

Image Captioning Using CNN & LSTM



Uday Kamal

Id : 1406041



Md Hasibul Amin

Id : 1406045



Rajib Al-Sabah

Id : 1406035



Abhishek Shushil

Id : 1406034



Presentation Outline

- Problem Statement
- Basic building blocks for the network
 - CNN
 - Transfer Learning
 - RNN
 - LSTM
- How do we wire them together?
- Code
- Other places this can be implemented
- Interaction & Questions

Problem Overview

Can a computer understand these pictures?



A yellow bus driving down a road with green trees and green grass in the background.



Living room with white couch and blue carpeting. The room in the apartment gets some afternoon sun.

Problem Overview

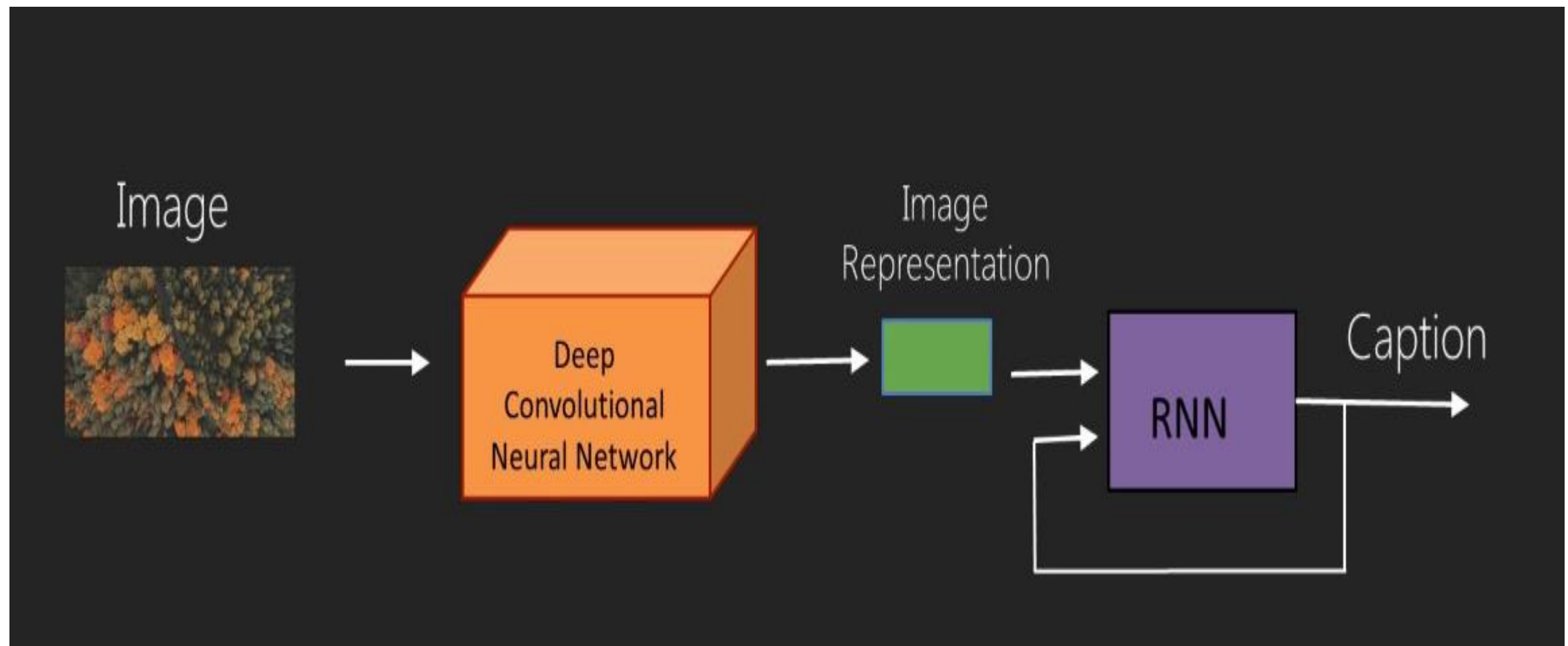
Human captions from the training set



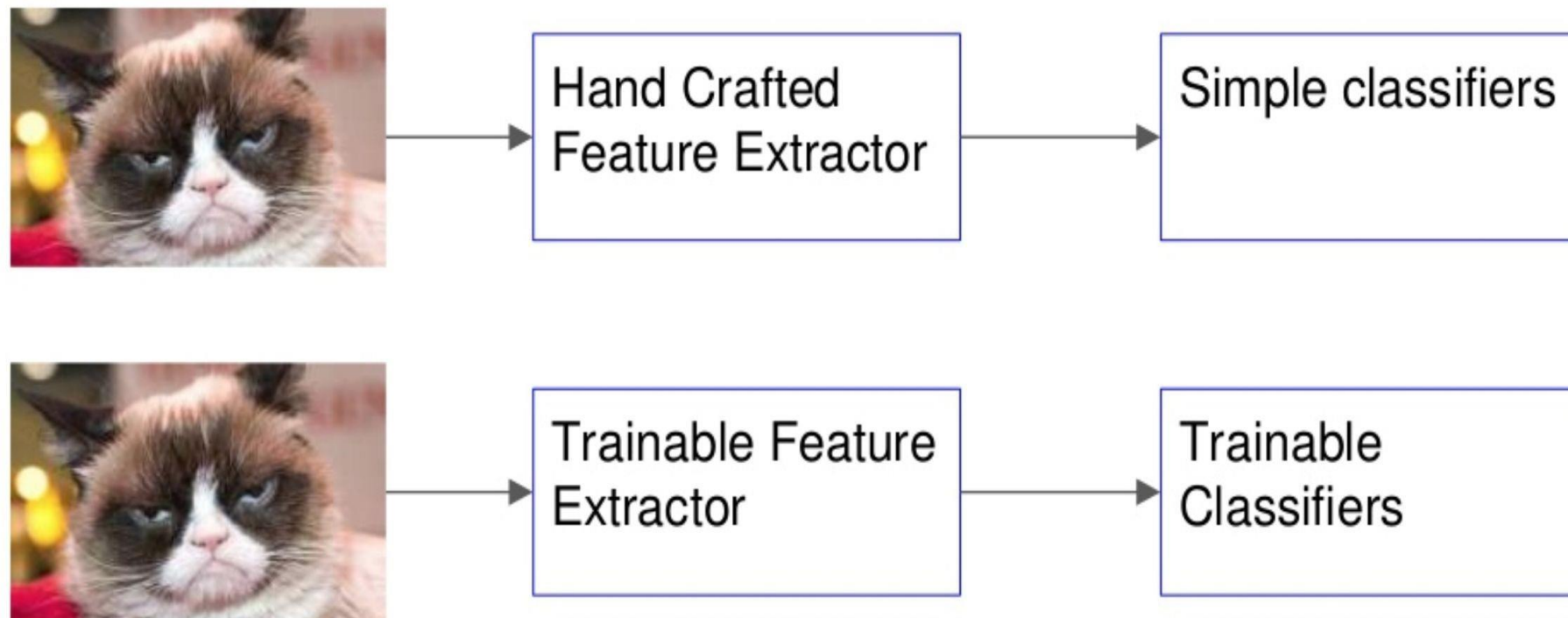
Automatically captioned



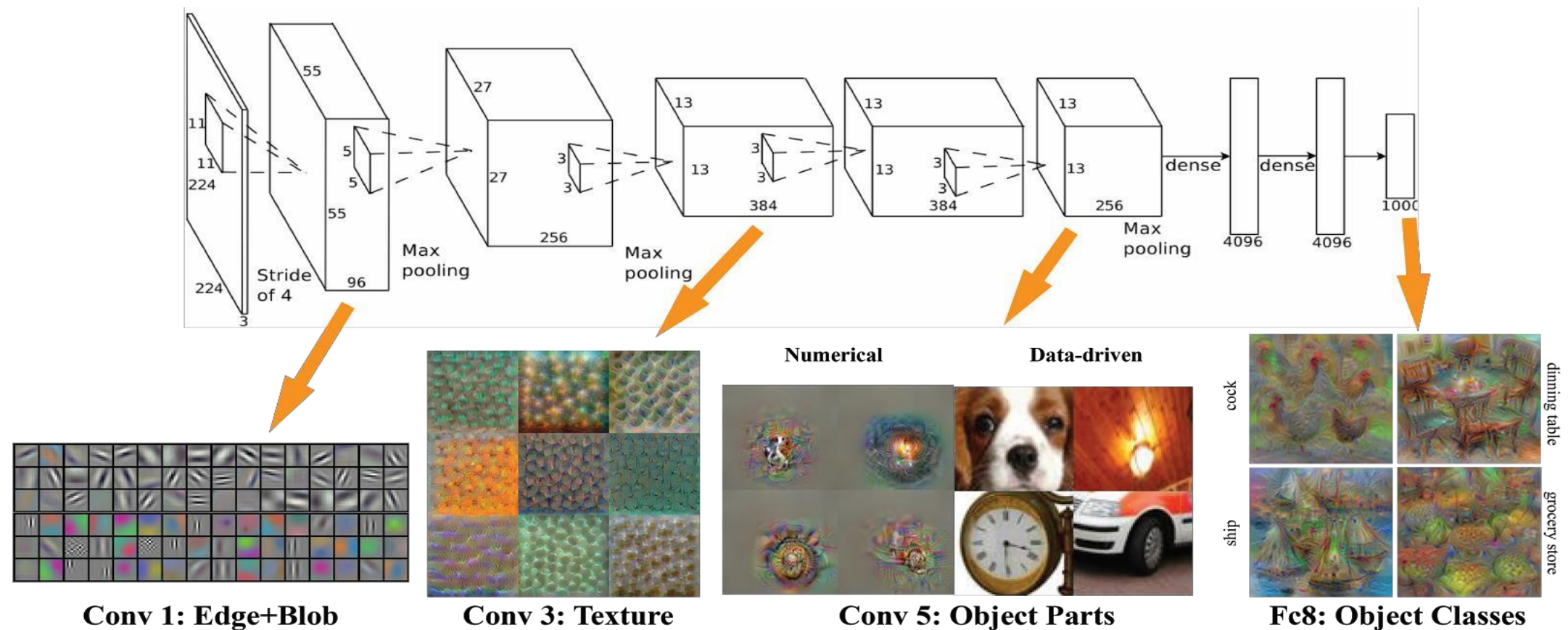
Overall Model:



Building Blocks for the Network: CNN



Building Blocks for the Network: CNN

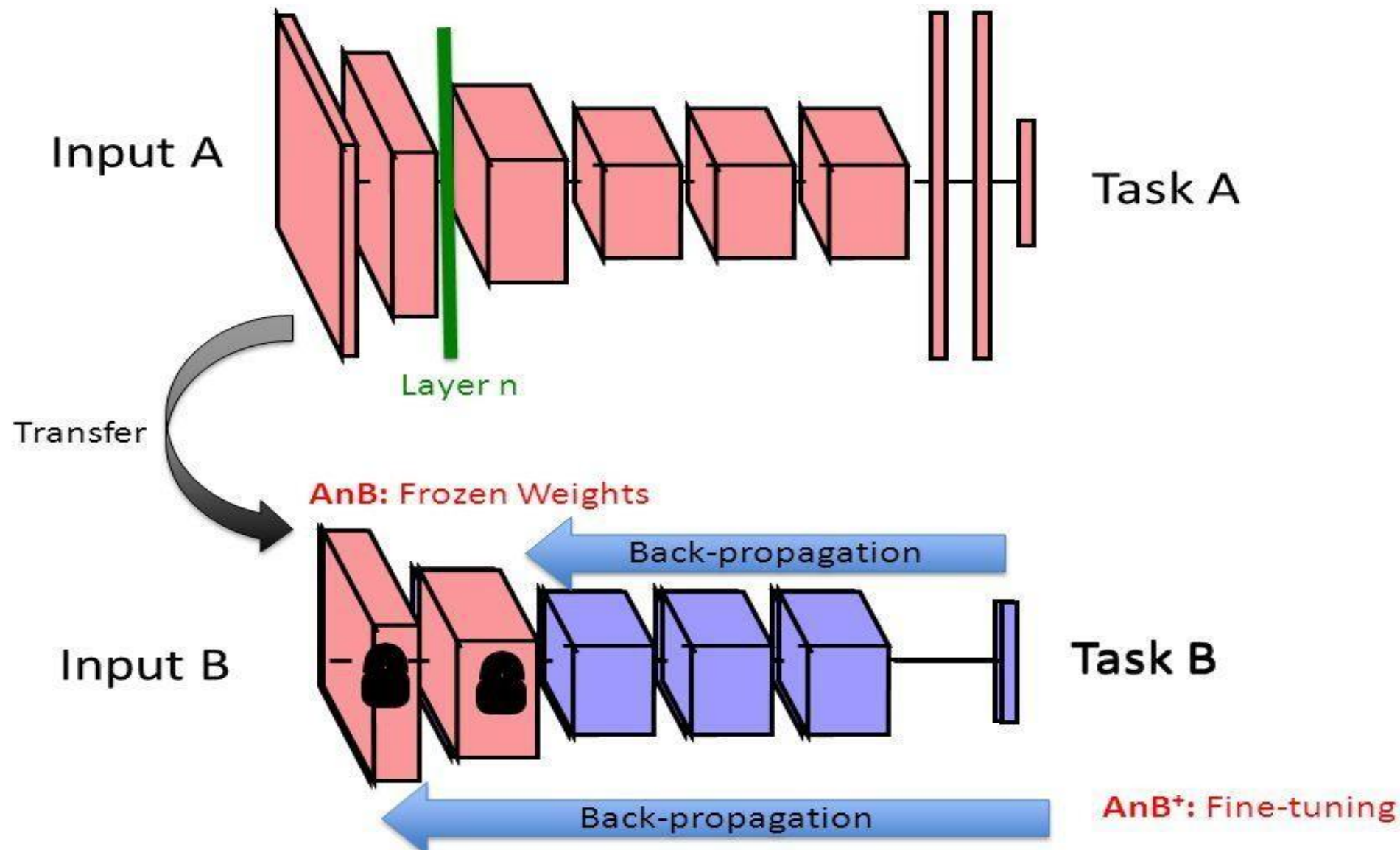


Convolution layer is a feature detector that automatically learns to **filter out the not needed information** from an input by using convolution kernel.

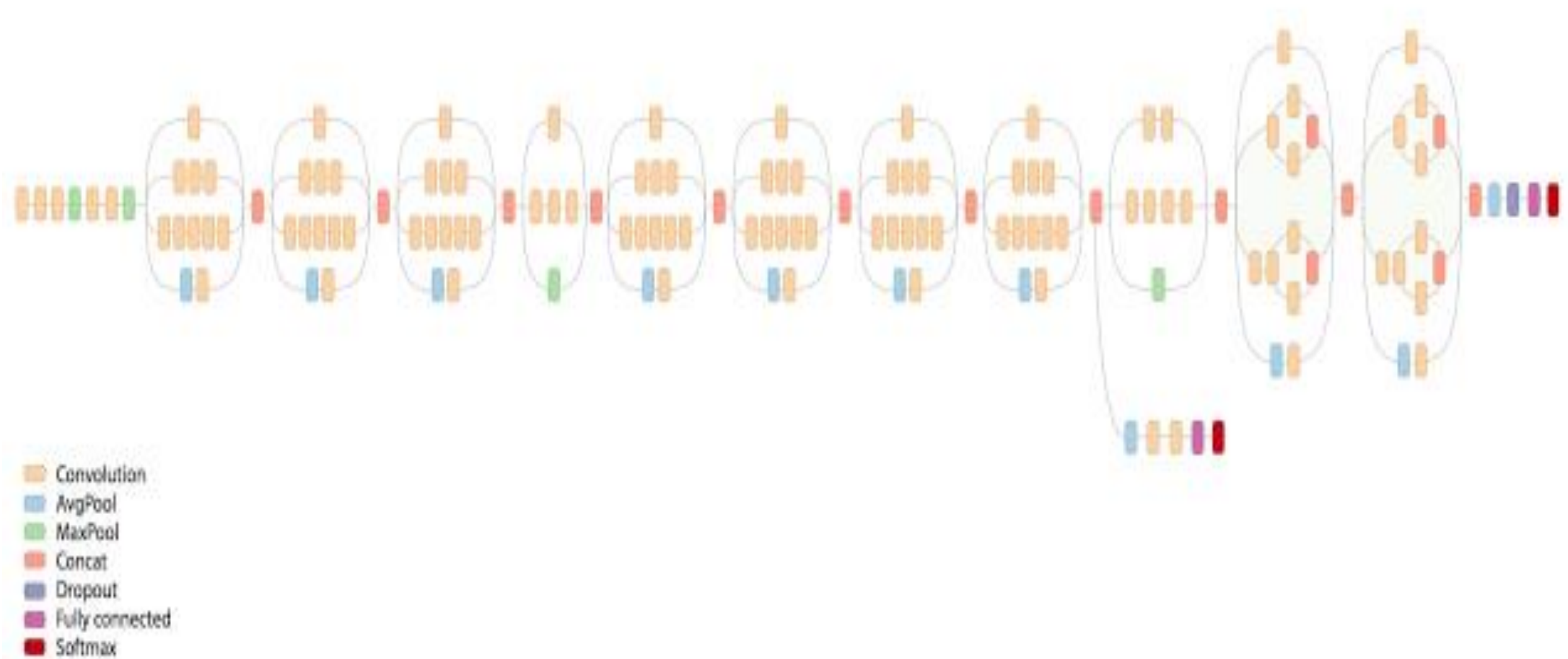
Pooling layers compute the max or **average value of a particular feature over a region** of the input data (*downsizing of input images*). Also helps to detect objects in some unusual places and reduces memory size.

Building Blocks for the Network: Transfer Learning

Transfer Learning Overview



Building Blocks for the Network: Inception V3



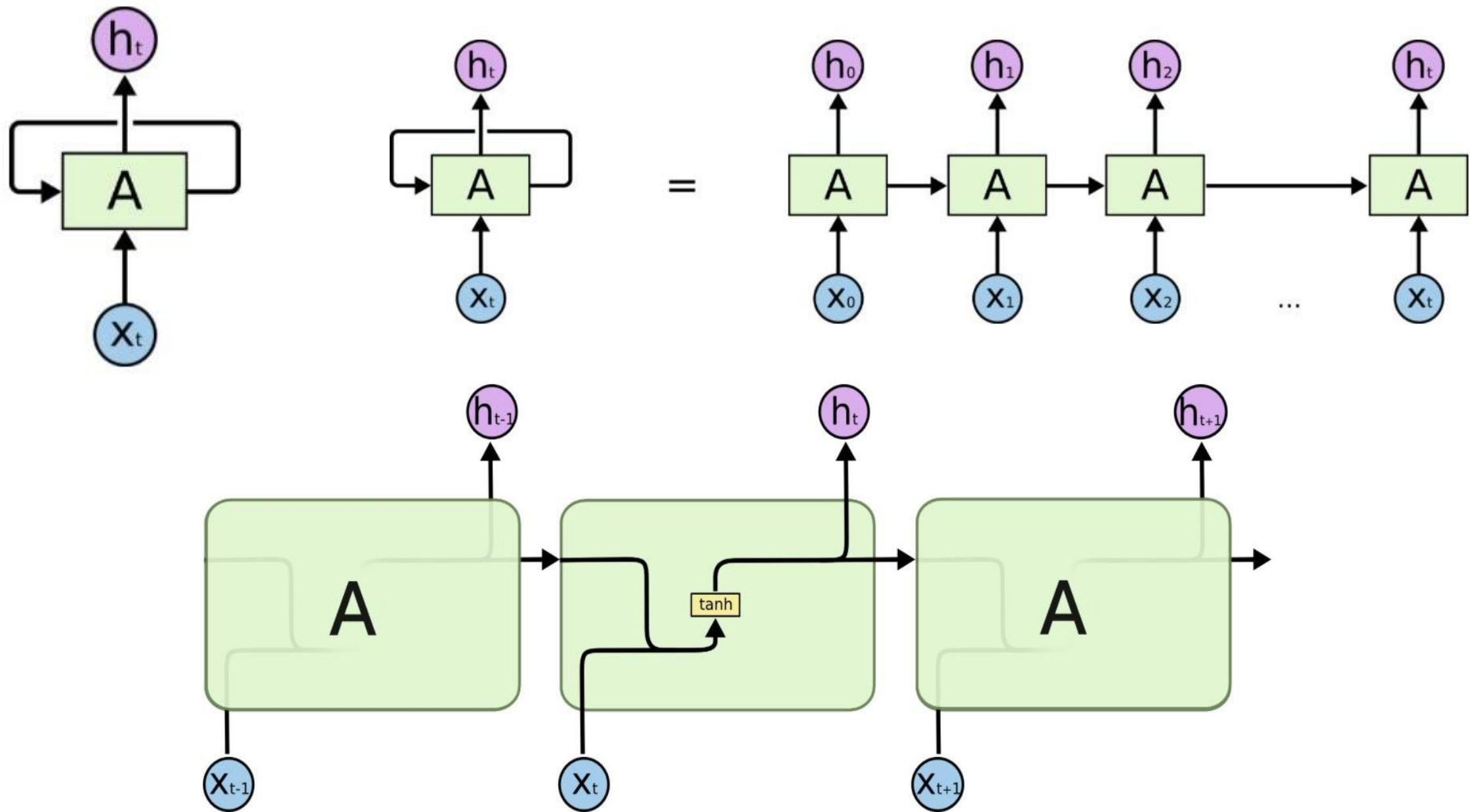
Building Blocks for the Network: RNN

- AS HUMANS WE UNDERSTAND CONTEXT
- EVERY SINGLE TIME WE DON'T RESET OUR UNDERSTANDING
- THOUGHTS HAVE PERSISTENCE
- TRADITIONAL NNS LIKE CNNs DON'T HAVE PERSISTENCE
- SPEECH RECOGNITION, LANGUAGE MODELING, TRANSLATION REQUIRES THIS PERSISTENCE

RNNs are **general computers which can learn algorithms to map input sequences to output sequences** (flexible-sized vectors).

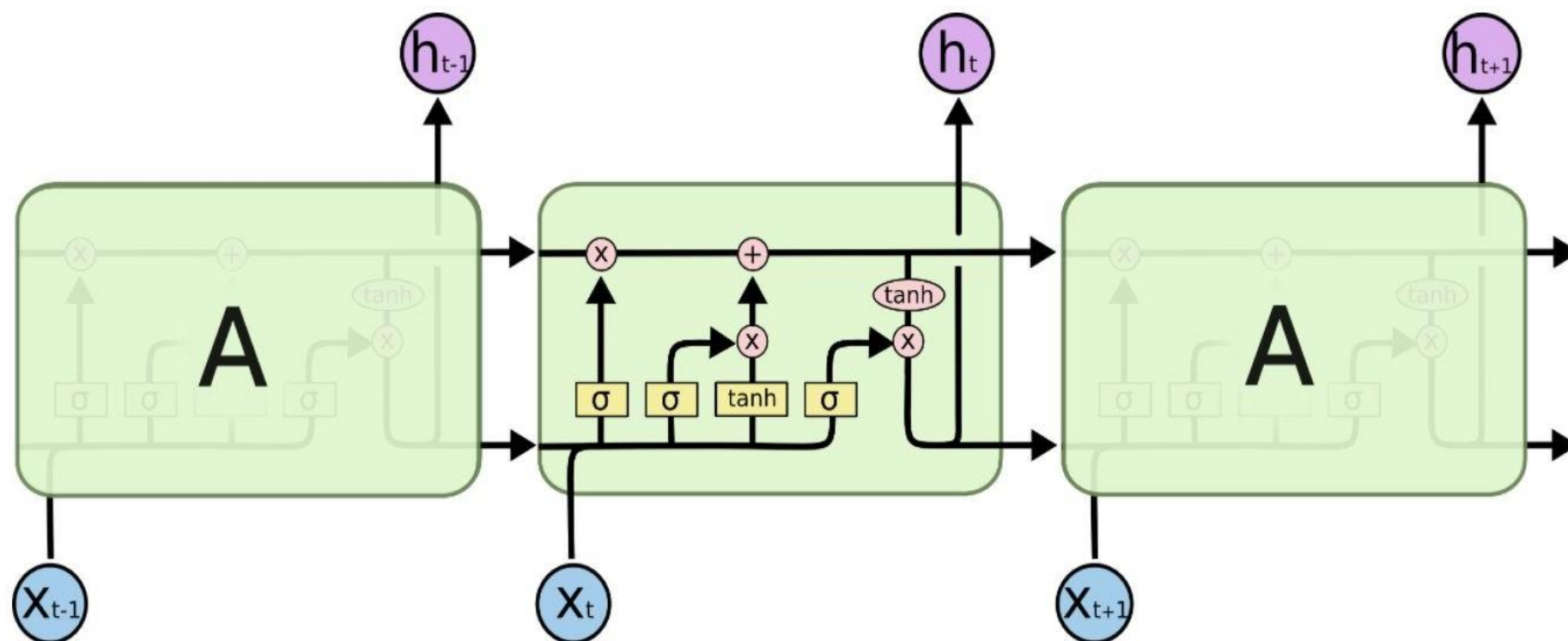
The output vector's contents are influenced by the entire history of inputs.

Building Blocks for the Network: RNN



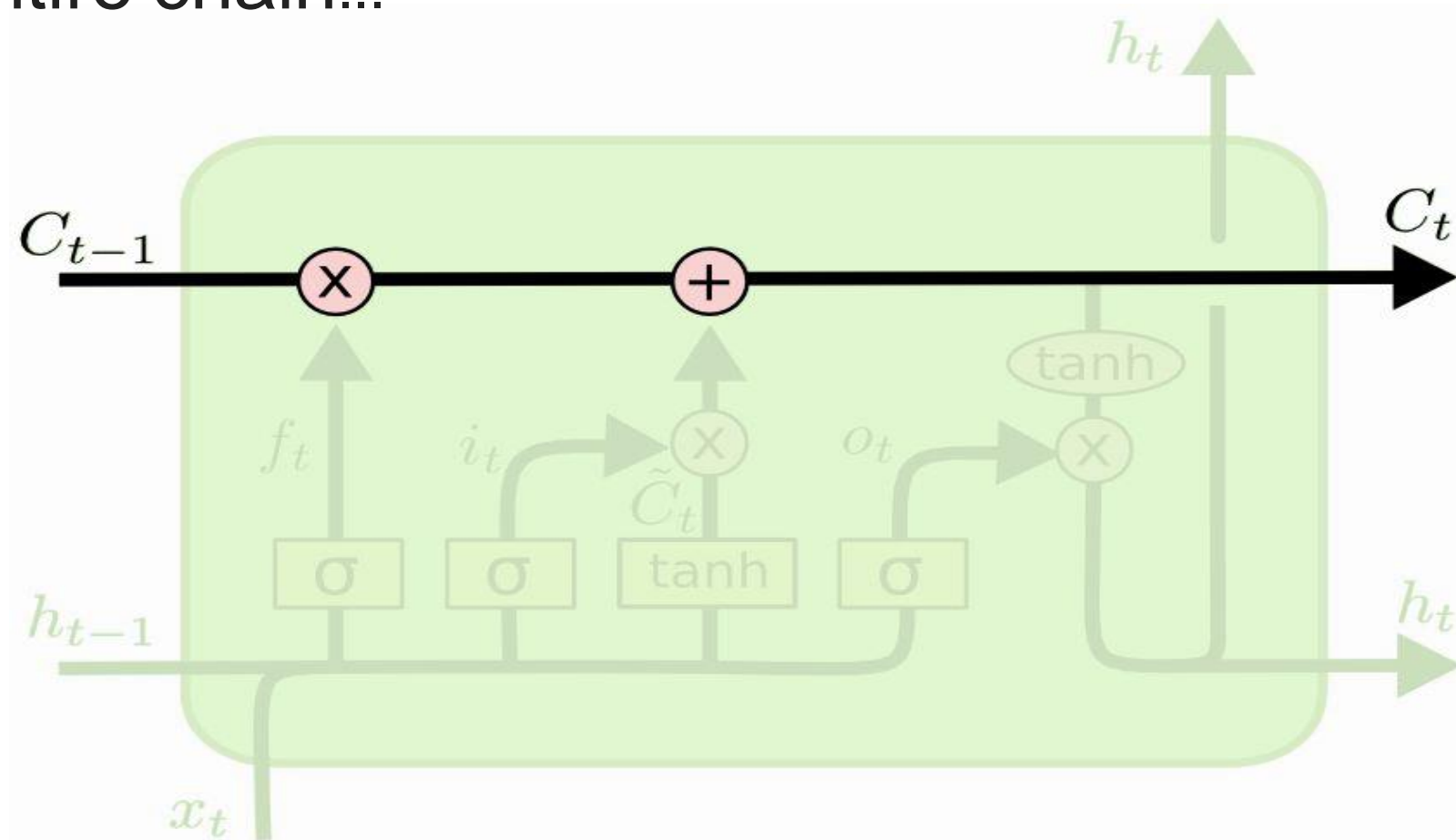
Building Blocks for the Network: LSTM

The LSTM units give the network **memory cells with read, write and reset operations**. During training, the network can learn when it should remember data and when it should throw it away.

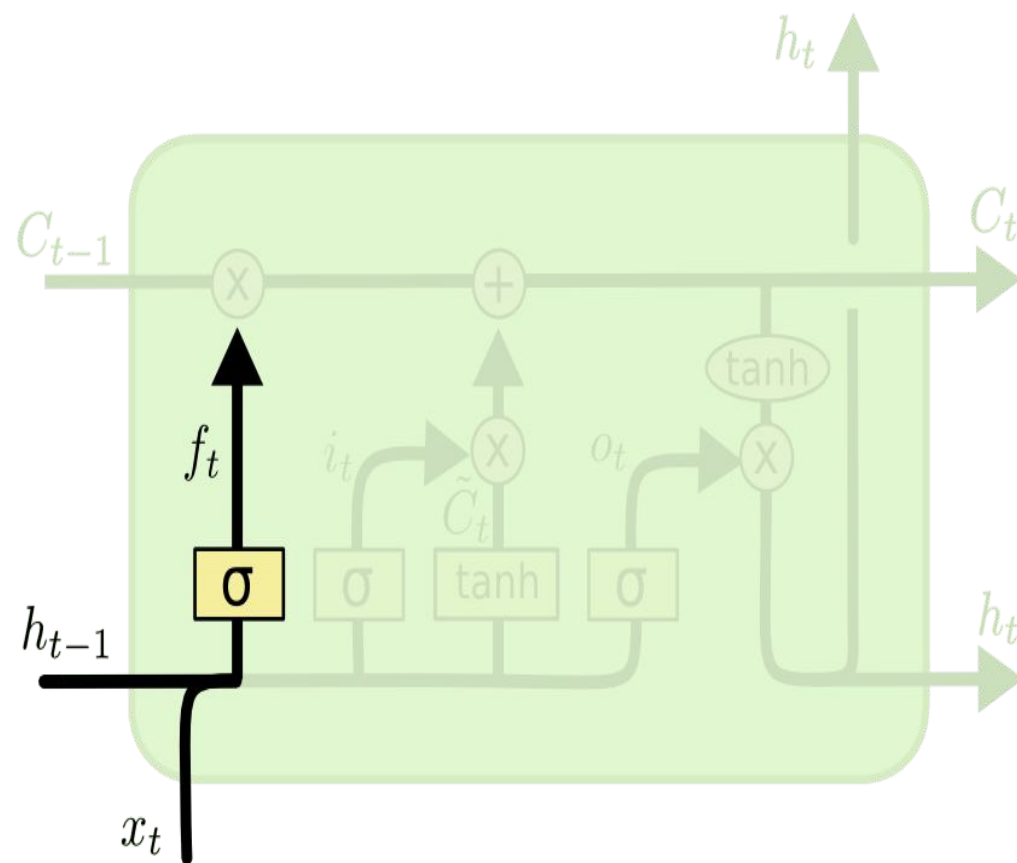


Building Blocks for the Network: LSTM

C_t is the cell state, which flows through the entire chain...



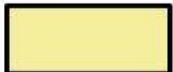
Building Blocks for the Network: LSTM




Forget Gate:


$$f_t = \sigma \left(W_f \cdot \underbrace{[h_{t-1}, x_t]}_{\text{Concatenate}} + b_f \right)$$

Concatenate


Neural Network
Layer

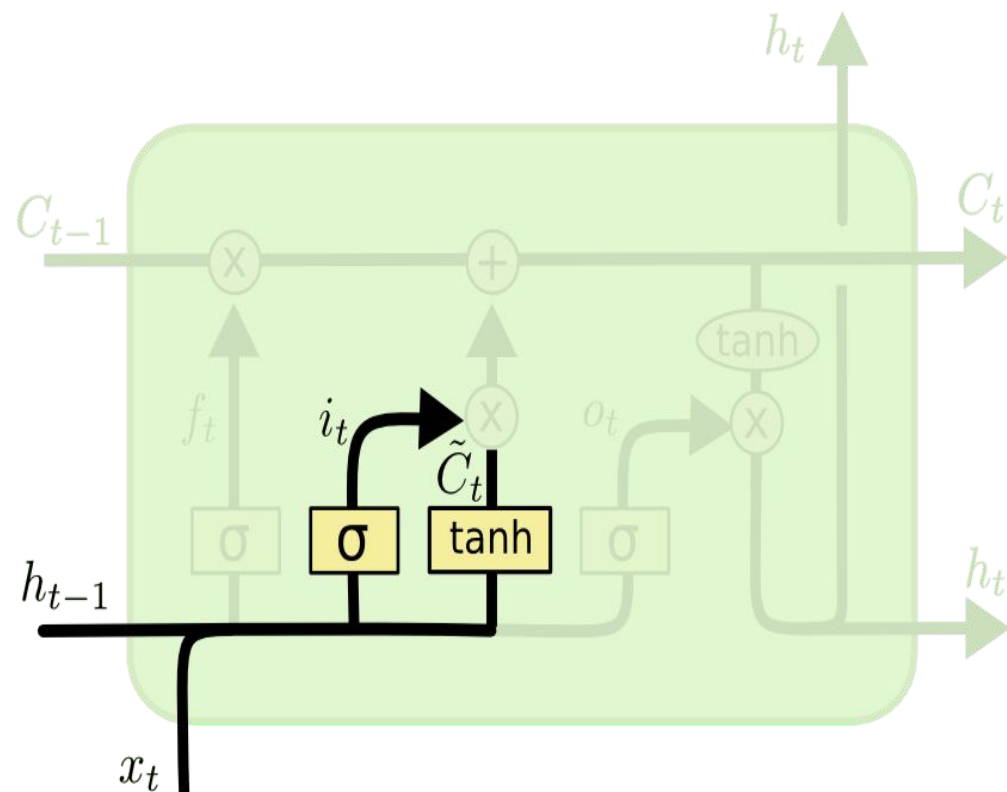

Pointwise
Operation


Vector
Transfer


Concatenate


Copy

Building Blocks for the Network: LSTM



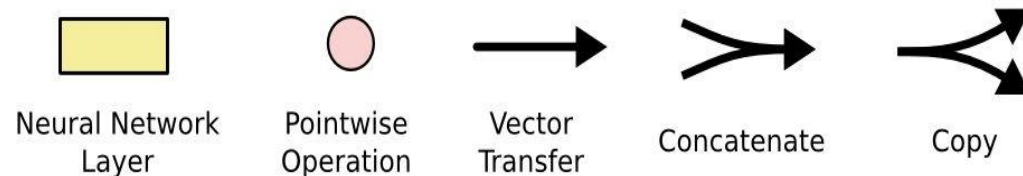
Input Gate Layer

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

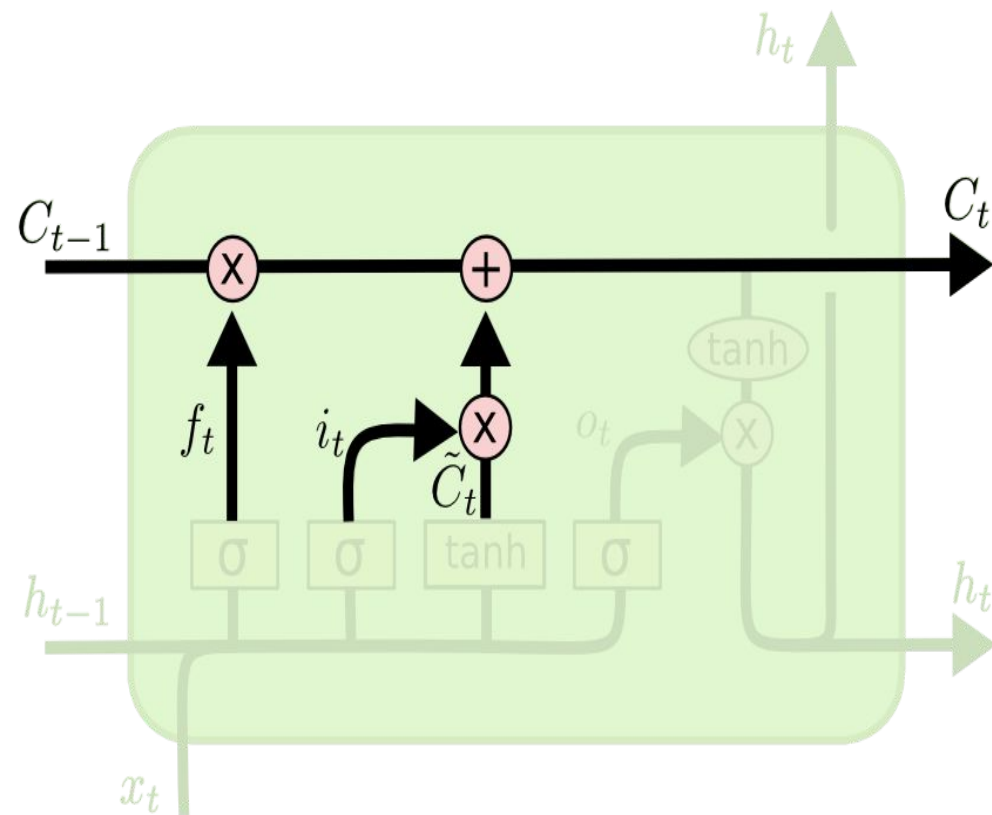
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

New contribution to cell state

$$\tilde{C}_t = \underbrace{\tanh(W_C \cdot [h_{t-1}, x_t] + b_C)}_{\text{Classic neuron}}$$



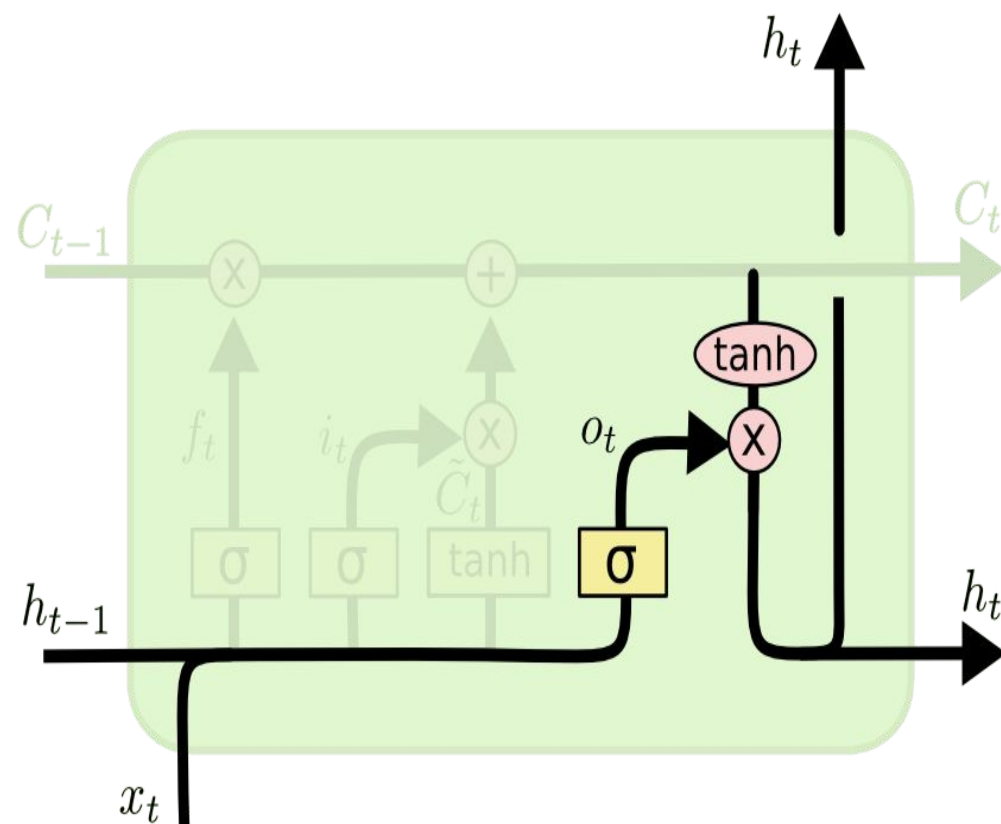
Building Blocks for the Network: LSTM



Update Cell State (memory):

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Building Blocks for the Network: LSTM



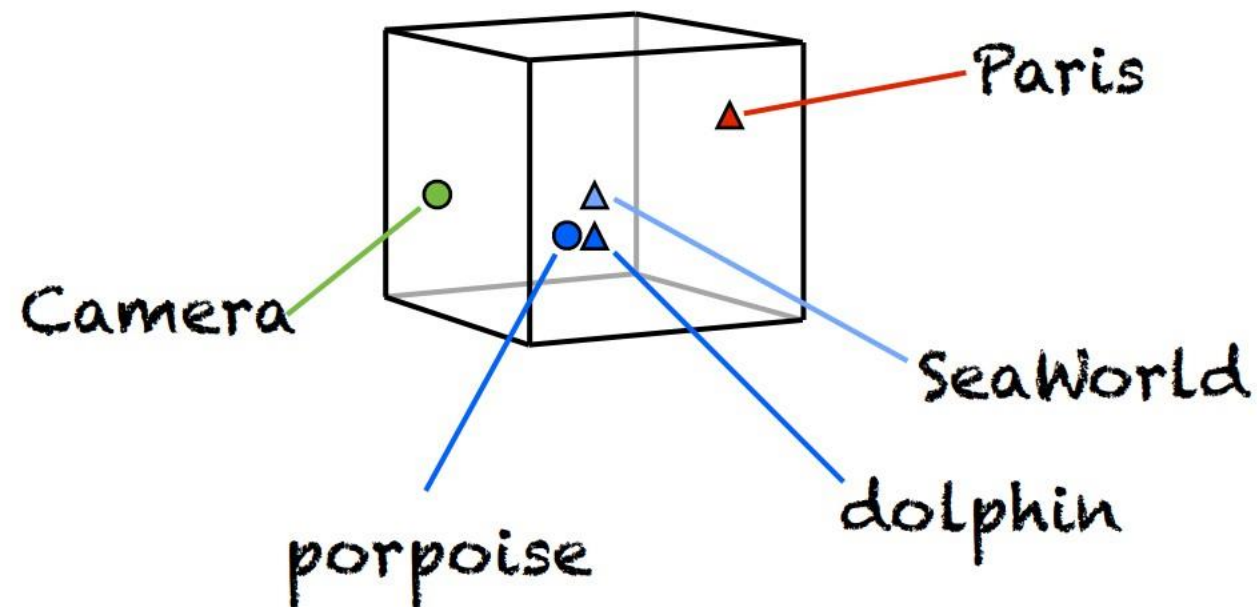
Output Gate Layer

$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

Output to next layer

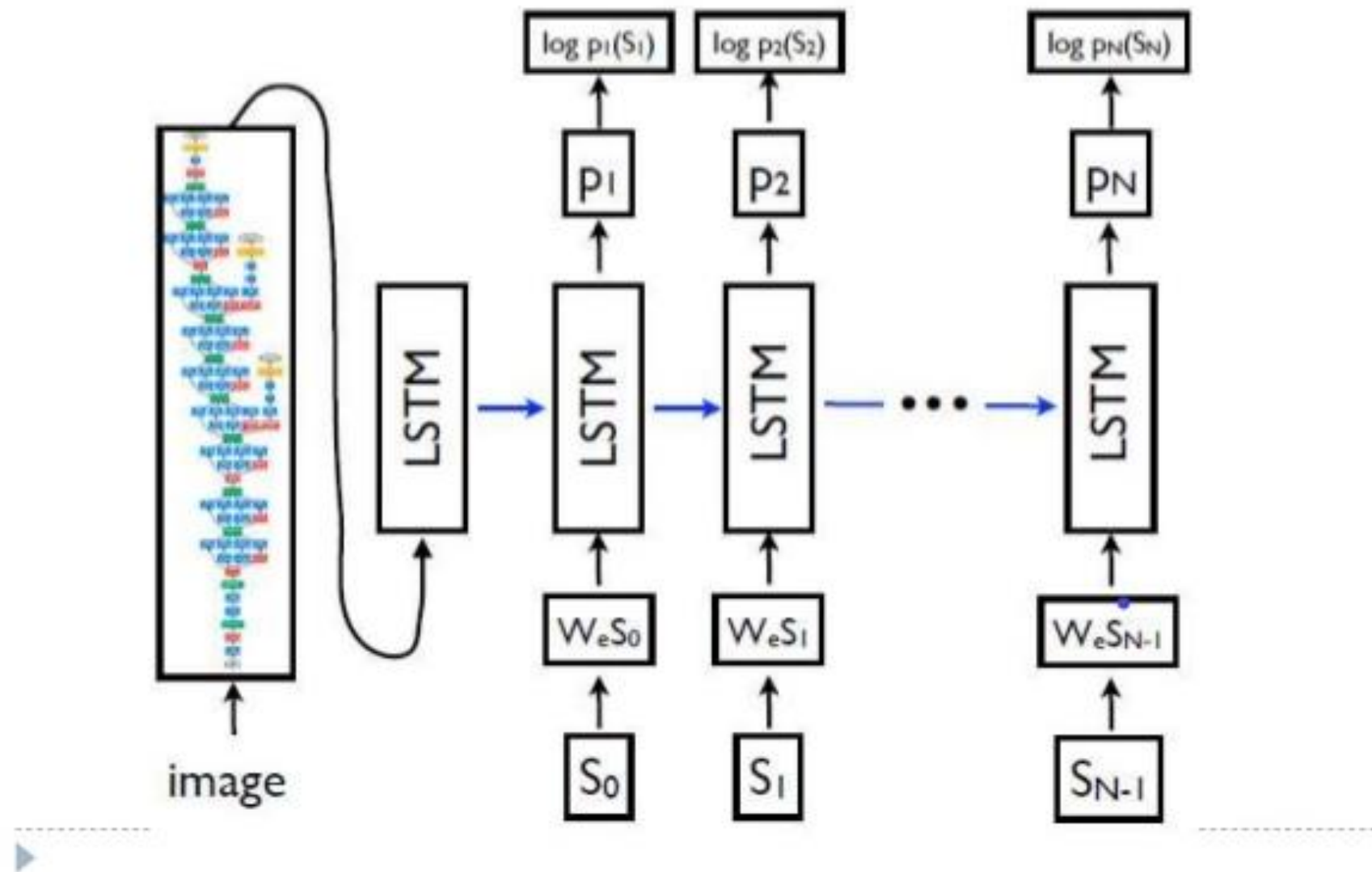
$$h_t = o_t * \tanh (C_t)$$

Building Blocks for the Network: Word Embedding



Embeddings are used to turn textual data (words, sentences, paragraphs) into high-dimensional vector representations and group them together with semantically similar data in a **vectorspace**. Thereby, **computer can detect similarities mathematically**.

Final Model:



Training Data:

Flickr8k Dataset

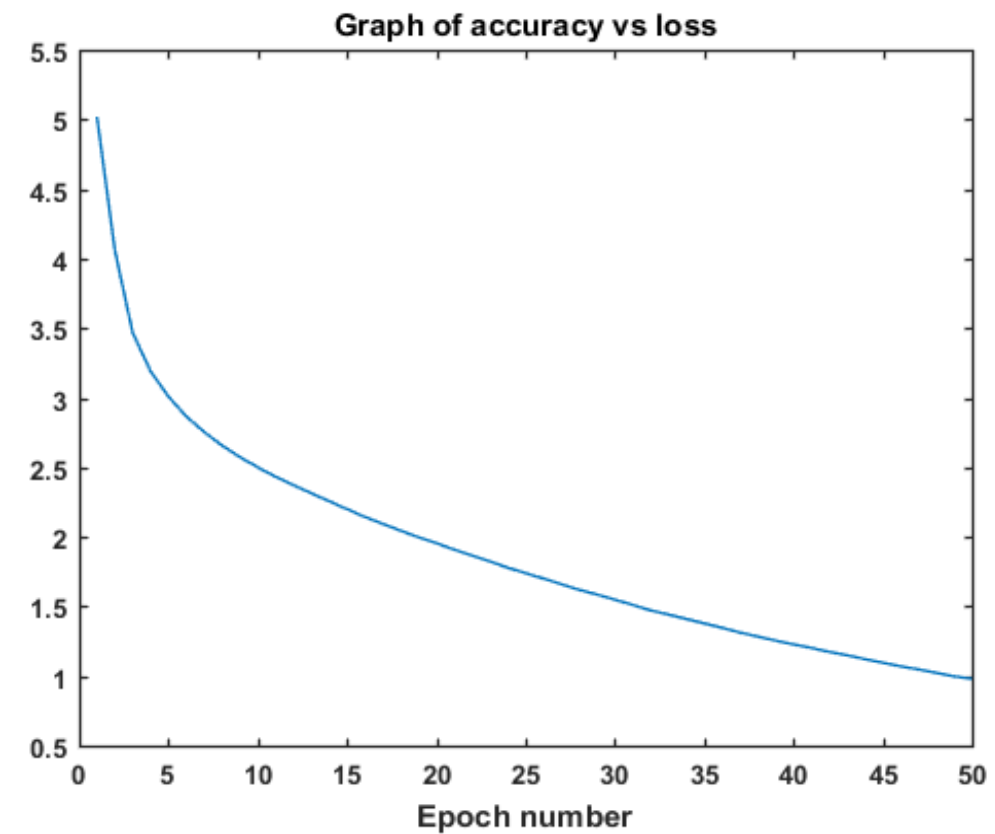
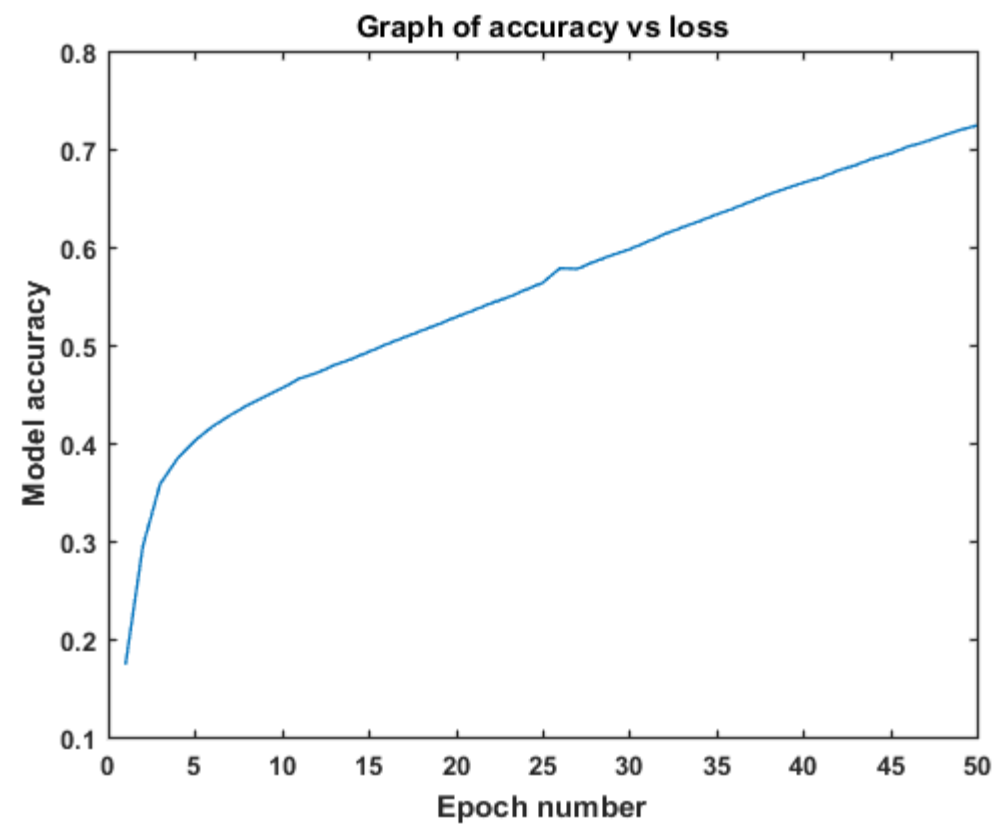
Dataset contains 8000 different images with 5 different human labelled captions.



The image is given 5 different captions:

- 1) A boy runs as others play on a home-made slip and slide.
- 2) Children in swimming clothes in a field.
- 3) Little kids are playing outside with a water hose and are sliding down a water slide.
- 4) Several children are playing outside with a wet tarp on the ground.
- 5) Several children playing on a homemade water slide.

Training History:



Model's Performance on Test Data:

Describes without errors



A person riding a motorcycle on a dirt road.

Describes with minor errors



Two dogs play in the grass.

Somewhat related to the image



A skateboarder does a trick on a ramp.



A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A little girl in a pink hat is blowing bubbles.

Model's Performance on Real Data:

Describes without errors



Three people are on a boat in the water

Describes with minor errors



Three people pose for a picture together

Somewhat related to the image



One man is sitting at a table in front of a restaurant



A soccer player prepares to kick the ball



A group of kids play in the water



A boy hits the ball at a baseball game .

Application:

- Visual to Text systems for blind people
- Search Engines for searching medical records based on content based caption
- Auto Tagging different imaging data
- Auto Video tagging and summary generation