Visual Question Answering
CS626 - Project Report

Raaghav Raaj, Vrinda Jindal, Yash Gupta

Department of Computer Science and Engineering,
Indian Institute of Technology Bombay, Mumbai, India
{180050082, 180050120, 180050121}@iitb.ac.in

Abstract. In this report, we present our project on Visual Question Answering. Visual Question Answering refers to the idea of taking an image and a natural language question related to it as an input and returning a natural language answer to the question. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. VQA is a prominent multi-modal research problem. To correctly answer visual questions about an image, the machine needs to understand both the image and question. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions.

Keywords: Visual Question Answering · VQA MSCOCO · LSTM-CNN Model · BiLSTM-CNN Model · Vis-BiLSTM

1 Introduction

The complex compositional structure of language makes problems at the intersection of vision and language challenging. In this project, we propose various models in the direction of solving the problem of visual question answering.
Due to the often excellent context provided language priors can give a false impression that machines does well when answering questions when they are only exploiting language priors to achieve high accuracy. We also try to establish the opposite by doing ablation studies through language only and vision only models.

2 Dataset

We deploy the VQA-COCO dataset for our model learning. It is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.
The questions generally seek very specific information, so answers are one word - numbers, yes/no. Also, many questions require common sense knowledge.

	Training	**Validation**
Images	82459	40504
Questions	215359	121512
Question Max Length	26	26

Fig. 1. Is there any animal in the image?(yes); What is the color of the bird?(yellow); How many birds are there?(1); Does the bird have feathers?(Yes); Is the bird alive?(Yes

3 Models

3.1 LSTM-CNN Model

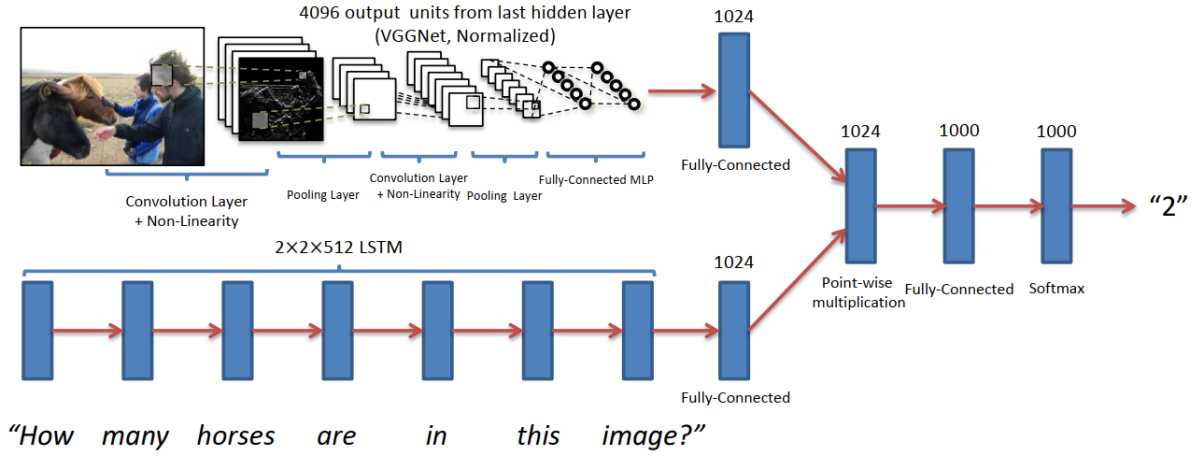

Fig. 2. LSTM-CNN Model (Image from reference)

- Image Channel: This channel provides an embedding for the image. These are 2-normalized activations from the last hidden layer of VGGNet19 (19 layers). Each of them is of size (1, 4096) per image.
- Question Channel: This channel provides an embedding for the question. The input vocabulary to the embedding layer consists of all the question words seen in the training dataset. An LSTM with two hidden layers is used to obtain 2048-dim embedding for the question. The embedding obtained from the LSTM is a concatenation of last cell state and last hidden state representations (each being 512-dim) from each of the two hidden layers of the LSTM. Hence 2 (hidden layers) x 2 (cell state and hidden state) x 512 (dimensionality of each of the cell states, as well as hidden states). This is followed by a fully-connected layer + tanh nonlinearity to transform 2048-dim embedding to 1024-dim.

Model Size - 16M parameters
We tried both multiplication and concatenation at the junction layer. (multiplication was selected)
We tried running with both with and without image normalisation. (normalisation selected)

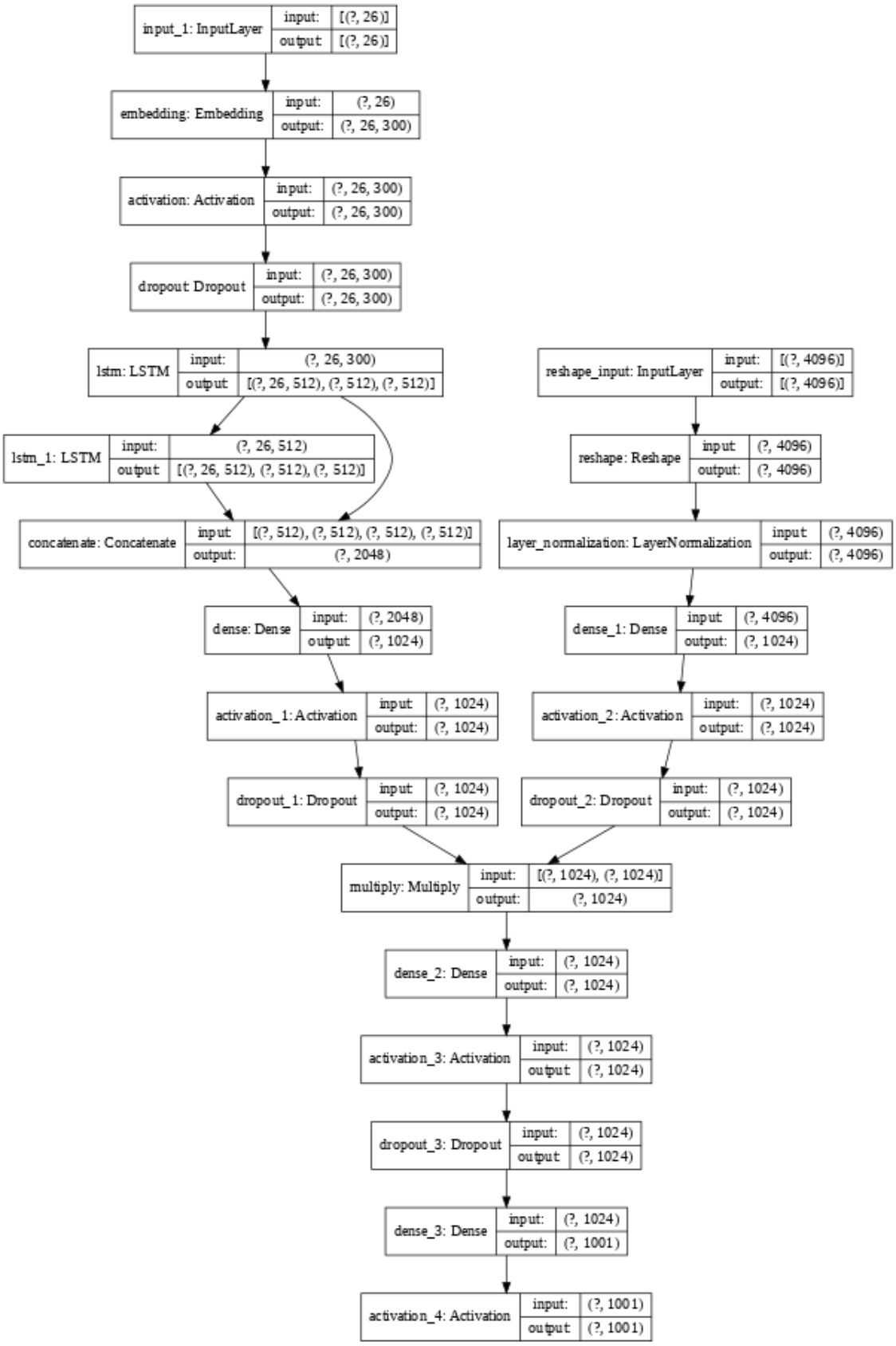

Fig. 3. LSTM-CNN Model

3.2 BiLSTM-CNN Model

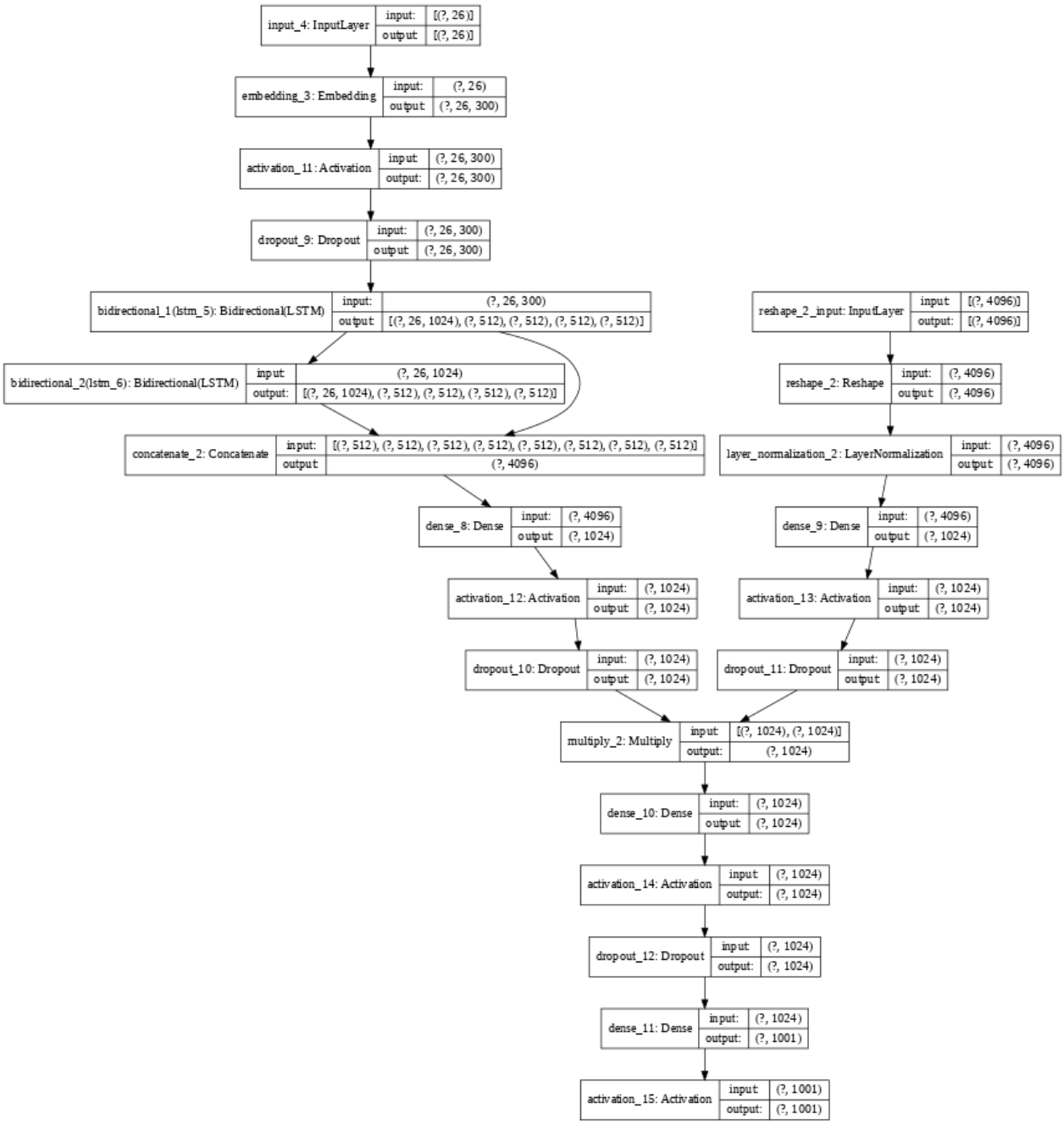

Fig. 4. BiLSTM-CNN Model

We replace LSTM layer with a BiLSTM layer. Now on concatenation of all hidden cell states and cell states we get a language output of 2 x 2 x 2 x 512 = 4096 features.
Model Size - 23M parameters

3.3 Vis-BiLSTM Model

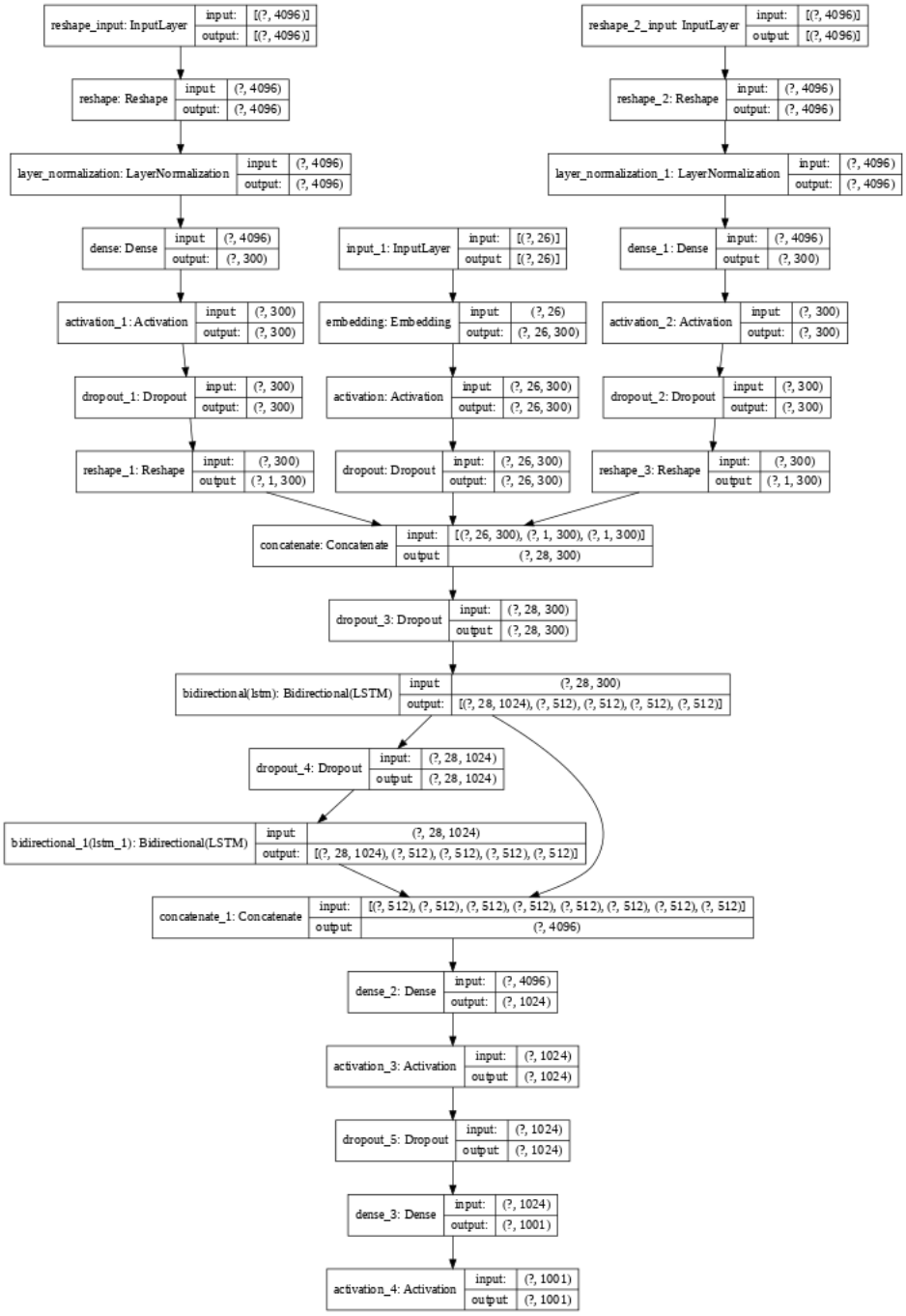

Fig. 5. Vis-BiLSTM Model

This model builds directly on top of the LSTM sentence model and is called the "VIS+LSTM" model. It treats the image as one word of the question.

Model Size - 21M parameters

4 Results

We perform open ended task of VQA. (number of classes = 1000).

4.1 Accuracies

Models	Validation Accuracy
LSTM-CNN	46.88%
BiLSTM-CNN	48%
Vis-BiLSTM	45.02%

4.2 Observations

- Accuracies are somewhat low. Some misclassifications shown. Eg Question ID 2342392, 1542221
- Questions that started with "Is there" or "Are" or "Does" and had yes/no as answer were answered correctly more often (55%). Eg Question ID 1082720, 1339952
- Questions that had very specific answers and required both common-sense knowledge and good image analysis showed poorer results. Eg Question ID 1104490, 868400

Image	Question	Top-5 Answers with Confidence
	Question ID: 730932 what sport is this ?	99.87 % baseball 000.1 % soccer 00.01 % playing baseball 000.0 % tennis 000.0 % football
	Question ID: 3699971 how many bikes in the image ?	40.74 % 2 21.74 % 1 14.28 % 3 08.61 % 4 04.62 % 5
	Question ID: 1082720 is there a cop in the image ?	93.09 % yes 06.91 % no 000.0 % background 000.0 % nowhere 000.0 % happy
	Question ID: 1339951 is this picture taken outside or inside ?	68.43 % outside 31.34 % inside 00.13 % zoo 00.05 % beach 00.01 % airport
	Question ID: 1104490 what caused the red eyes on the guy in the green shirt ?	69.92 % coffee 04.46 % candle 02.38 % soda 002.1 % skull 001.9 % wine
	Question ID: 868400 what cartoon character does this hydrant most remind you of ?	100.0 % hello kitty 000.0 % toy 000.0 % cross 000.0 % statue 000.0 % baby

4.3 Ablation Study on LSTM-CNN Model

Clearly, both models worse than the original. This is quite expected since all information needed for giving correct answer. Also, language-alone performs better than vision-alone possibly due to exploiting

	Language-Alone Model	Vision-Alone Model
Test-Dev Accuracy	24%	23%

subtle statistical priors about the question types (e.g. “What color is the banana?” can be answered with “yellow” without looking at the image).

5 Conclusion

In this paper, we consider the image QA problem and present our end-to-end neural network models. Our model shows a reasonable understanding of the question and some coarse image understanding, but it is still very naiıve in many situations. Image question answering is a fairly new research topic, and the approach we present here has a number of limitations. First, our models are just answer classifiers. Ideally we would like to permit longer answers which will involve some sophisticated text generation model or structured output. But this will require an automatic free-form answer evaluation metric. Second, we are only focusing on a limited domain of questions. However, this limited range of questions allow us to study the results more in depth. Lastly, it is also hard to interpret why the models output a certain answer.

6 Challenges

- Dataset Size: the zip file for dataset alone is 12 GB. Due to limited free cloud storage available, everything had to be stored on different accounts. Due to this, a custom data-generator was coded.
- Limited Compute: We tried increasing model sizes but they take too much time for one epoch and frequently go out of RAM.
- Word Embeddings: Rather than training a separate embeddings layer, we can use pre trained embeddings. We tried using spacy, but the data loader is crashing as of now.
- Overfitting: Most of the models display some kind of overfitting or the other.

7 Future Possibilities

- Accuracies: slightly on the lower side. Lot of scope of hyperparameter tuning and then training on more epochs. Some of it maybe attributed to different ways of calculating (our just matches the answer with the most common one).
- Glove Embeddings can be used for encoding rather than training a separate embeddings layer.
- Attention: several ways to incorporate attention can be tried.

References

1. Visual Question Answering `https://arxiv.org/pdf/1505.00468.pdf`
2. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering `https://arxiv.org/pdf/1612.00837.pdf`
3. VQA - Deep Learning `https://medium.com/@anuj_shah/visual-question-answering-2350eea072df`
4. VQA - Deep Learning `https://towardsdatascience.com/deep-learning-and-visual-question-answering-c8c8093941bc`

8 Appendix - LSTM Thereotical Details

- Activation Gates - sigmoid activation (red), tanh activation (blue)
- Forget Gate - decides which info gets thrown/kept, takes addition of hidden and input state, outputs number between 0/1 (forget/keep)

- Input gate - decides which current info is important, (0/1) (important/ not important), multiplied with regularized info to get important info
- Cell State - memory of the network, info gets added/removed with help of gates, updates by multiplying with forget gate (forget useless info) and adding by input gate (add useful info)
- Output Gate - takes in hidden, output and cell state and outputs hidden state (for carry over as well as decoding)

Using Bidirectional LSTMs, you feed the learning algorithm with the original data once from beginning to the end and once from end to beginning.