

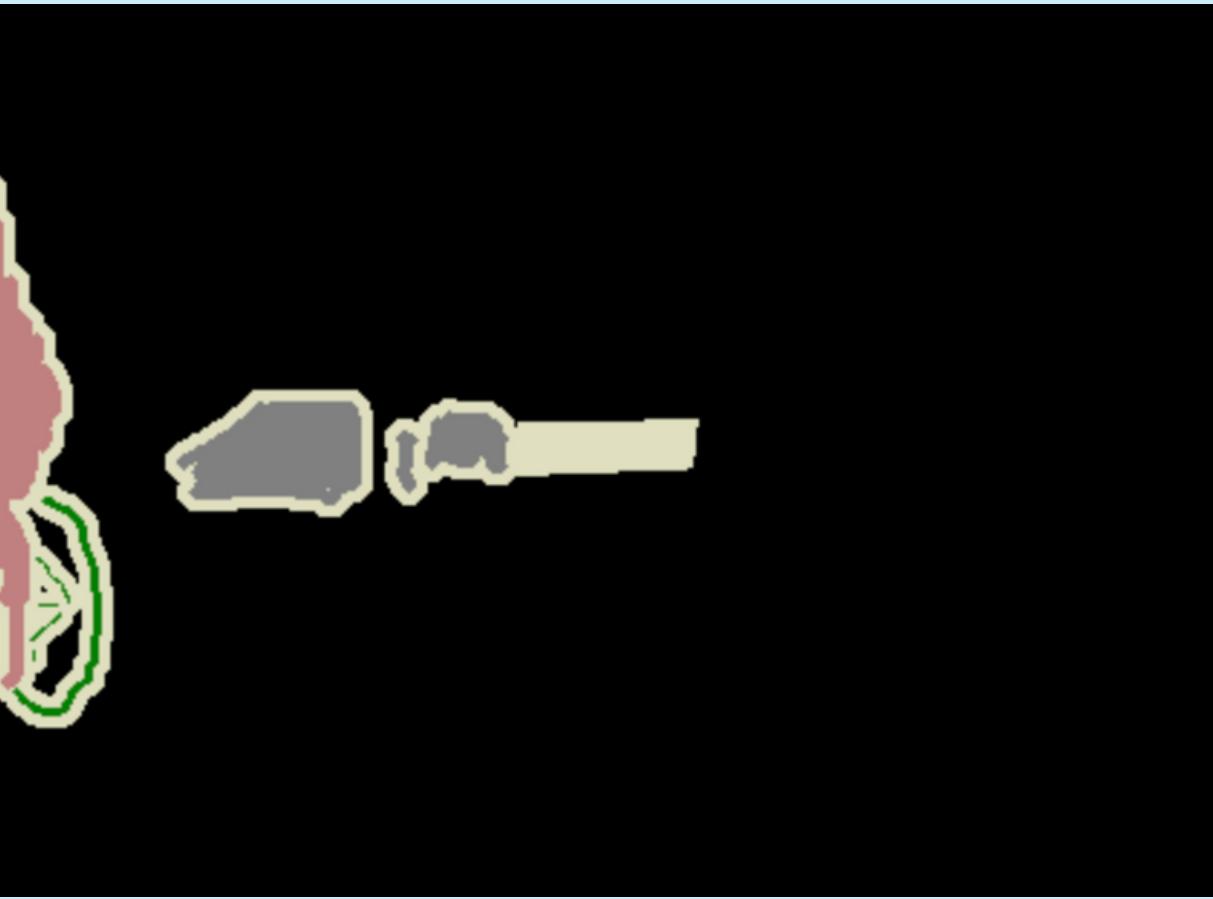
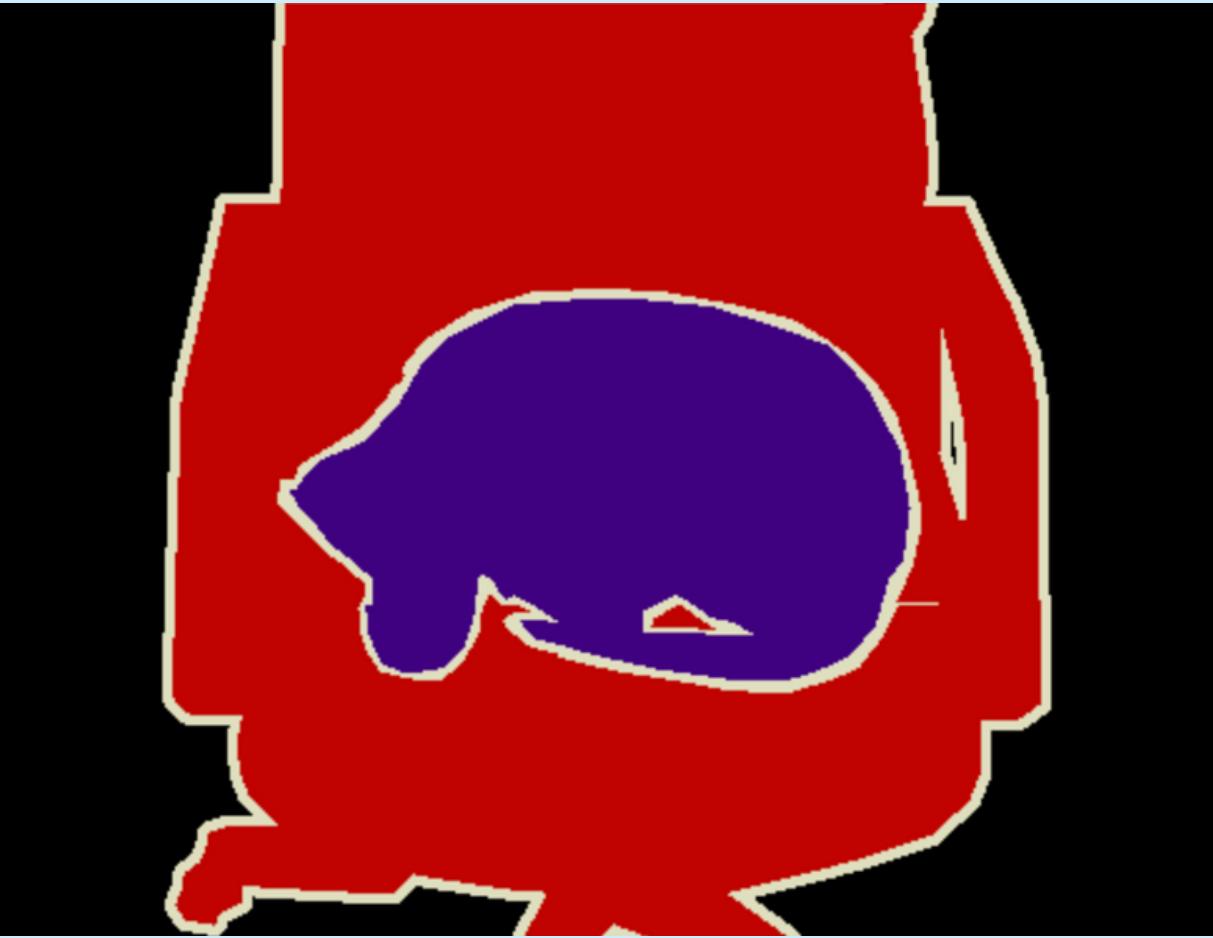
Show and Tell

*A Neural Image Caption
Generator*

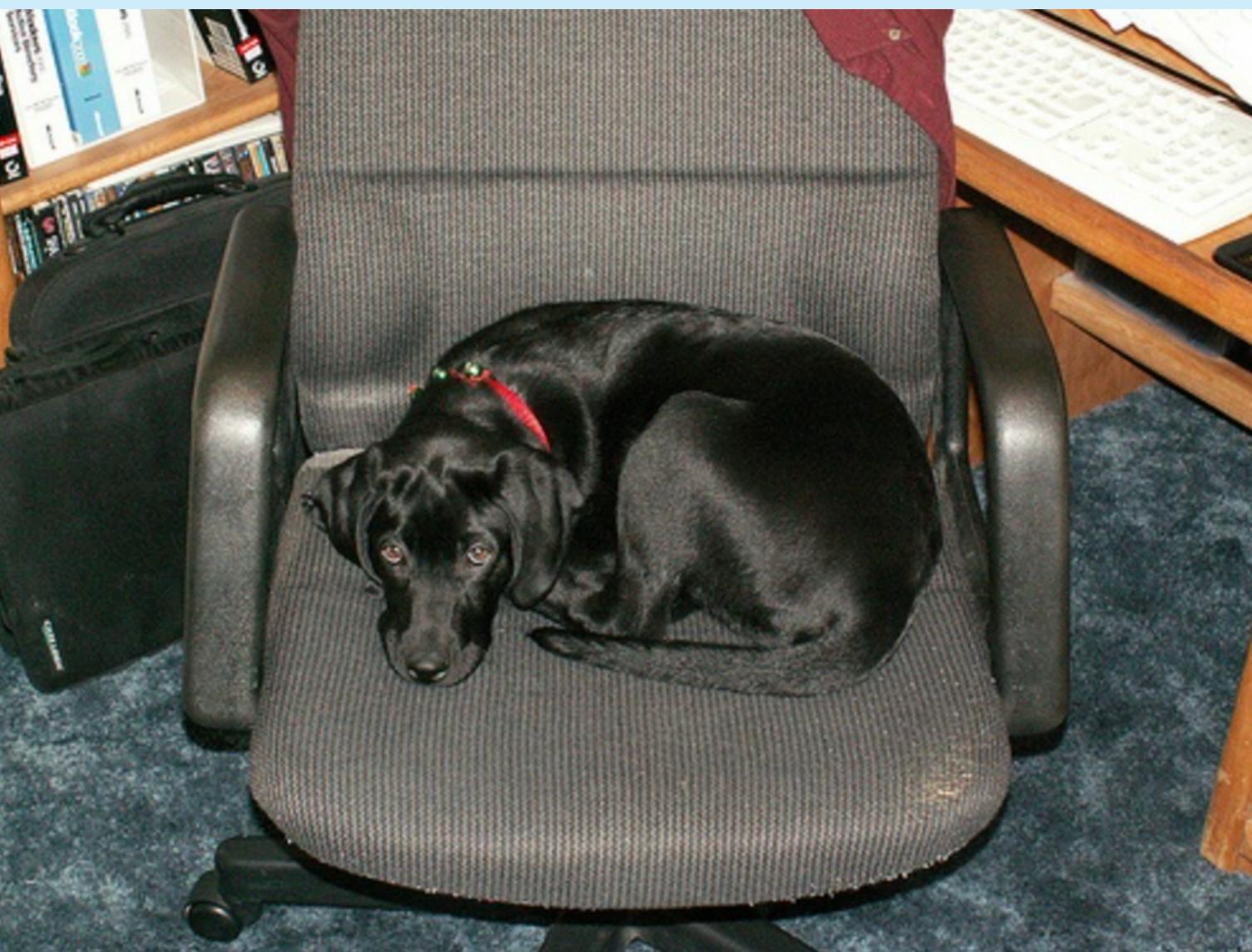
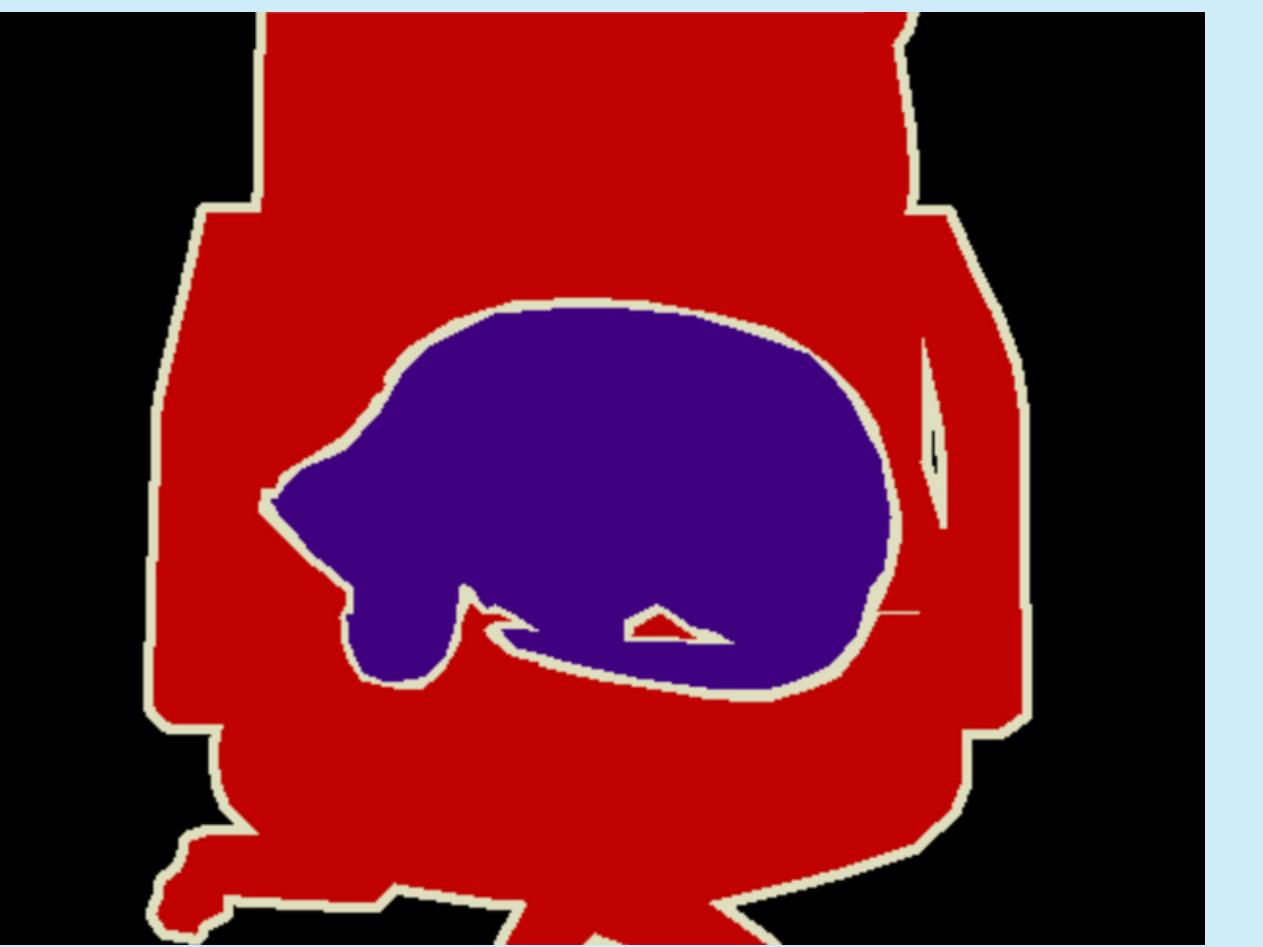
Paper Implementation, AMMI '21
by 'Lekan Raheem



**What
Do
You
See?**



**What
Do
You
See?**



The Process of Image Captioning



The Process of Image Captioning

1

Object Recognition

What objects are in
our picture?



The Process of Image Captioning

1

Object Recognition

What objects are in
our picture?

2

Object Description

Color, posture, size
etc.



The Process of Image Captioning

1

Object Recognition

What objects are in
our picture?

2

Object Description

Color, posture, size
etc.

3

Position of Objects

Where is each
object?



The Process of Image Captioning

1	2	3	4
Object Recognition	Object Description	Position of Objects	Relate all Objects in Image
What objects are in our picture?	Color, posture, size etc.	Where is each object?	The relationship!



Possible Captions...

1

Three people are in a canoe
on a calm lake with the sun
reflecting yellow

2

Three people are on a
boat in the middle of
the water while the sun
is in the back

3

Three people in a boat
float on the water at
sunset

4

Sunset with three people
in a boat on the lake, one
holds a paddle



A photograph of a person's hand holding a magnifying glass. The hand is positioned on the left side of the frame, with the magnifying glass held vertically. The background is a solid dark teal color. The magnifying glass has a black handle and a clear circular lens.

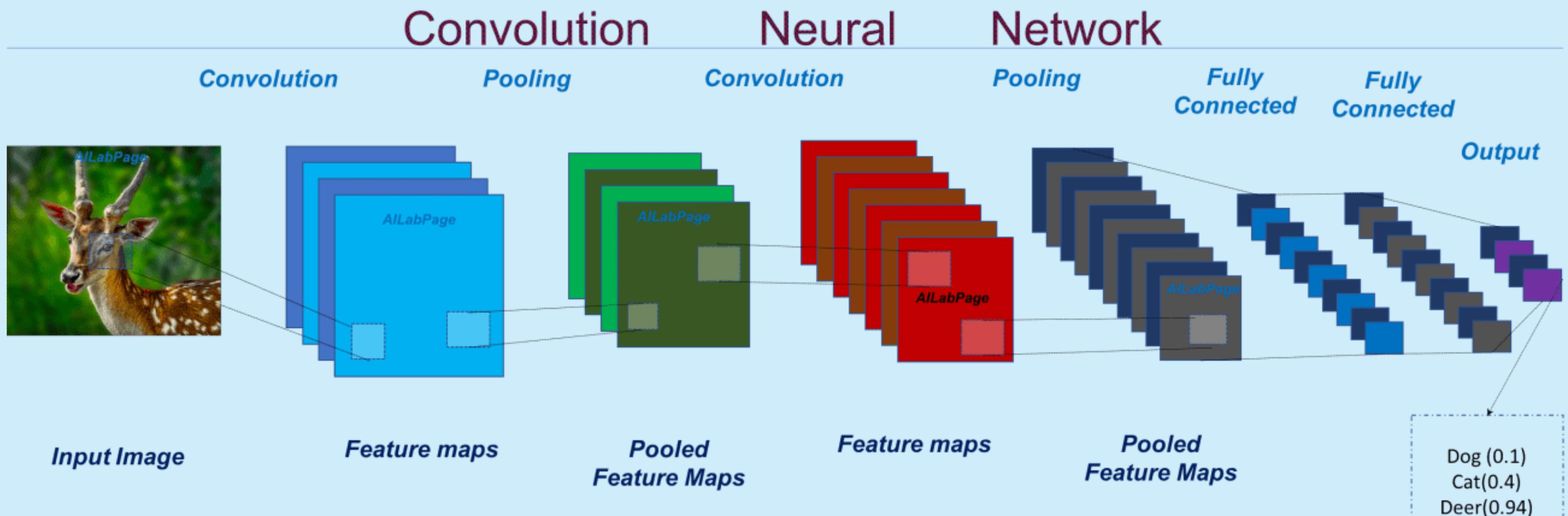
Model?

How do we do this?

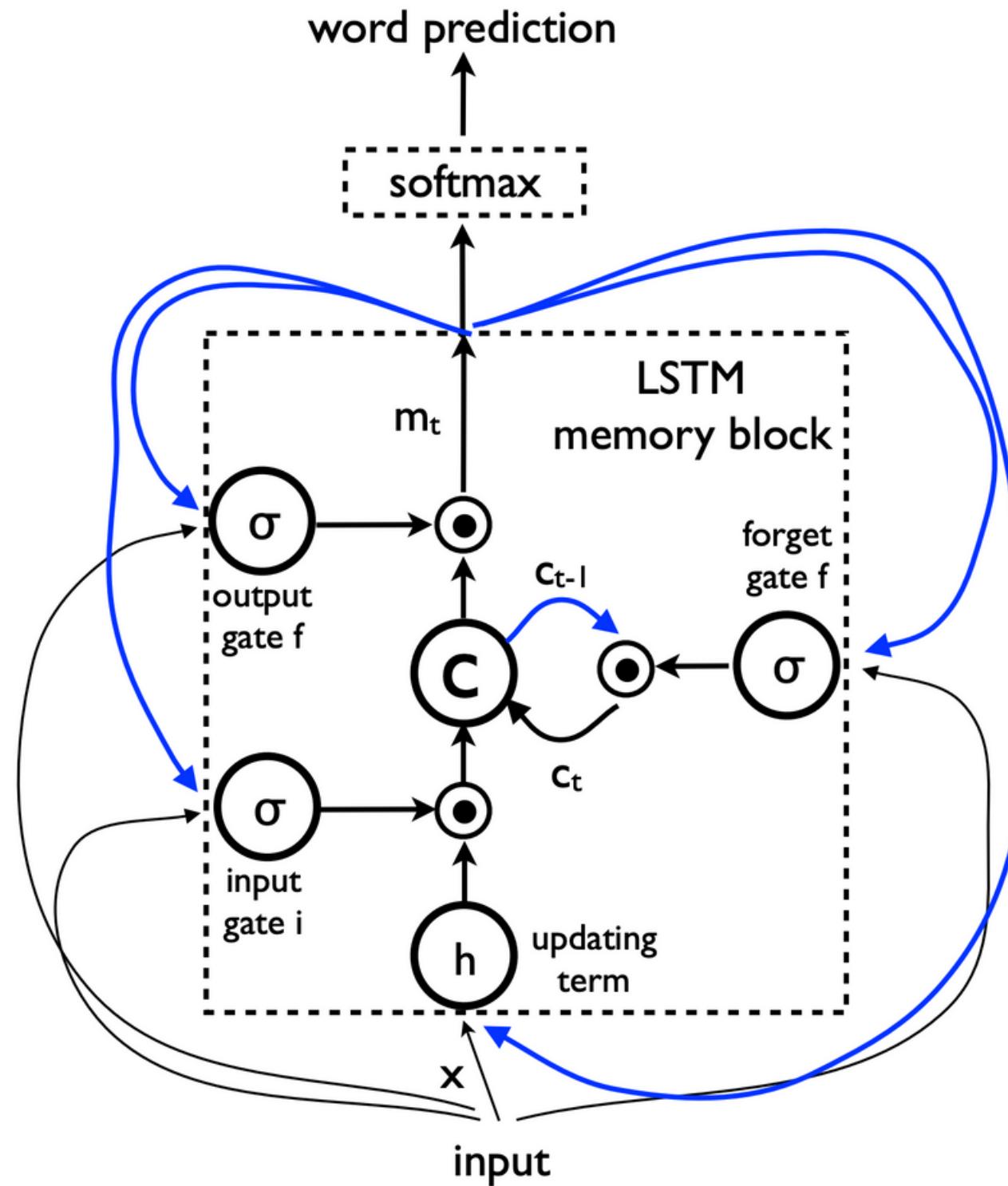
The Process of Caption Generation

1 Pretrained on ImageNet Object recognition	2 Feature Extraction features of image	2 Preprocess Captions tokenize, SOS, EOS	3 Recurrentlly Train Show feature map iteratively
---	--	--	--

Feature Extraction



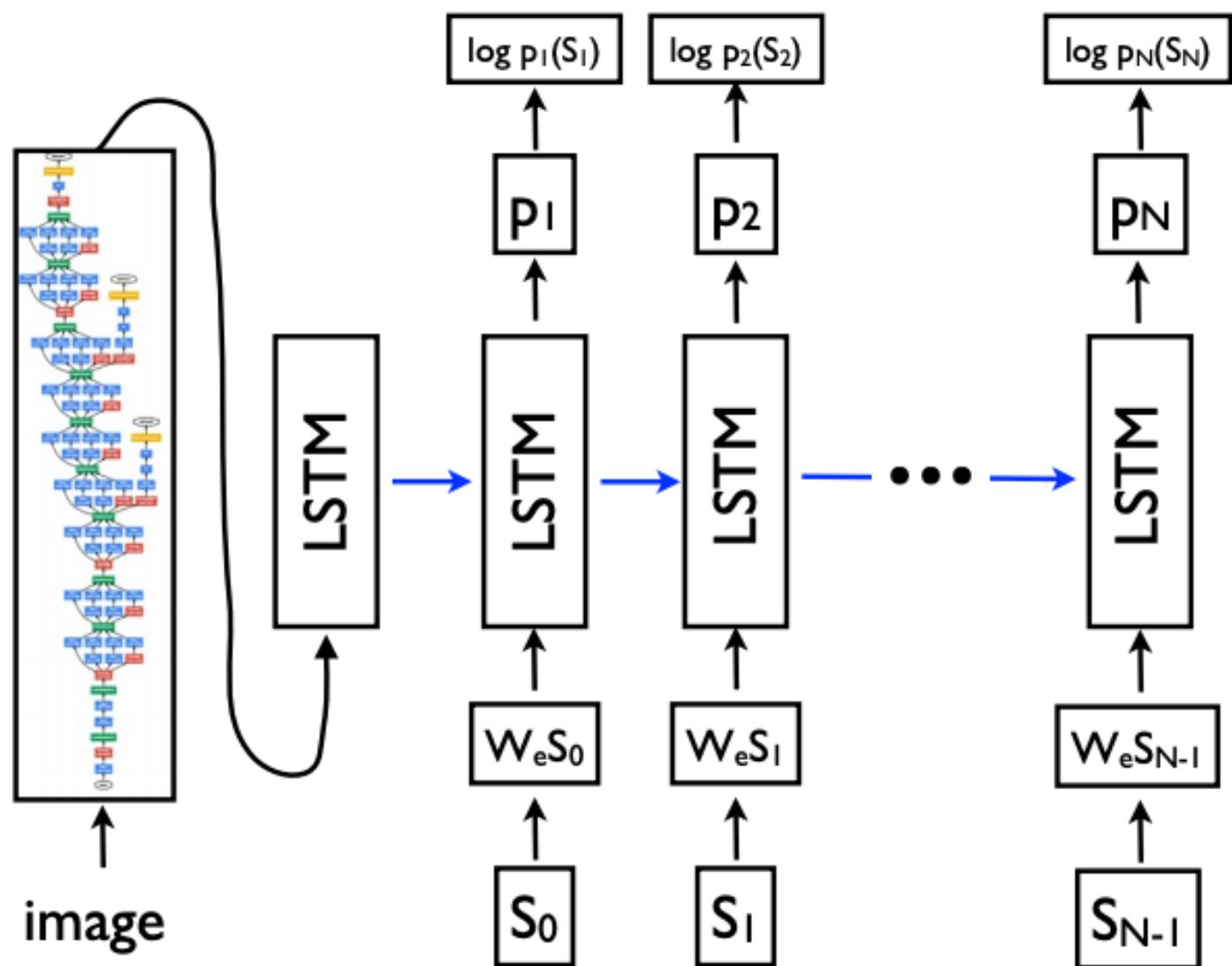
Generating each word



$$\begin{aligned} i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1}) \\ f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \\ o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \\ m_t &= o_t \odot c_t \\ p_{t+1} &= \text{Softmax}(m_t) \end{aligned}$$

Iterative Training

We train from SOS till EOS



$$x_{-1} = \text{CNN}(I)$$

$$x_t = W_e S_t, \quad t \in \{0 \dots N-1\}$$

$$p_{t+1} = \text{LSTM}(x_t), \quad t \in \{0 \dots N-1\}$$

Really!

Thats all it takes?



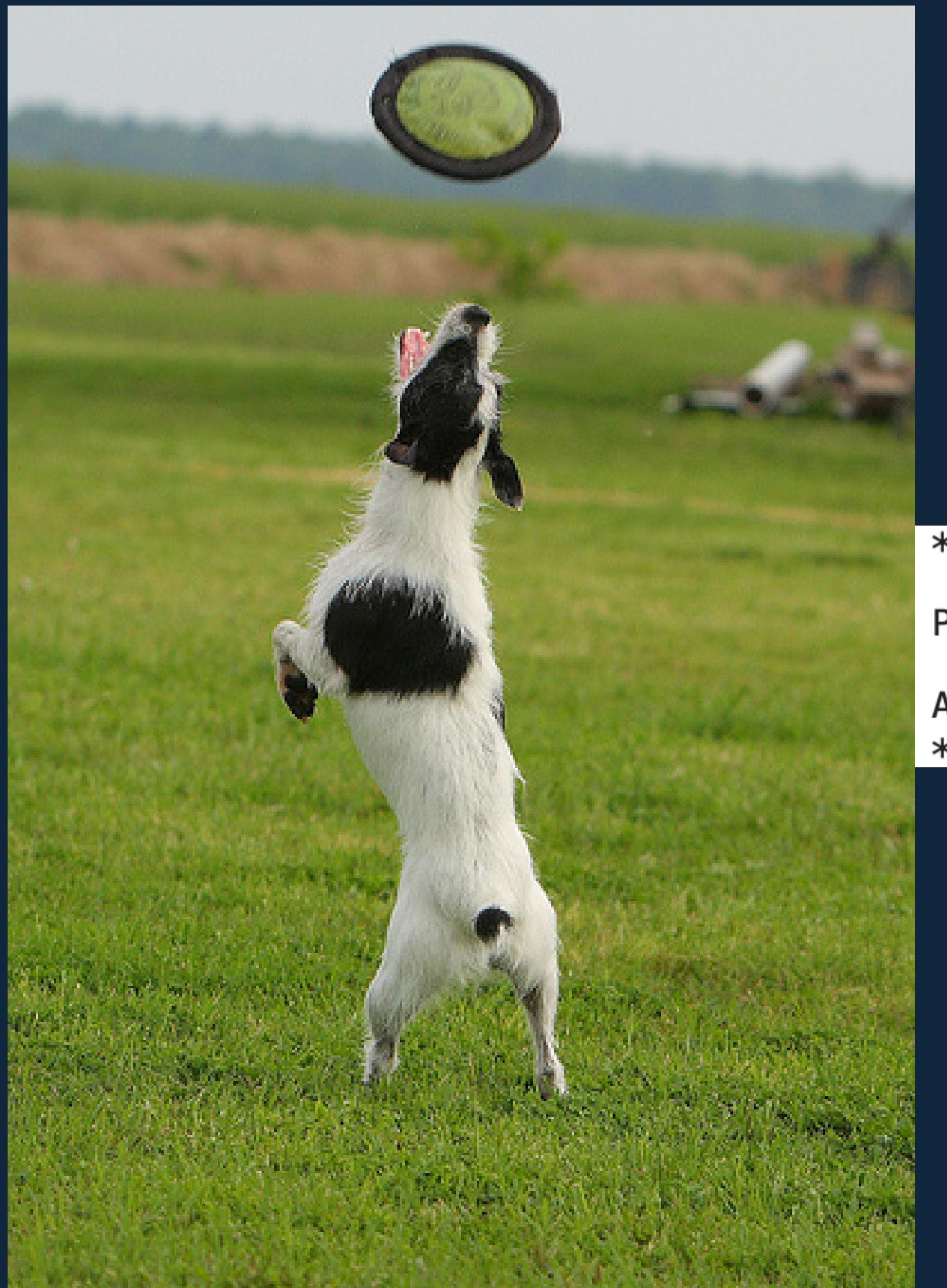




That was Close!

Predicted caption -> <sos> brown dog is sitting <eos>

Actual caption -> <sos> brown dog is sitting in some long grass <eos>



Not really!

Predicted caption -> startseq black and white dog is jumping over hurdle end

Actual caption -> startseq black and white dog is playing with ball on lawn



Misfire!

Predicted caption -> startseq man in red shirt is standing on skateboard in the snow

Actual caption -> startseq girl holds something while three dogs beg endseq



Arghhhhhh!



Predicted caption -> <sos> The man is climbing up the rock <eos>

Actual caption -> <sos> A person is climbing a cliff wall , over a rocky shore , using a rope <e

A close-up photograph of a woman's face, partially obscured by a black hijab. Her eyes are closed, and her hands are clasped together in a prayerful gesture at the bottom of the frame. The background is a solid dark blue.

Thank You!

Questions