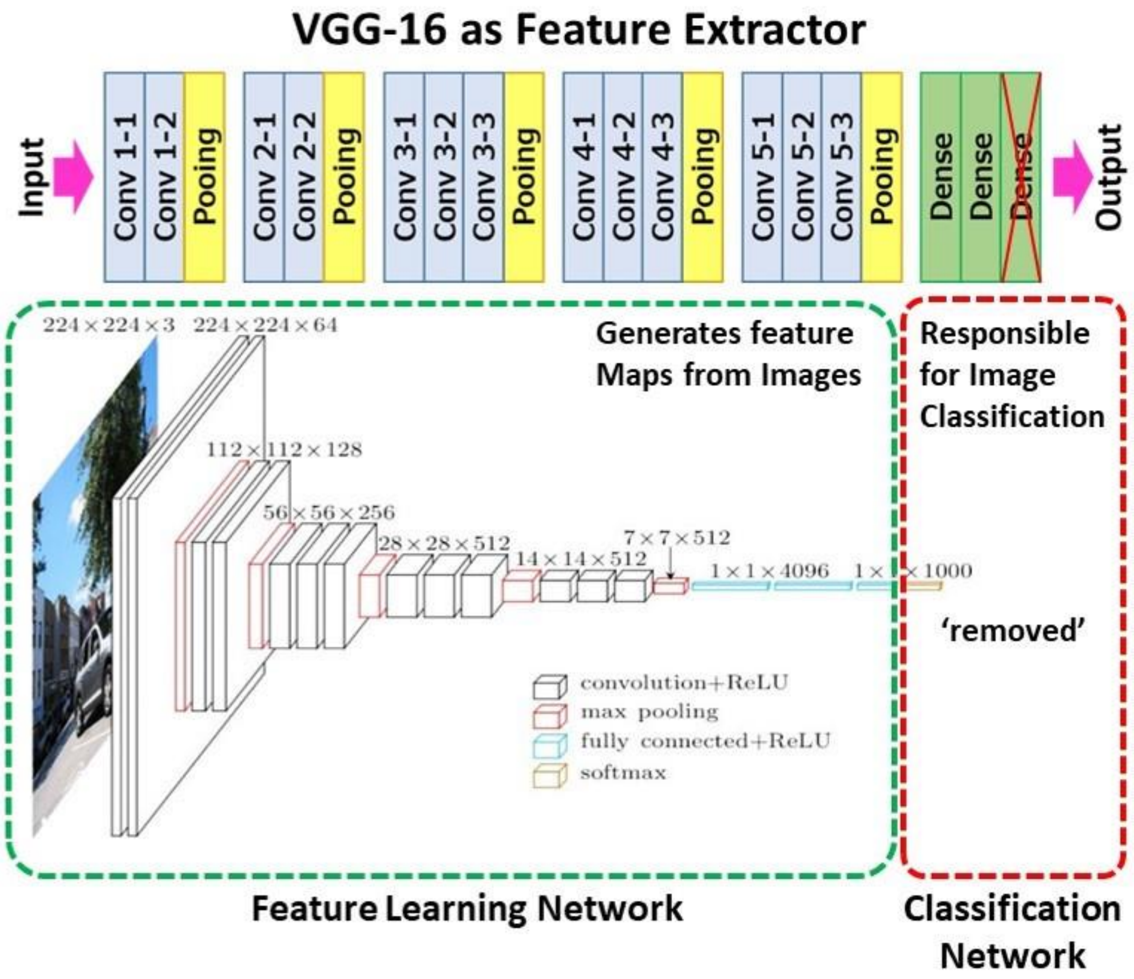


- VGG-16 [11] CNN extracts 4096-dimensional image feature vector from a single image frame of the real-time video captured by the vision enabled eye wear of a visually impaired person. It is made of 2 Networks [11]

- Feature learning network consists of multiple convolution & pooling layers which generates the image feature maps

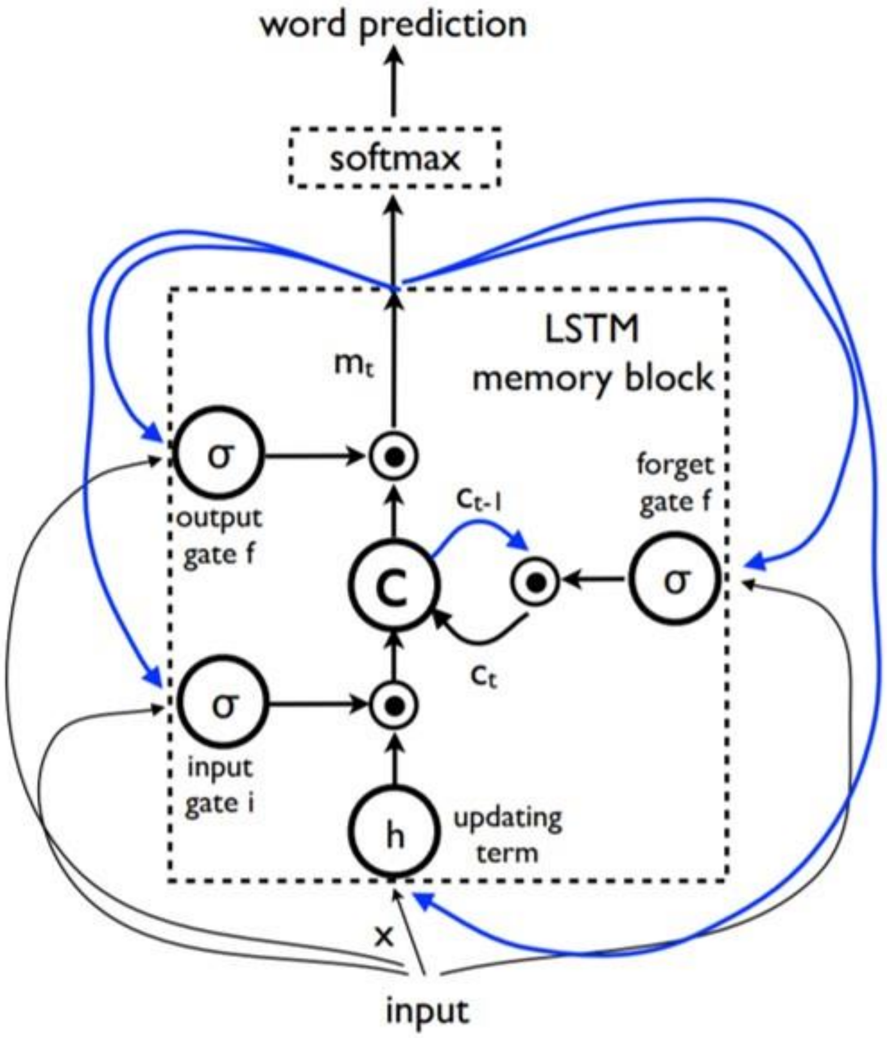
- Classification Network is used for image classification & contains multiple dense layers & a single output layer (originally tuned for classification of images into 1000 different classes).

Image Feature Extraction using VGG-16



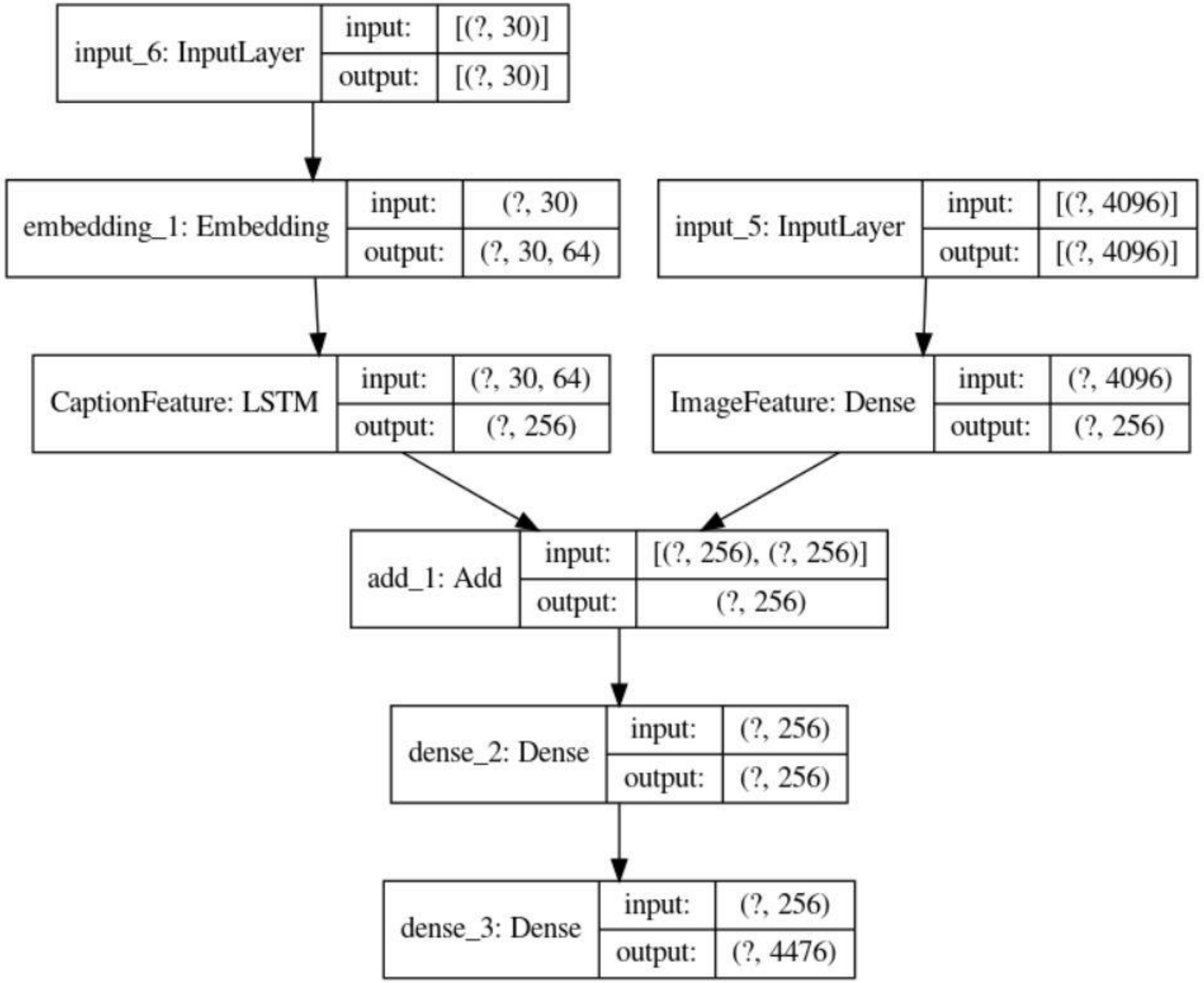
The classification network is removed & the VGG-16 net is employed as a feature extractor. Later these obtained feature vectors are fed as an input into the first layer of the Long Short Term Memory (LSTM) for language generation.

Long Short Term Memory (LSTM)

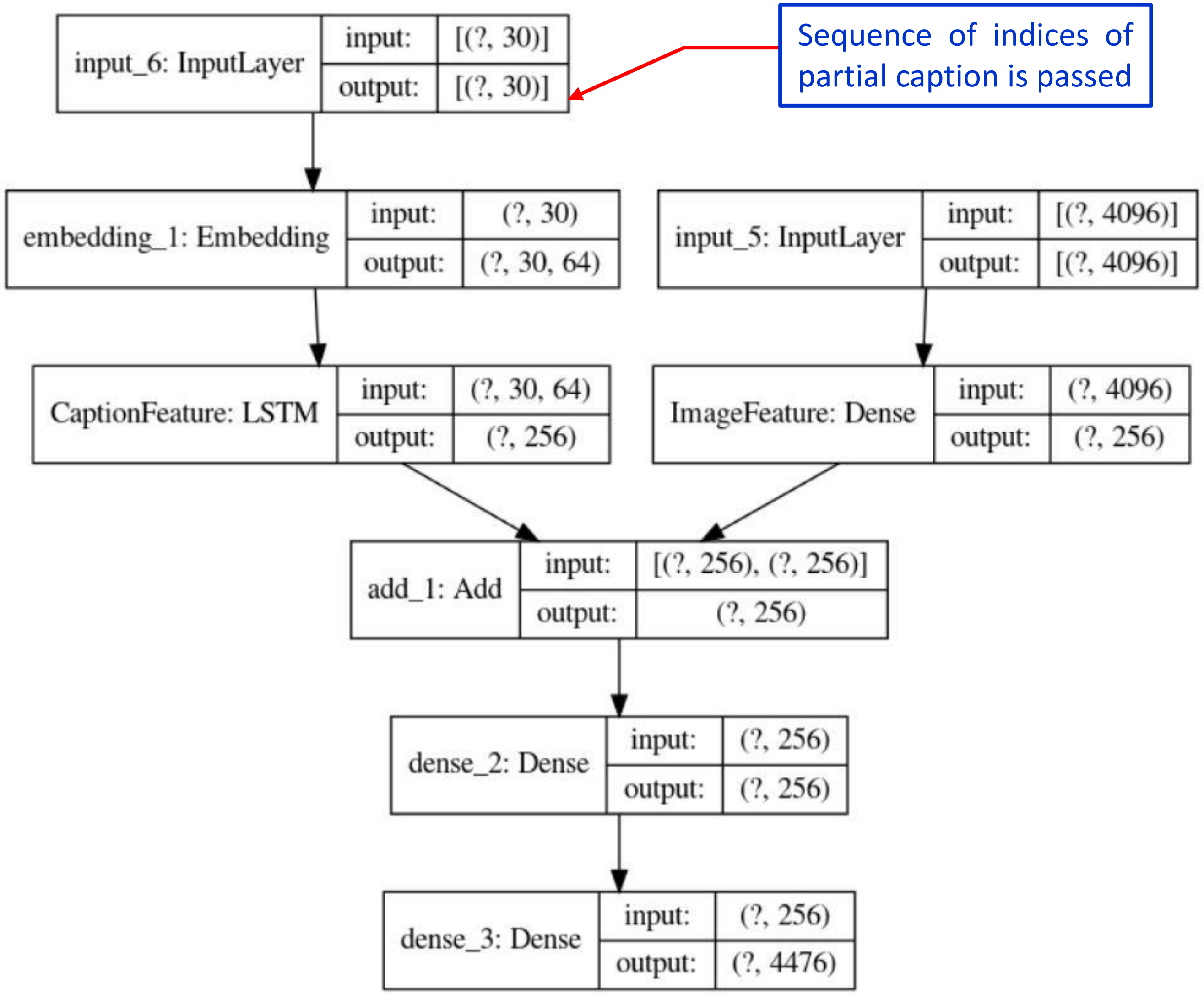


- LSTM [13] is a specialized RNN used for natural language generation.
- Although RNNs are more efficient in text generation tasks, they encounter vanishing/ exploding gradient problems resulting from propagating the gradients down through many layers of recurrent networks.
- Since LSTMs use memory units which not only allows the network to learn and forget previous hidden states but also when to update hidden states when given new information, they do not have these gradient problems.

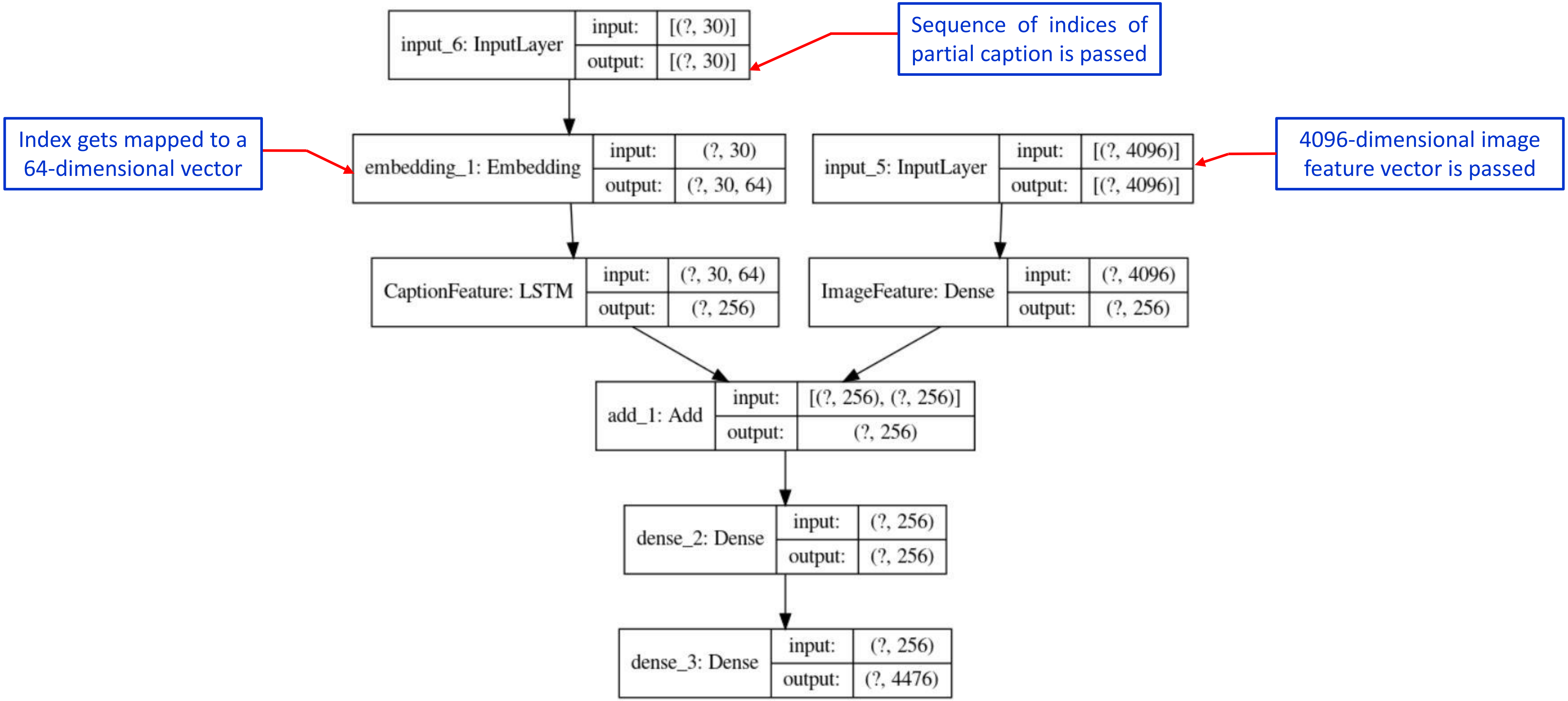
Model Architecture for Language Generation



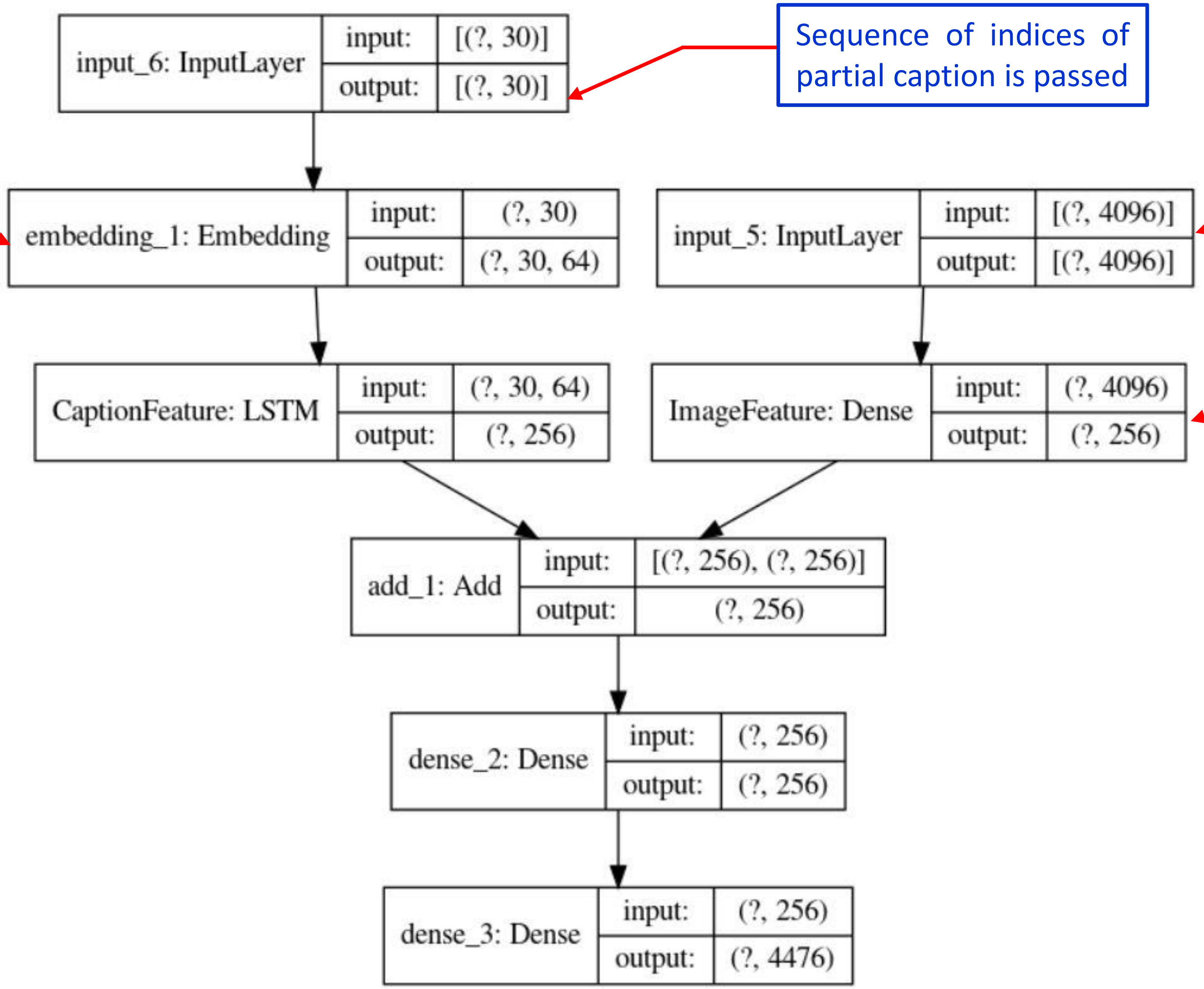
Model Architecture for Language Generation



Model Architecture for Language Generation



Model Architecture for Language Generation



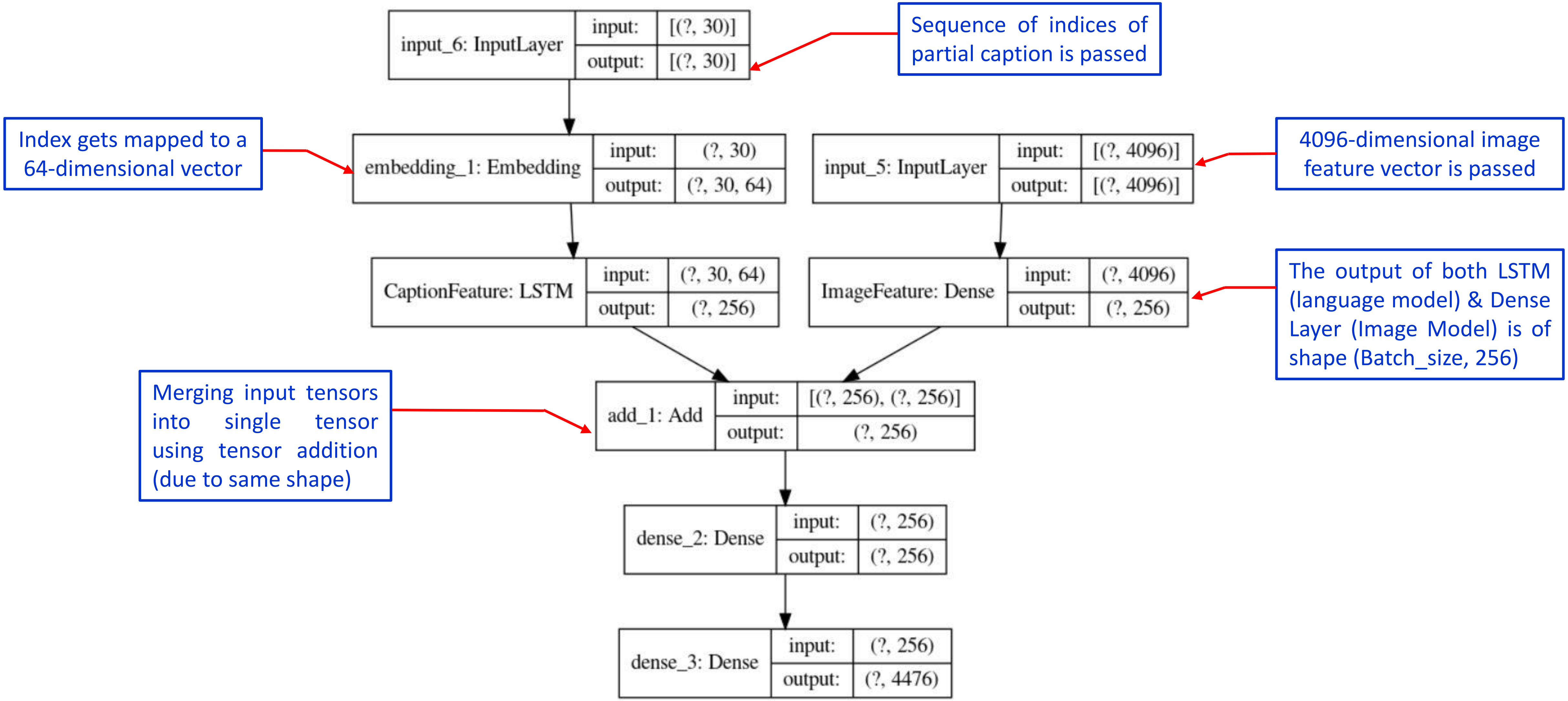
Sequence of indices of partial caption is passed

Index gets mapped to a 64-dimensional vector

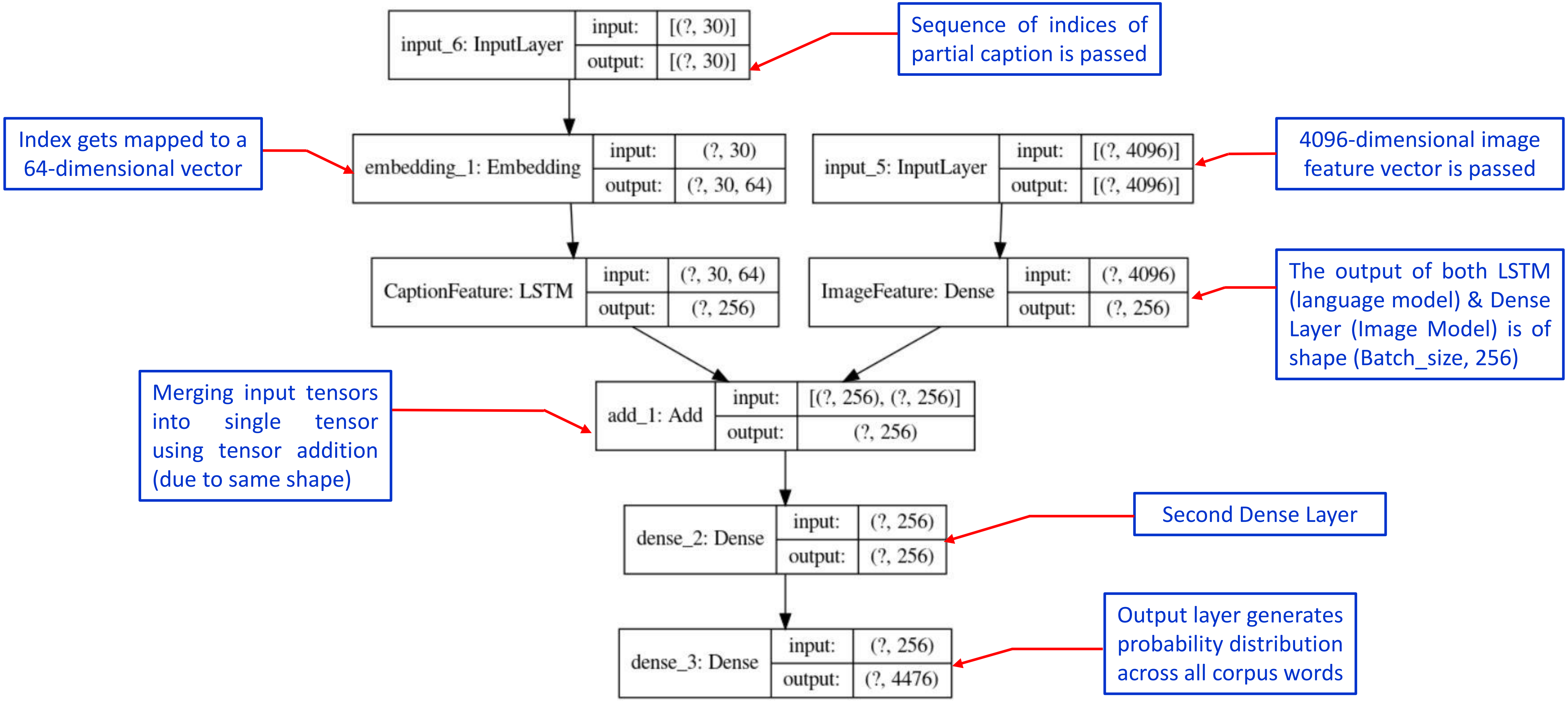
4096-dimensional image feature vector is passed

The output of both LSTM (language model) & Dense Layer (Image Model) is of shape (Batch_size, 256)

Model Architecture for Language Generation



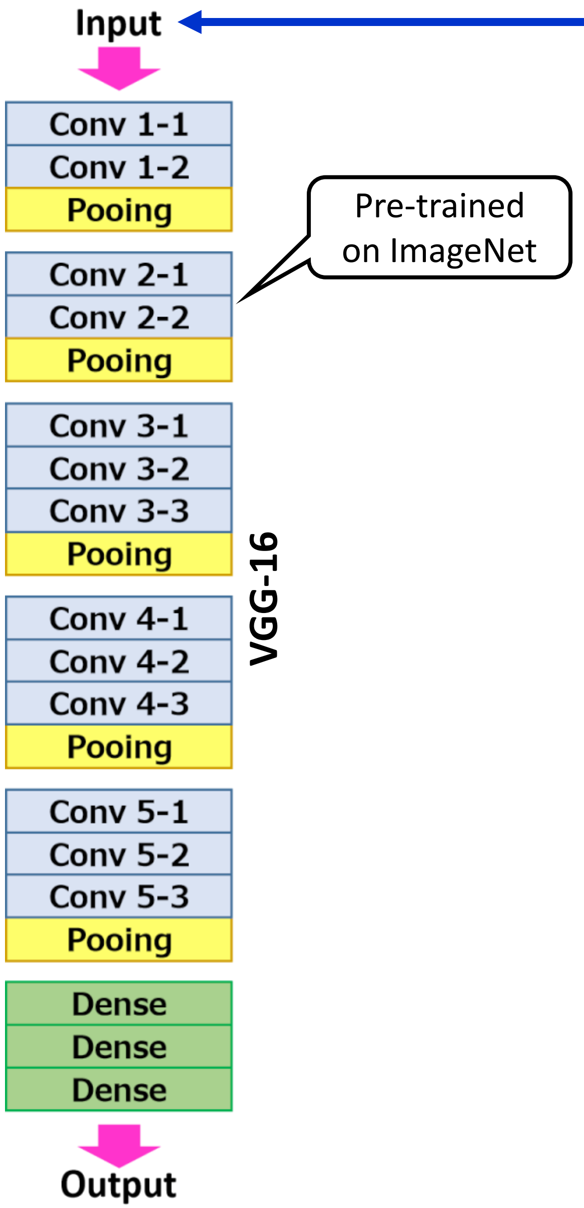
Model Architecture for Language Generation



Scene Description Generation

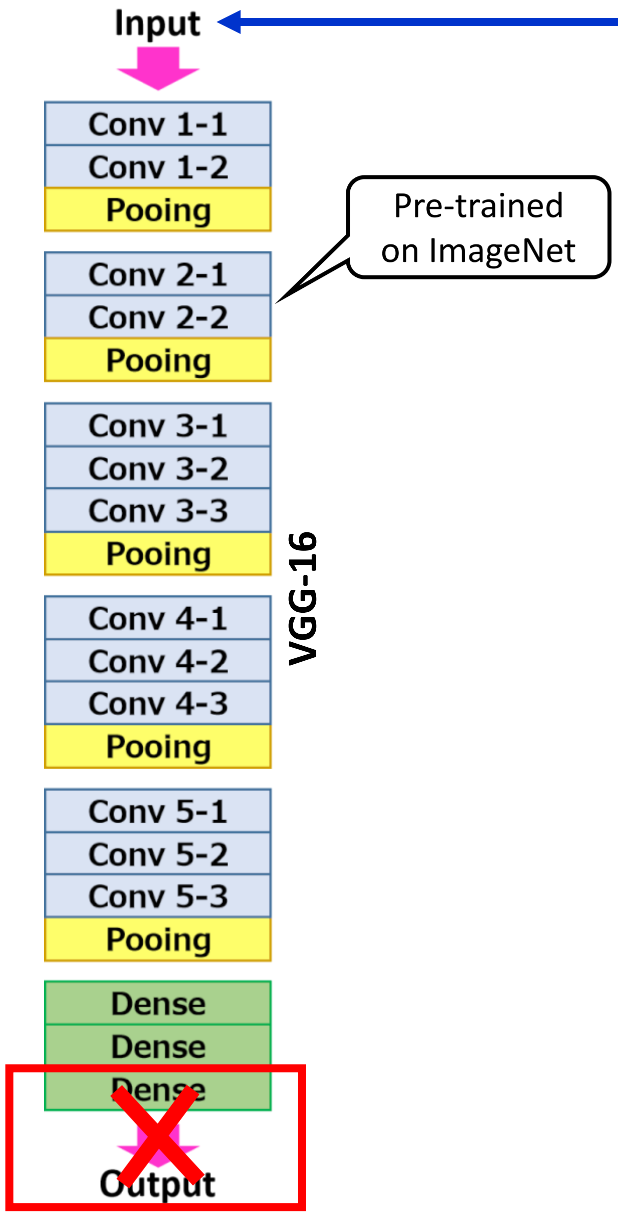
test image ‘I’

Flickr8K Image



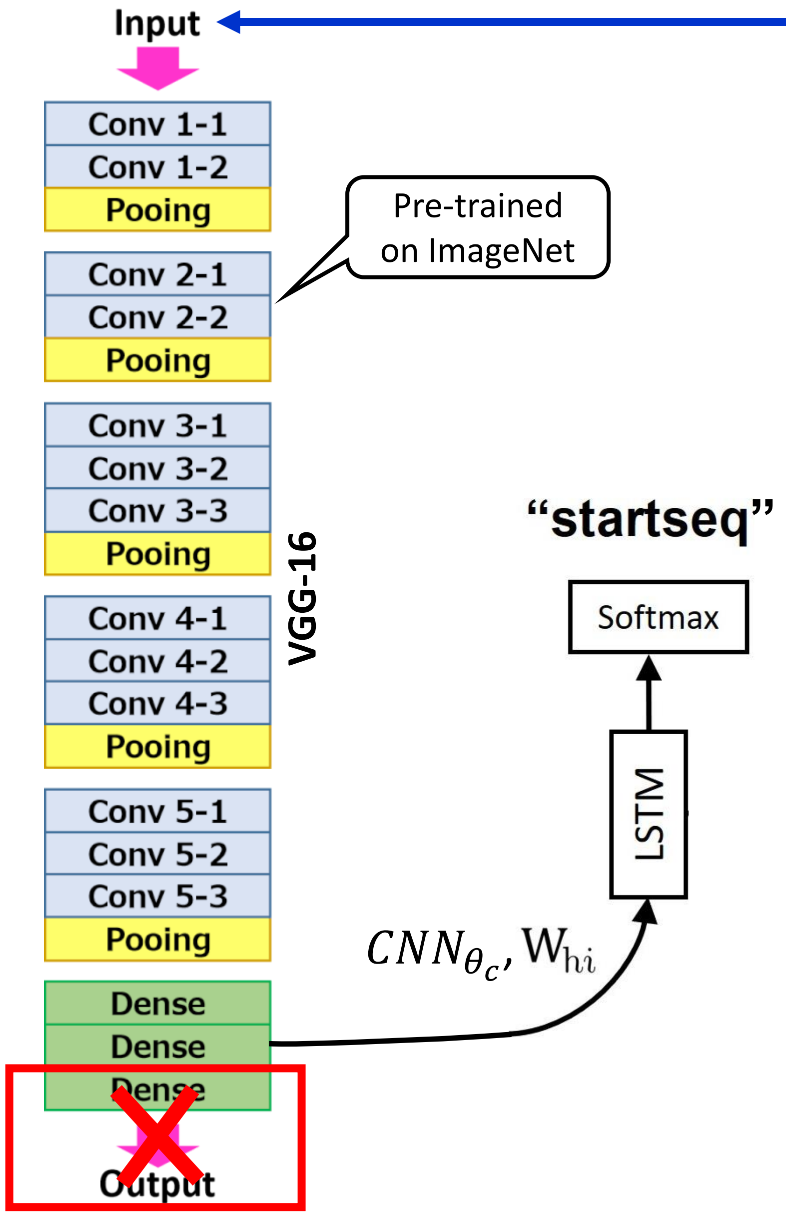
Flickr8K Image

test image ‘I’



Flickr8K Image

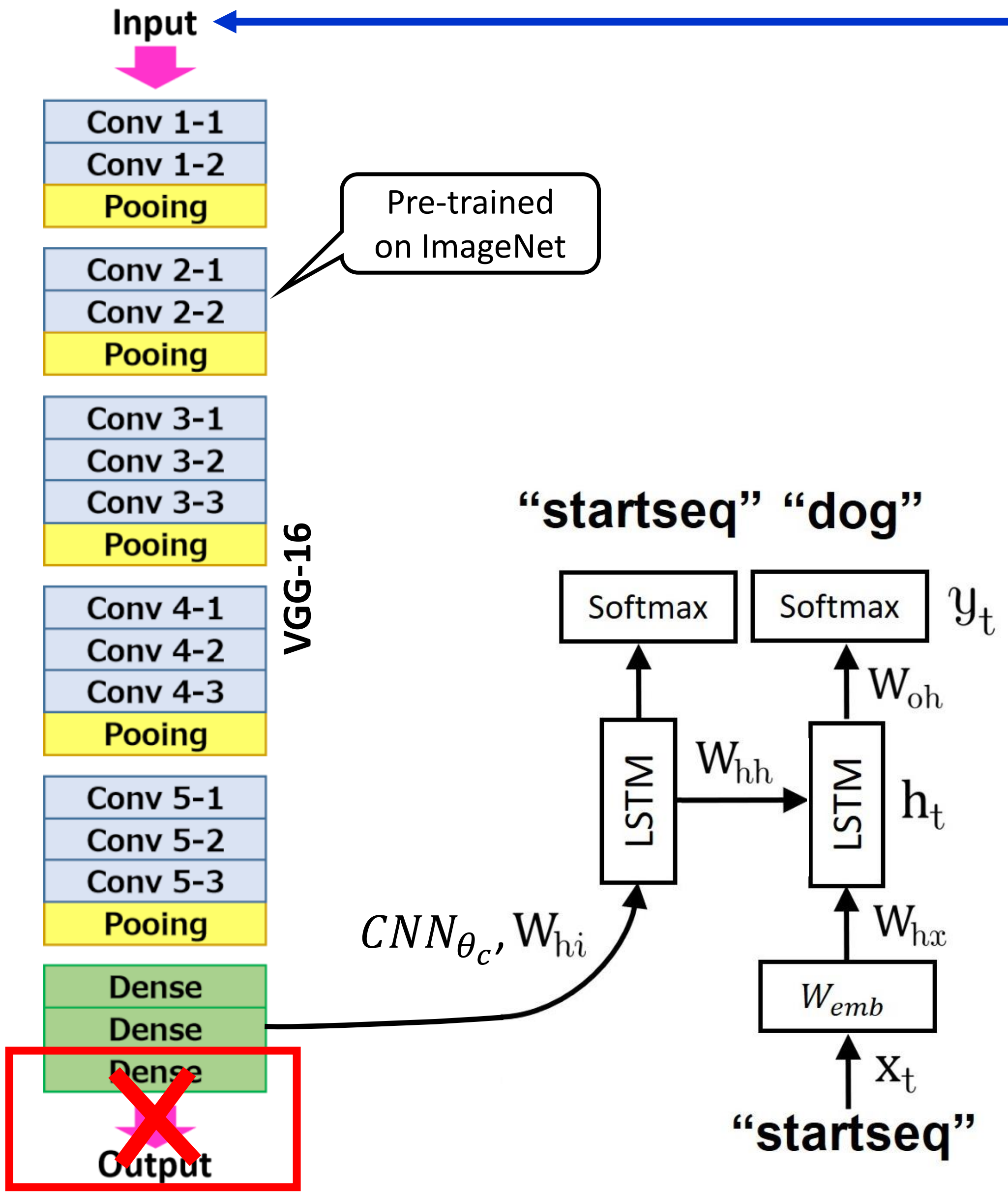
test image 'I'



Flickr8K Image

test image 'I'

LSTM takes image 'I' & sequence of vectors $(x_1, \ldots, x_T)$ as input



Flickr8K Image

The LSTM computes a sequence of hidden states $(h_1, \dots, h_t)$ & sequence of outputs $(y_1, \dots, y_t)$ by the recurrence relation for t = 1 to T [10]:

$$b_v = W_{hi}[CNN(I)] \tag{1}$$

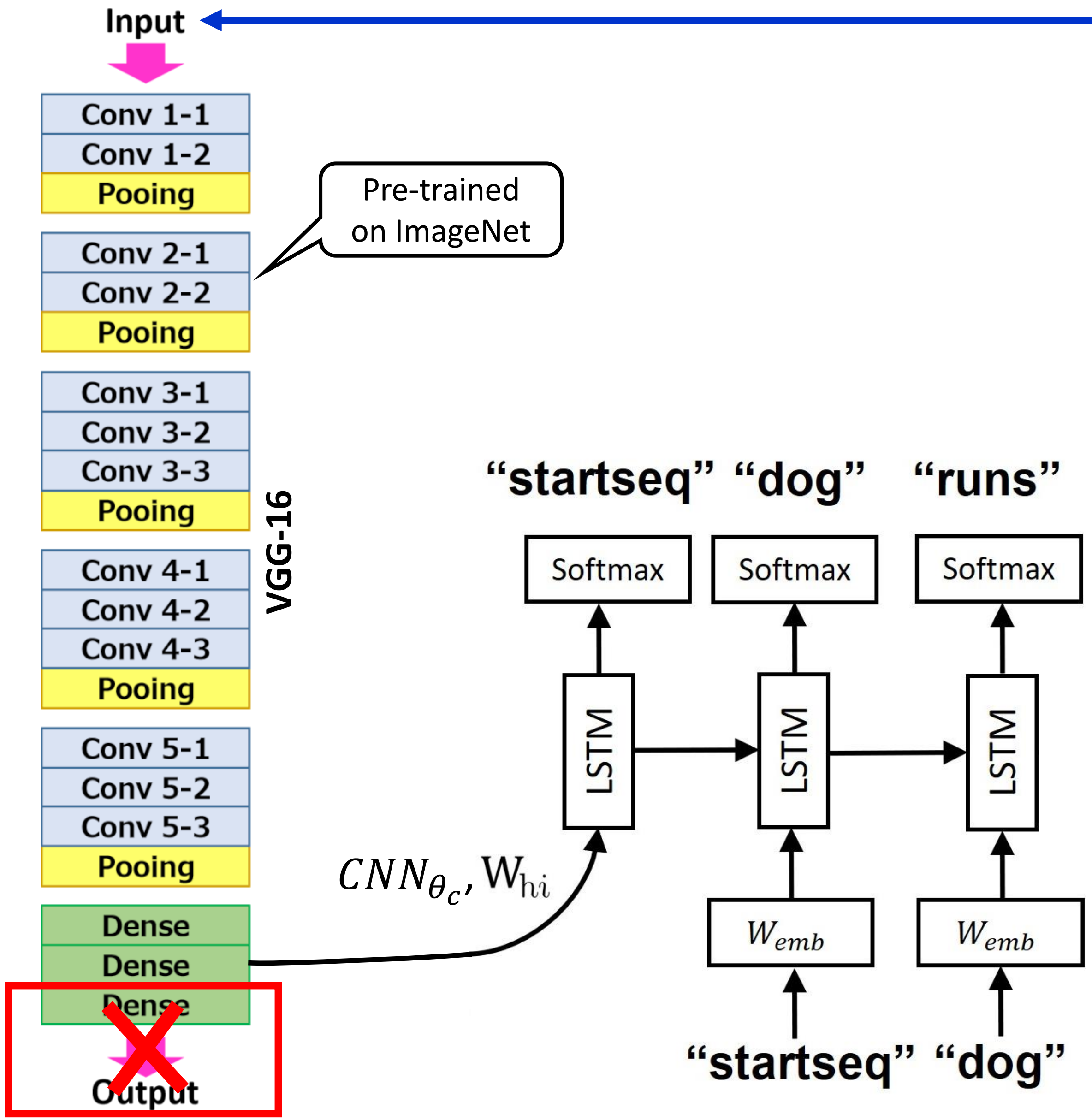
$$h_t = f(W_{hx}x_t + W_{hh}ht - 1 + b_h + 1(t = 1)\; o\; b_v) \tag{2}$$

$$y_t = Softmax(W_{oh}h_t + b_o) \tag{3}$$

where W_{hi} , W_{hx} , W_{oh} , x_i , b_h , and b_o are learnable parameters & $CNN(I)$ depicts the image feature vectors extracted by VGG-16. The LSTM model was trained to correctly predict the next word (y_t) based on current word (x_t), & previous context (h_{t-1}). Initially,

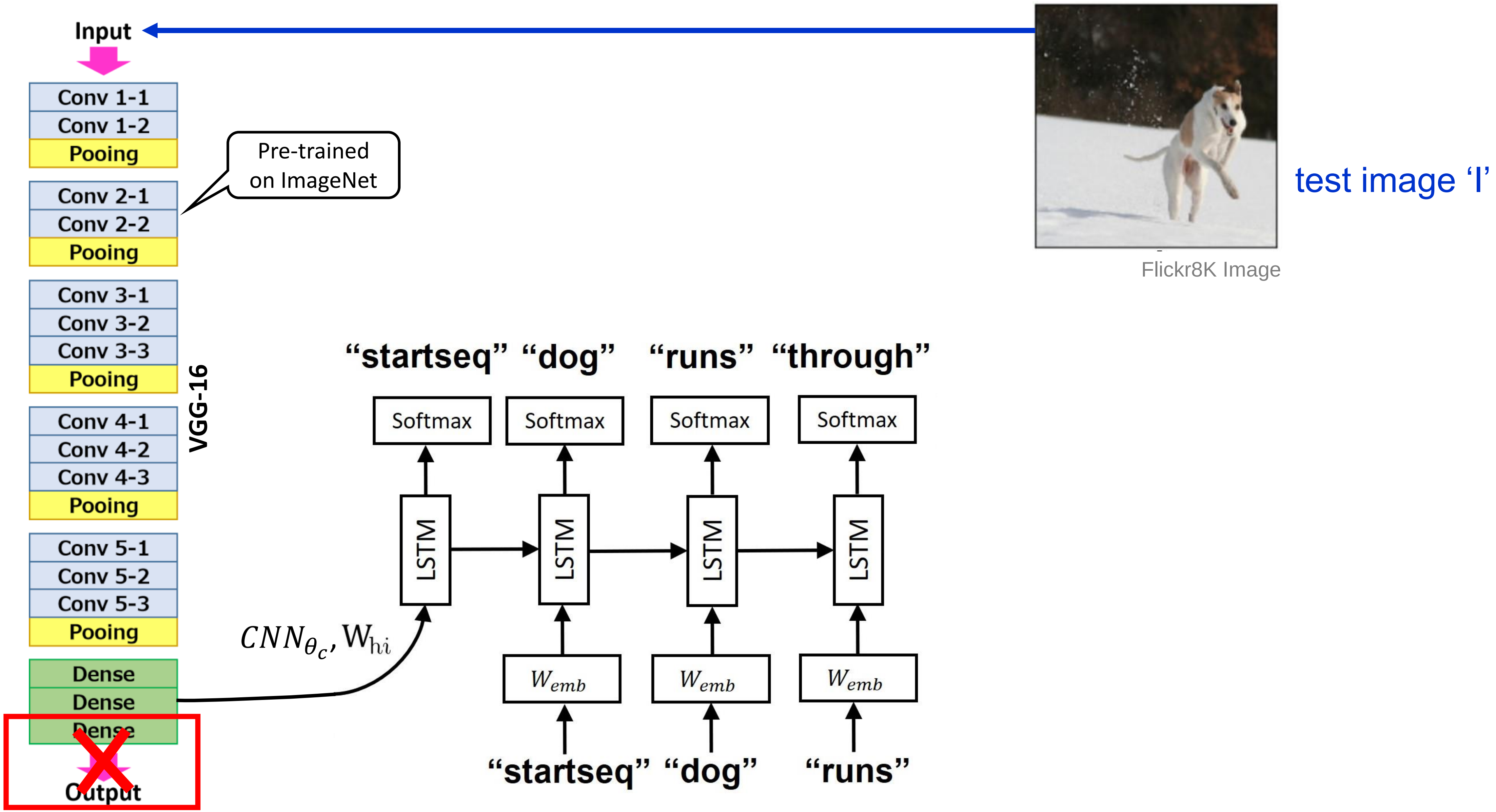
$$h_o = 0 \;\; and \;\; x_1 = < STARTSEQ > \tag{4}$$

& desired layer y_1 is set as the first word in sequence.

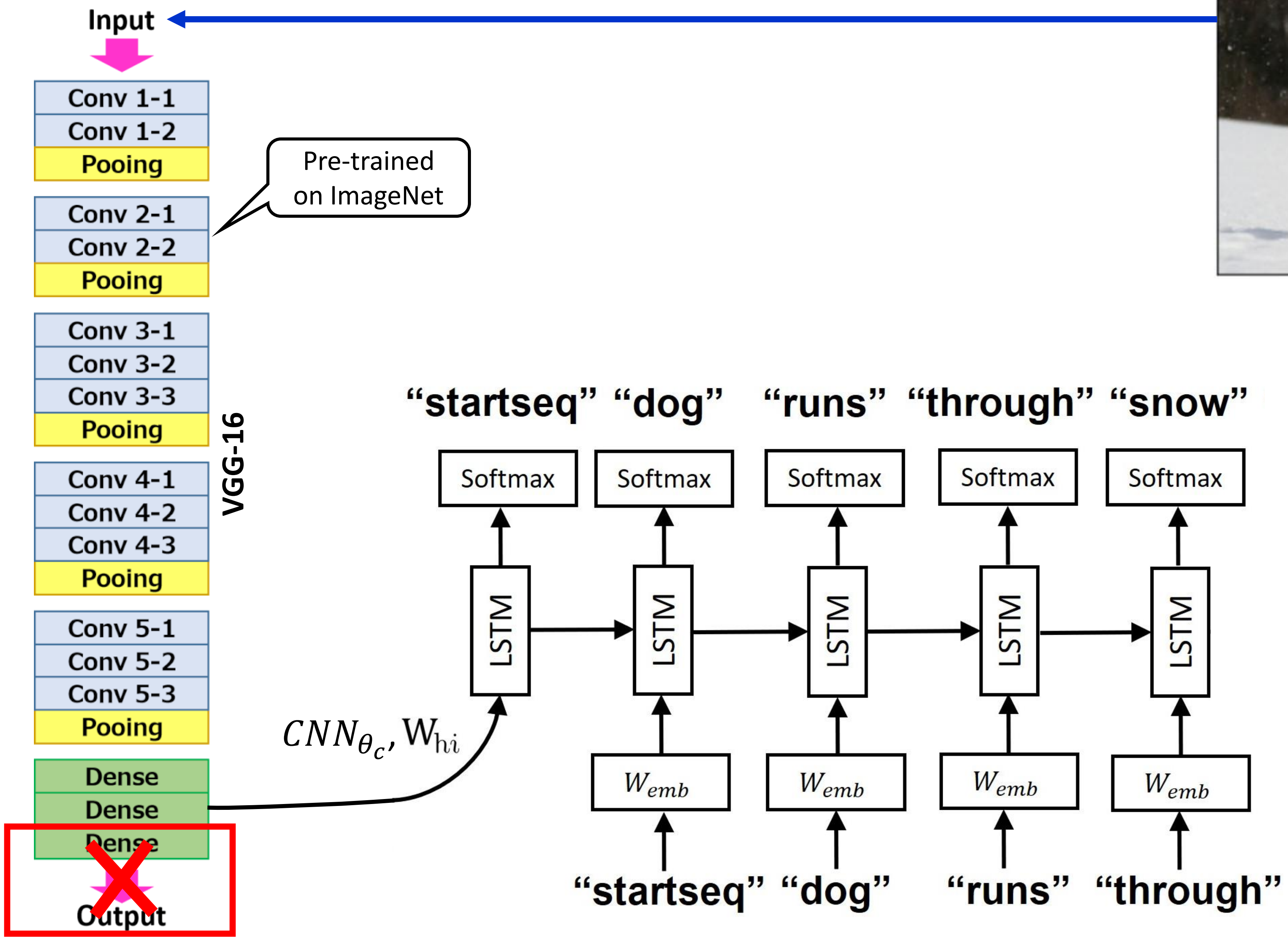


Flickr8K Image

test image ‘I’

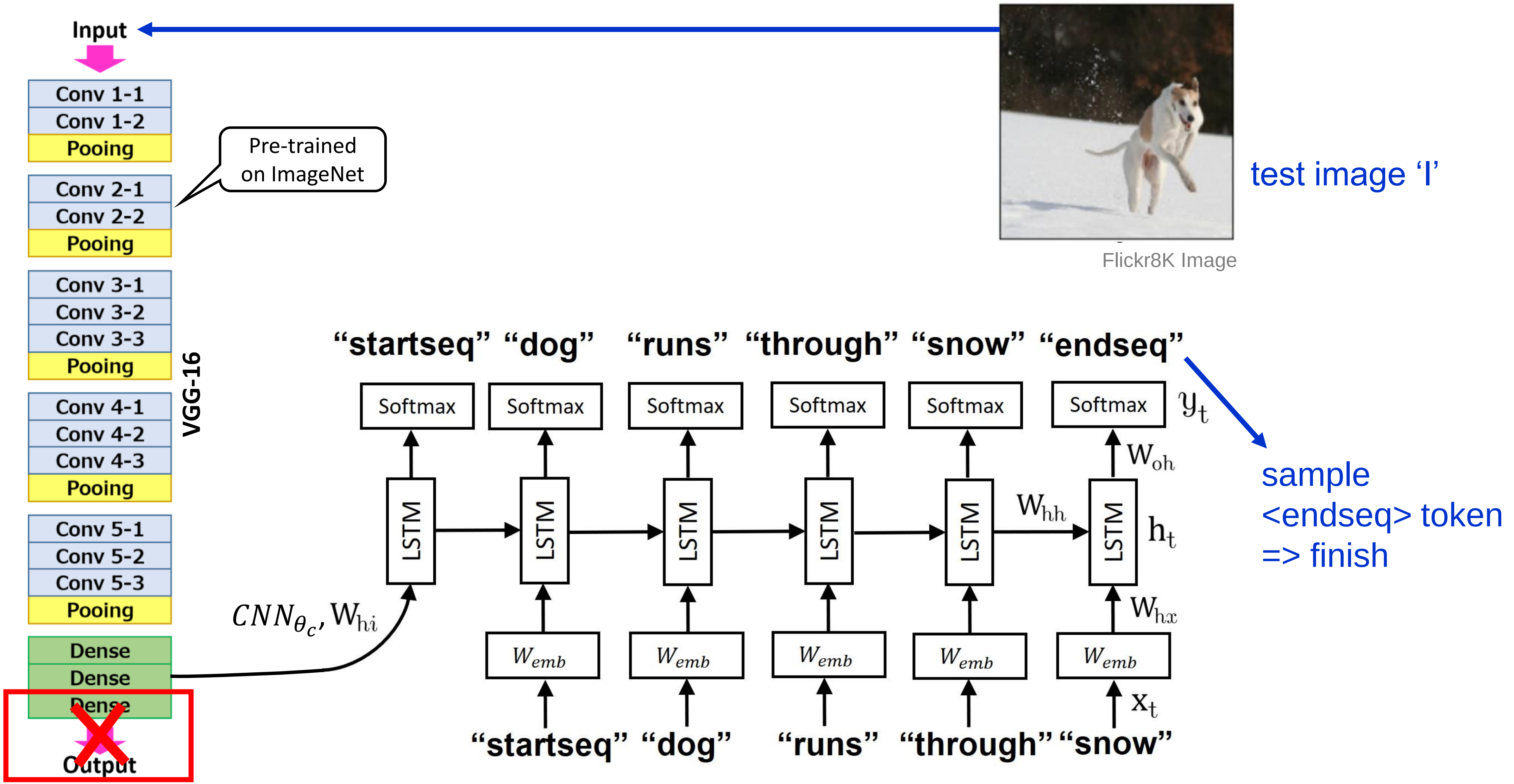


Flickr8K Image

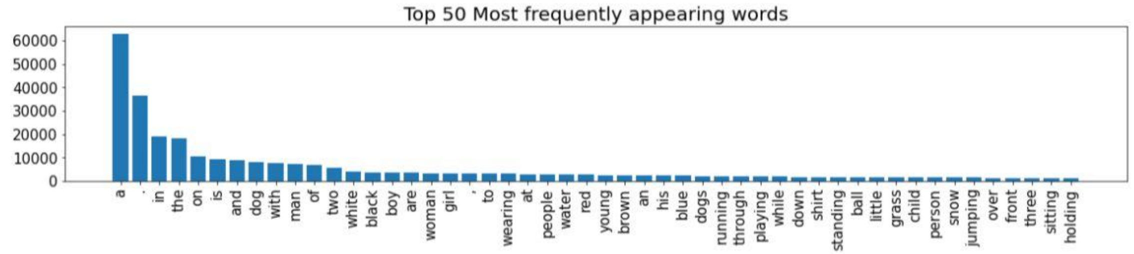


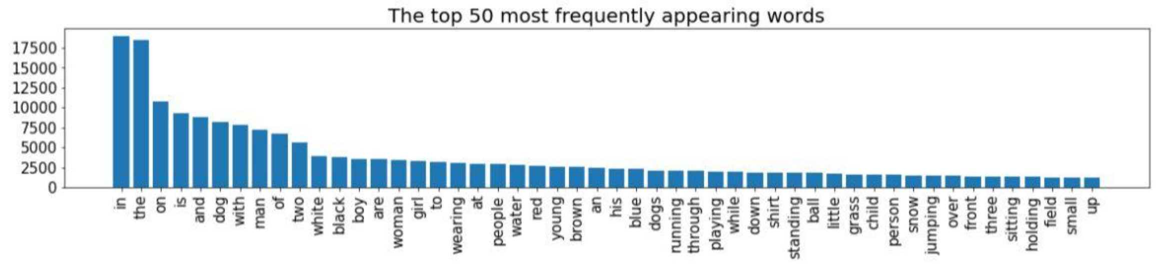
test image ‘I’

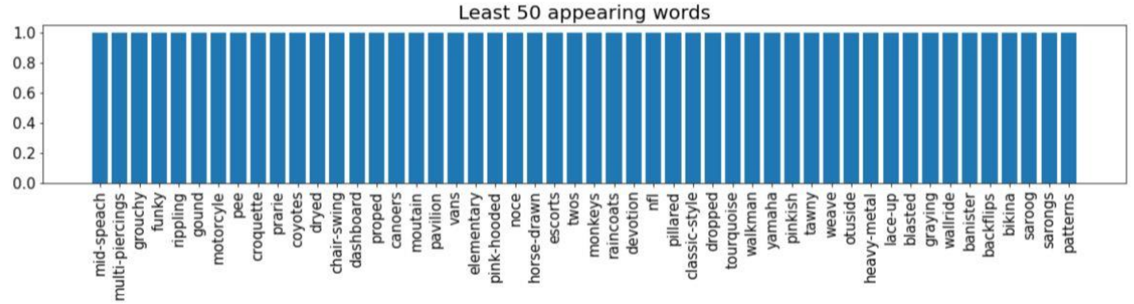
Flickr8K Image

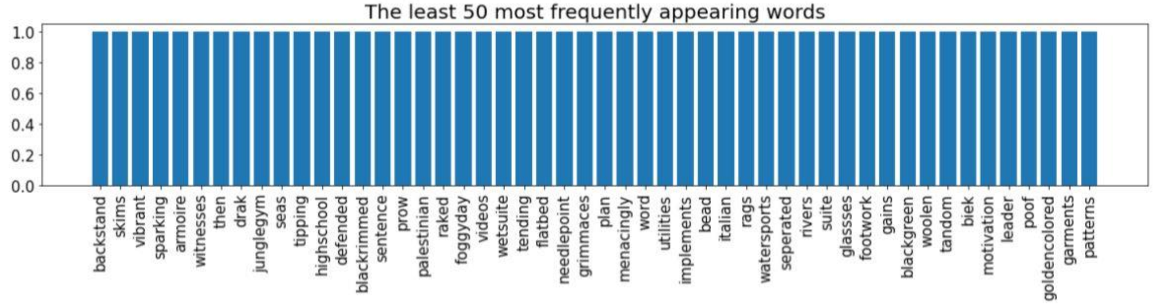

Input
Conv 1-1
Conv 1-2
Pooing
Conv 2-1
Conv 2-2
Pooing
Conv 3-1
Conv 3-2
Conv 3-3
Pooing
Conv 4-1
Conv 4-2
Conv 4-3
Pooing
Conv 5-1
Conv 5-2
Conv 5-3
Pooing
Dense
Dense
Dense
Output
VGG-16
Pre-trained on ImageNet
test image ‘I’
Flickr8K Image
“startseq” “dog” “runs” “through” “snow” “endseq”
Softmax
y_t
W_{oh}
LSTM
W_{hh}
h_t
W_{hx}
W_{emb}
x_t
CNN_{θ_c}, W_{hi}
“startseq” “dog” “runs” “through” “snow”
sample
<endseq> token
=> finish

Training Model on Flickr 8K Dataset



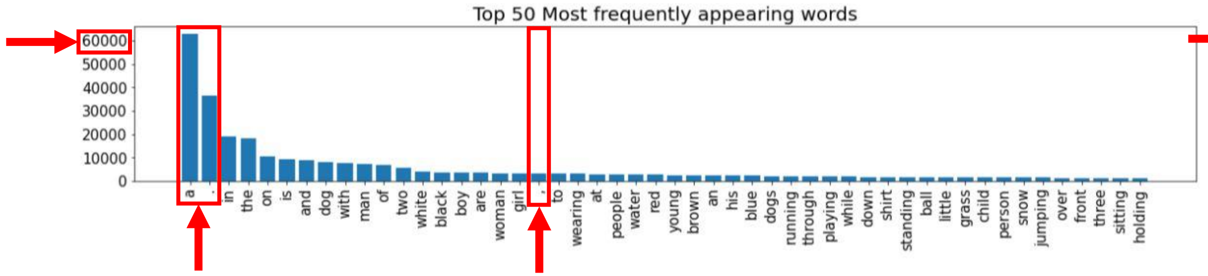


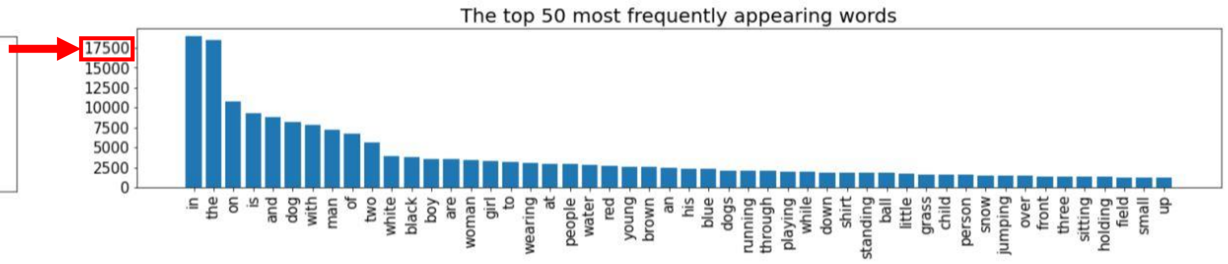


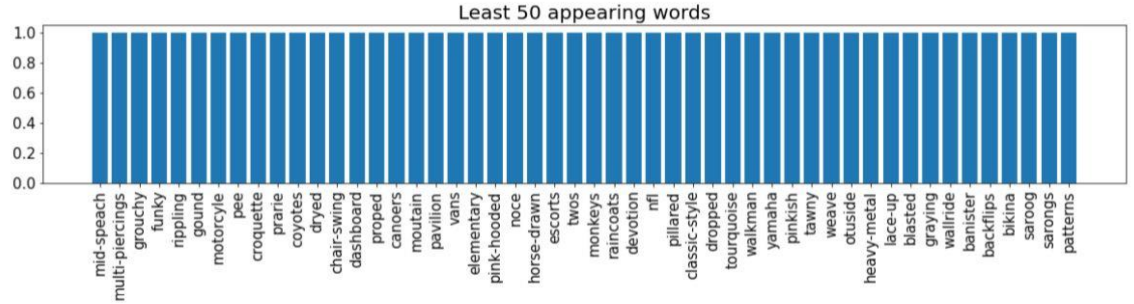


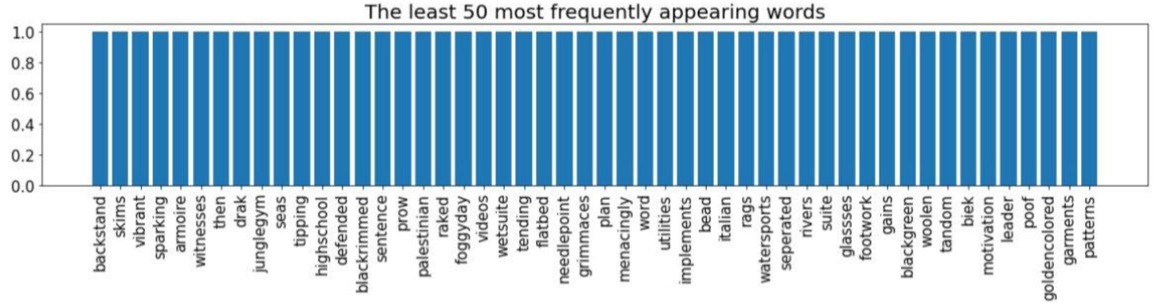
- The dataset is in the form [image->caption]
- Corpus of unique words is created & top 50 most frequently & least appearing words are analysed.
- ~8000 images with 5 captions each (= 6000 training images, 1000 validation images, 1000 test images)
- After data cleaning i.e., removing punctuation, single characters & numeric values which do not provide any relevant info useful in caption generation.
- Special tokens <STARTSEQ> & <ENDSEQ> are added to the beginning & end of each caption.

Training Model on Flickr 8K Dataset









- The dataset is in the form [image->caption]
- Corpus of unique words is created & top 50 most frequently & least appearing words are analysed.
- ~8000 images with 5 captions each (= 6000 training images, 1000 validation images, 1000 test images)
- After data cleaning i.e., removing punctuation, single characters & numeric values which do not provide any relevant info useful in caption generation.
- Special tokens <STARTSEQ> & <ENDSEQ> are added to the beginning & end of each caption.

Results & Discussion

1. Generation of Scene Descriptions in Real time-

Our neural scene description generator performs well giving rich descriptions on test set images. The model is successfully created by training on Flickr 8K dataset & run on test data on which it is found to generate rich meaningful captions.

True: man climbing rock wall

Predicted: man in harness is climbing from rock

BLEU: 0.8091067115702212

True: black and white dog is running through the field

Predicted: black and white dog is running through the grass

BLEU: 0.8633400213704505

True: lightcolored dog runs on the beach

Predicted: dog runs on the beach

BLEU: 0.8187307530779819

True: people stand inside rock dome

Predicted: group of people sit on rock

BLEU: 0.7598356856515925

Results & Discussion

1. **Generation of Scene Descriptions in Real time-**

Our neural scene description generator performs well giving rich descriptions on test set images. The model is successfully created by training on Flickr 8K dataset & run on test data on which it is found to generate rich meaningful captions.

Person Object Activity Surrounding

True: man climbing rock wall

Predicted: man in harness is climbing from rock

BLEU: 0.8091067115702212

Person Activity Surrounding

True: black and white dog is running through the field

Predicted: black and white dog is running through the grass

BLEU: 0.8633400213704505

True: lightcolored dog runs on the beach

Predicted: dog runs on the beach

BLEU: 0.8187307530779819

Person Activity Surrounding

True: people stand inside rock dome

Predicted: group of people sit on rock

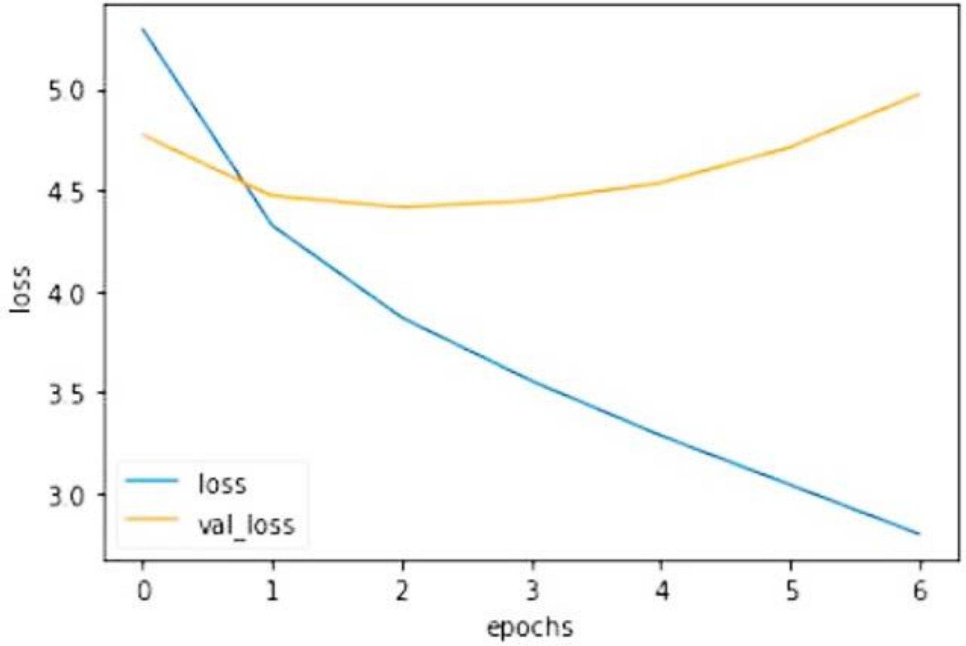
BLEU: 0.7598356856515925

Person Activity Surrounding

Results & Discussion

2. Training Loss & Validation Loss Curves -

The shape and dynamics of a learning curve is used to diagnose the behavior of a model and in turn suggest the type of configuration changes that may be made in the hyperparameters to improve the learning performance. From the learning curves obtained on training the model for 7 epochs on the Flickr 8K dataset, it can be inferred that the training data fits the validation set data easily.



Results & Discussion

3. BLEU Score as Evaluation Metrics-

Bilingual Evaluation Understudy [14] is used as the evaluation metrics. It is a measure of how descriptions generated by a model matches the 5 captions present in the dataset as true cases. BLEU n-gram precision p_n is defined as the sum of n-gram matches for every hypothesis sentence ‘S’ in the test corpus ‘C’ as shown below:

$$p_n = \frac{(\Sigma_{S \in C} \Sigma_{ngram \in S} Count_{matched}(ngram))}{(\Sigma_{S \in C} \Sigma_{ngram \in S} Count(ngram))} \quad (5)$$

The richness of an description generated depends on how high is the BLEU score in comparison with human evaluation. It has been found that the accuracy of our model closely compares with other similar models. (B-n represents the BLEU Score that uses upto n-grams.)

Flickr 8K [15]				
Model	B-1	B-2	B-3	B-4
Nearest Neighbour	-	-	-	-
Google NIC [8]	63	41	27	-
Karpathy et al. [10]	57.9	38.3	24.5	16.0
Our Model	53.1	28.8	16.4	8.8

Results & Discussion

4. **Audio Generation to Aid the Visually Impaired-**

- The description generated is converted into audio signals & played back to the visually impaired person. It was found to be coherent.
- It was also found that the system can work on real-time basis & read out the scene descriptions for the benefit of a visually impaired person.
- Although the model is capable of extracting large number of pictures on real time basis and generating captions, the recitation speed of the captions to the visually impaired person limits the use of all the generated captions on each image frame.
- The recitation speed depends on many factors including choice of natural language, age, culture, etc., & found to be varying from person to person.




Results & Discussion

5. Repetition

- Furthermore, it is noticed that in case the scene around the visually impaired person is not changing fast enough like when sitting on a park bench or roadside with people walking down the street, resulting descriptions repeat frequently.

True: little girl covered in paint sits in front of painted rainbow with her hands in bowl

Predicted: group of children are sitting on the snow

BLEU: 0.21874242445215208

True: boy smiles in front of stony wall in city

Predicted: two men are sitting on the street

BLEU: 0

- Some images which are not accurately captioned (where BLEU score < 0.25) are also encountered.

Conclusion

An end-to-end human-centric model has been implemented based on deep recurrent architecture which generates scene descriptions using images from real-time videos captured through the vision enabled eye wear of a visually impaired person. These captions are then recited to the person on real time basis which provides him with a better sense of understanding about the environment around him.

Future Work-

Customized Model for User: training the model on a dataset created out of the videos captured by his/her eye wear over a period of time. Expected to increase the accuracy further dynamically.

Repetitive descriptions: due to scene not changing fast enough. Can be handled by incorporating a small feedback-mechanism which compares every newly generated caption with the earlier one & produce a standard music/ silence in case caption is repetitive by a pre-determined frequency. This will eliminate irritation due to repetitive recitations.

Further Challenge: To study & give online feedback on human emotions during face to face interactions.

References

[1] "World Report on Vision", World Health Organization, 2019.
[2] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in ICML-14, pp. 595–603, 2014
[3] J. L. Elman, "Finding structure in time," Cognitive Science, vol. 14, no. 2, pp. 179–211, 1990.
[4] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, Every picture tells a story: Generating sentences from images. in Proc. European Conf. Computer Vision, 2010.
[5] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. "Babytalk: Understanding and generating simple image descriptions," in Proc. Conf. Computer Vision and Pattern Recognition, pp. 1601–1608, 2011.
[6] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in Proc. 15th Conf. Computational Natural Language Learning, pp. 220–228, 2011.
[7] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," J. Artif. Intell. Res., vol. 47, pp. 853–899, 2013.
[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555, 2014.
[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv preprint arXiv:1502.03044, 2015.
[10] Andrej Karpathy, Li Fei-Fei.Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3128–3137.A, 2015.
[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.
[12] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp. 311–318, 2002.
[15] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 139–147, 2010.

Deep Recurrent Architecture based Scene Description Generator for Visually Impaired

Aviral Chharia, Rahul Upadhyay

achharia_be18@thapar.edu *rahul.upadhyay@thapar.edu*

<\thankyou

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

12 ICUMT