深度學習基礎概論

0513

目錄

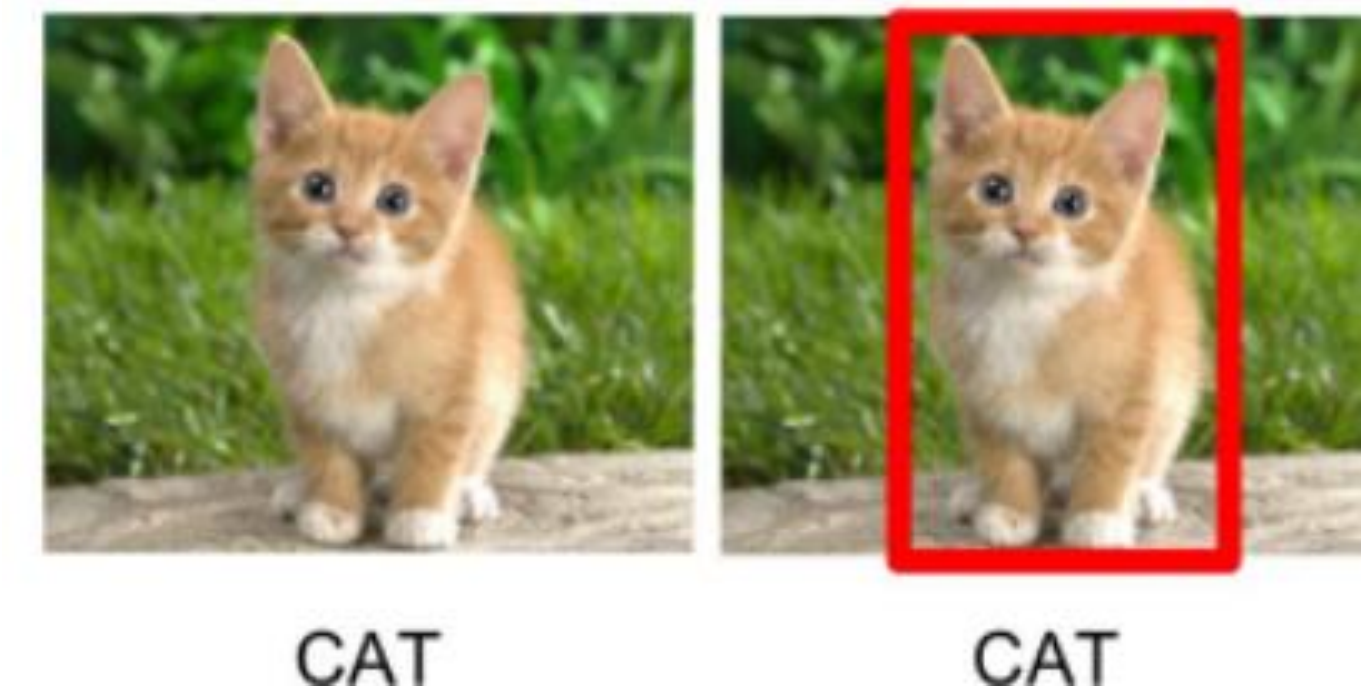
- Image caption with attention

- Code

Image caption

- 捕捉圖片上某些特徵

- E.g. 邊界徵測



Attention

- 考慮字跟字之間的關係

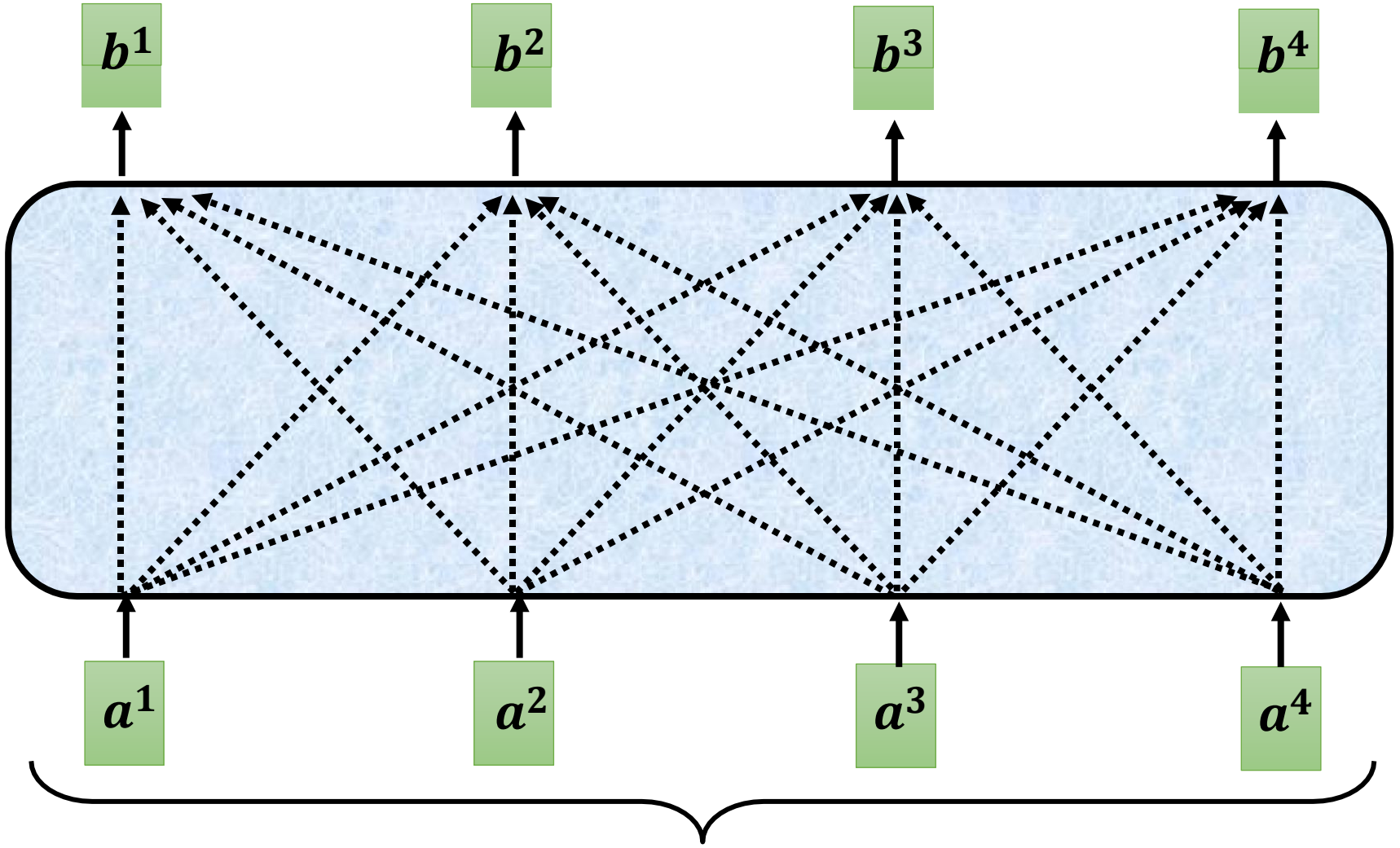


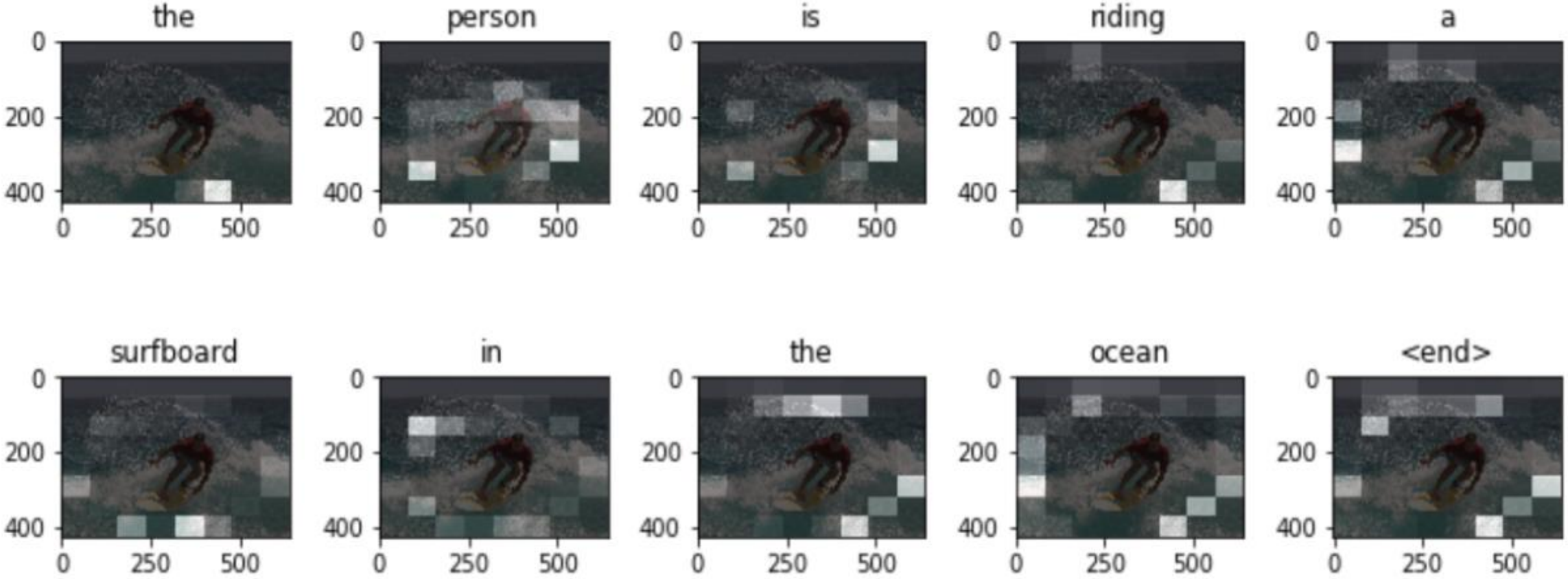
Image caption with attention

- 本質上是圖片訊息到文本間的翻譯，也就是圖片和文字之間的關係，

The man is riding a surfboard in the ocean

Image caption with attention

Prediction Caption: the person is riding a surfboard in the ocean <end>



白色的區塊是該字關注的區塊，越白代表關注度越高，也就是 attention 越大

Image caption with Attention

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

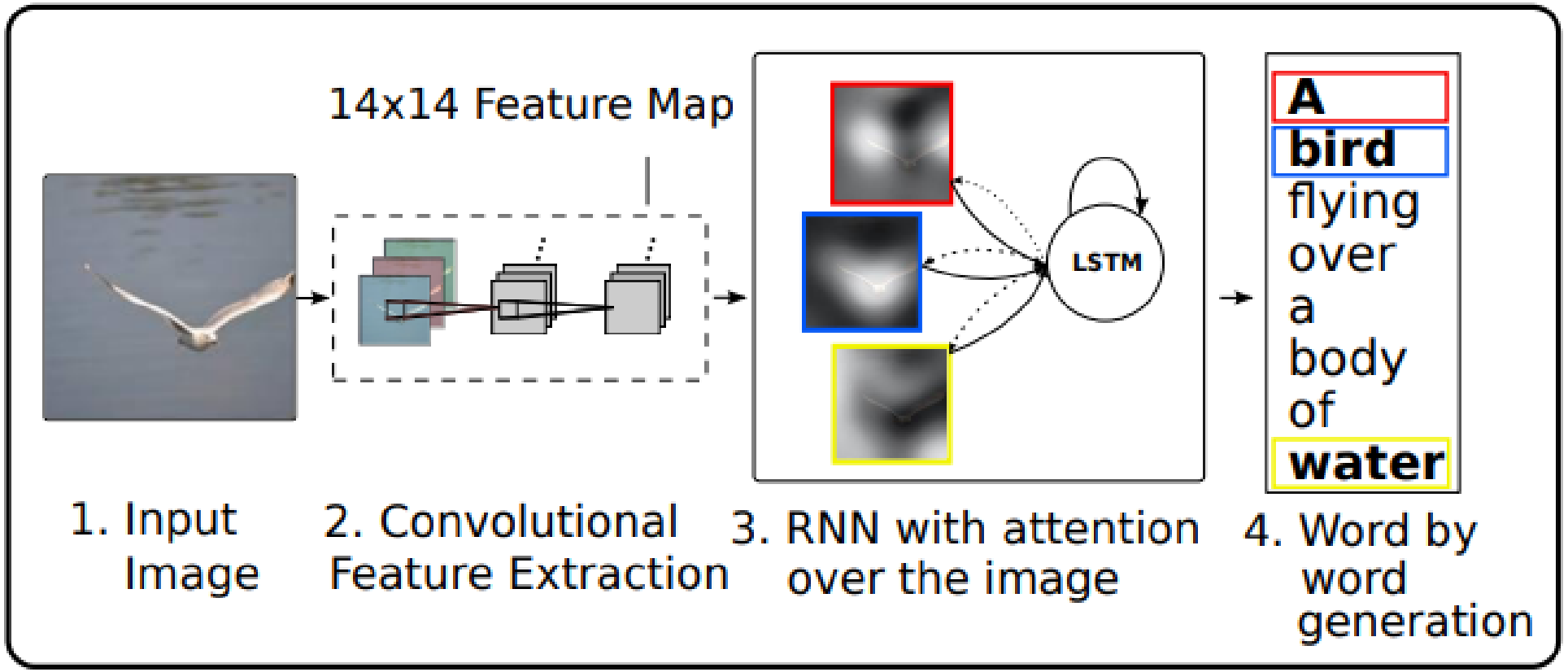
A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

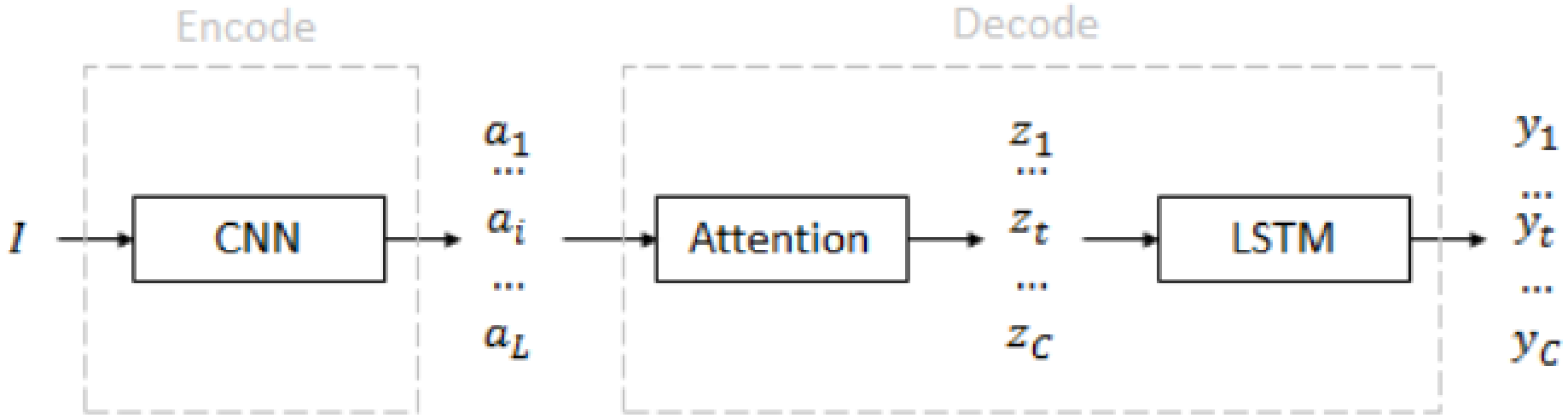
A giraffe standing in a forest with trees in the background.

Model

- encoder – decoder framework



Model



annotation(特徵) : $\{a_1, \dots, a_L\}$ a_i 是一個 K 維的 vector，K 是字典檔的大小， L 是 pixel 的數量

context(上下文) : $\{z_1, \dots, z_c\}$ z_t 是一個 K 維的 vector,c 是output 句子長度

output : $\{y_1, \dots, y_c\}$

Model

- CNN model ， 抽取圖片的特徵
- Attention : 為圖片的每個區塊給予一個權重，也就是該區塊的關注度
- LSTM ，output 一段文字

Encoder

將 Input 的圖像 reshape 到 244*244，特徵向量直接從 VGG 中的 conv5_3 層抽取，為 14*14*512 維，使用 lower level 的特徵是因為我們在乎的是局部的特徵，而非整張圖片

區域數量 $L = 14 \times 14 = 196$, 維度 $D = 512$

$$a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \ \mathbf{a}_i \in \mathbb{R}^D$$

Attention

α_t 為權重維度維 $L = 196$，紀錄每個 pixel 的權重

z_t 也是一個 D 維的向量

Decoding 是逐個單字進行的，所以 attention 會進行很多次，且依靠前一期的資訊

$$z_t = \alpha_t^T \cdot \mathbf{a}$$

Attention

h_t 為 LSTM 的 hidden layer

f_{att} 是 model 的 attention function

我們將 α_{ti} 當成個 pixel 的權重，所以過 softmax 使合為 1

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.$$

Attention

Attention function

- Hard attention : 隨機抽取 hidden layer 去做 attention，比較不容易做 backpropagation
- Soft attention : 使用所有該時點所有的 hidden layer 去做 attention ，容易進行 backpropagation
- ϕ is a function that returns a single vector

$$\hat{\mathbf{z}}_t = \phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right)$$

Dncoder

LSTM model

$Z_t = \phi(\{a_i\}, \{\alpha_i\})$ CNN 的output 經過 attention 後 vector

h_{t-1} : 前一時點 hidden layer vector

$E_{y_{t-1}}$: 前一時點的 output 經過 embedding 後的 vector

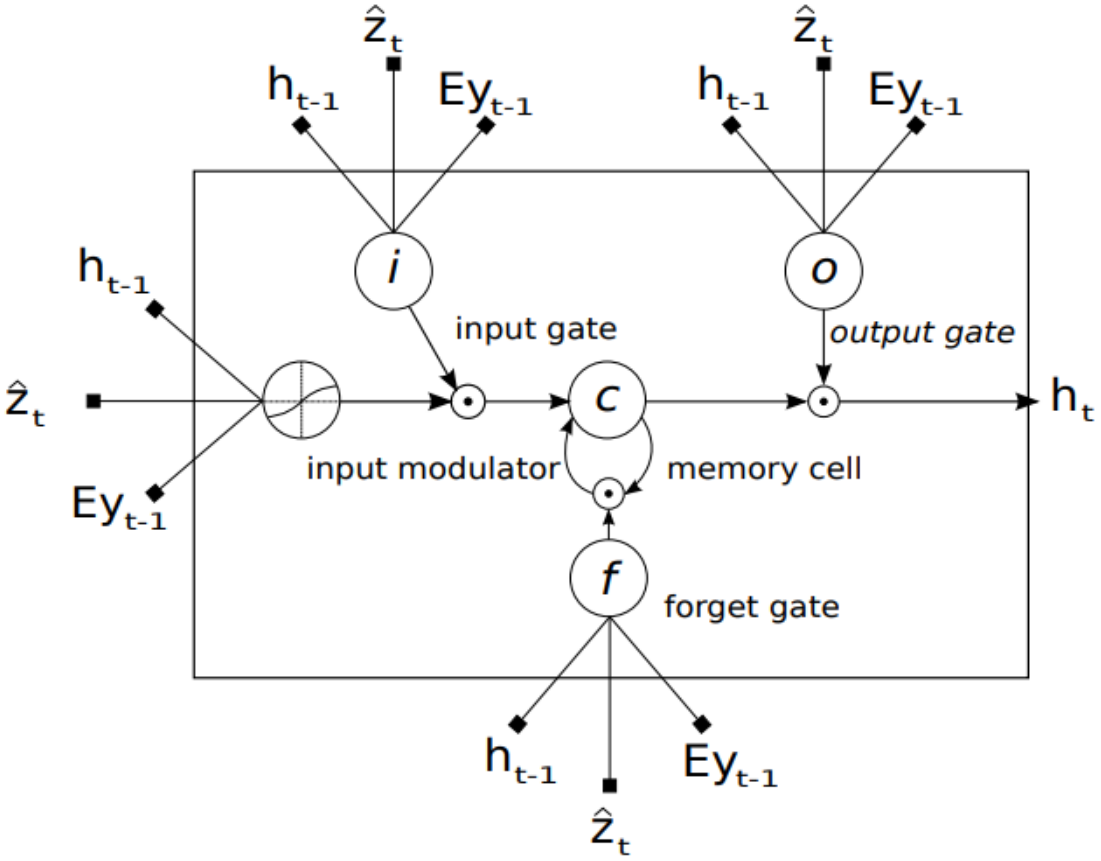


Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and
a hat on a skateboard.

A person is standing on a beach
with a surfboard.

A woman is sitting at a table
with a large pizza.

A man is talking on his cell phone
while another man watches.

Code