

# Sapienza Training Camp 2021

Building an Image Search Engine

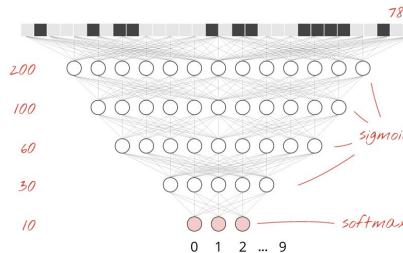
2 - 4 September, 2021

# Recap: What did we learn so far?

# Recap: Dense and convolutional layers

- Dense layer:

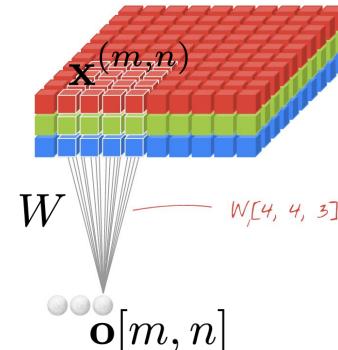
$$\mathbf{o} = g(W\mathbf{x} + w_0), \quad \text{where } \mathbf{x} \in \mathbb{R}^d$$



- Convolutional layer:

$$\mathbf{o}[m, n] = g(W\mathbf{x}^{(m, n)} + w_0), \quad \text{where } \mathbf{x}^{(m, n)} = \mathbf{x}[m:m+D, n:n+D]$$

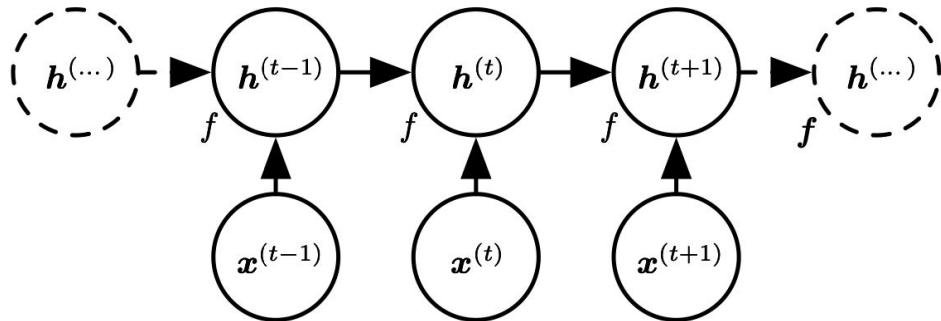
- **weight sharing:** same weights used for all local windows



Images from ["TensorFlow, Keras and deep learning, without a PhD"](#)

# Recap: Recurrent layer

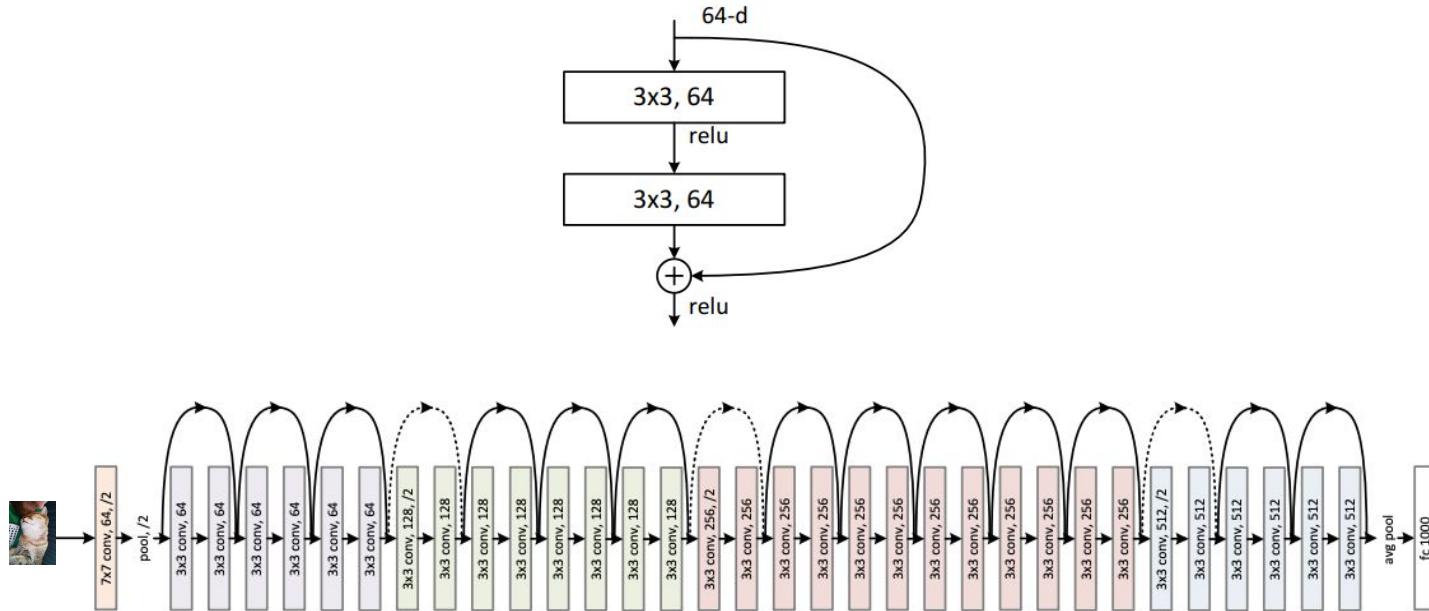
- Recurrent layer
  - **weight sharing** across time steps



$$\mathbf{h}^{(t)} = g(W\mathbf{h}^{(t-1)} + U\mathbf{x}^{(t)} + w_0)$$

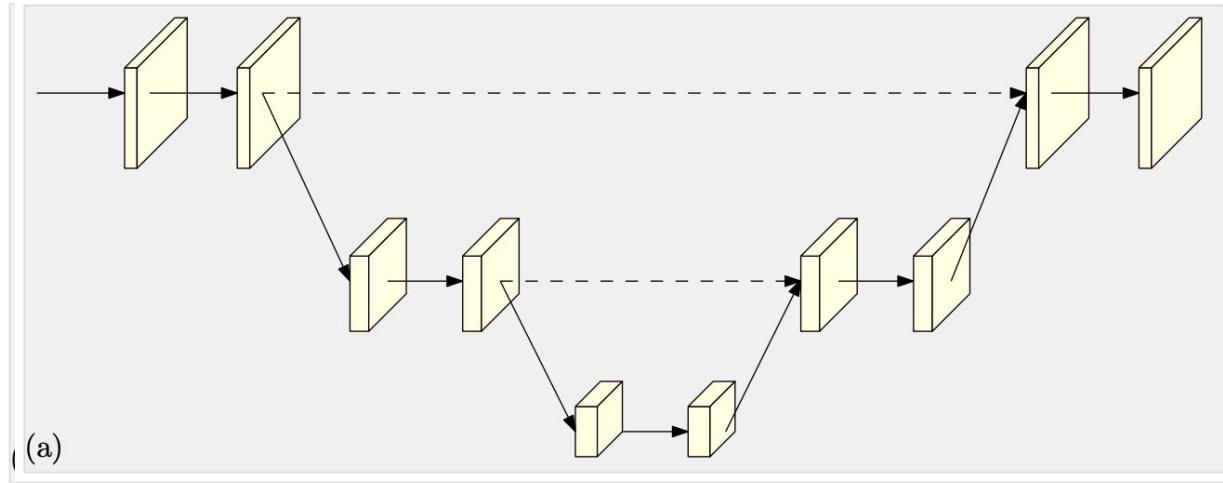
where  $\mathbf{h}^{(t)} \in \mathbb{R}^K$

# Recap: Neural network architectures



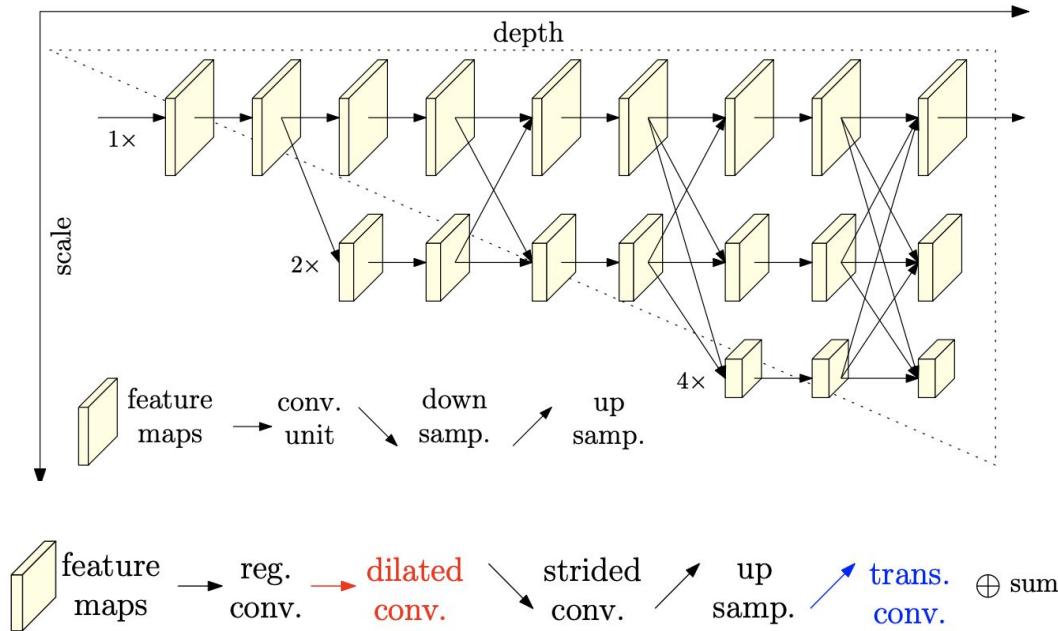
He et al., Deep Residual Learning for Image Recognition, CVPR'16

# Hourglass CNN



 feature maps → reg. conv. → dilated conv. → strided conv. → up samp. → trans. conv. ⊕ sum conv.

# HRNet



# Roadmap

Image from the photo collection

$I_k$



Neural Network (CNN)

Image encoding

$\varphi_k$

Recurrent neural network (RNN)

$T_k$

Image description

= "dog next to person on the bike near street crossing"

Define similarity function. Order images according to similarity to the query.

$Q$

Query: "person walking with a dog on the beach"



$$\text{sim}(Q, T_1) > \text{sim}(Q, T_2)$$

# More complex RNN units: GRU and LSTM

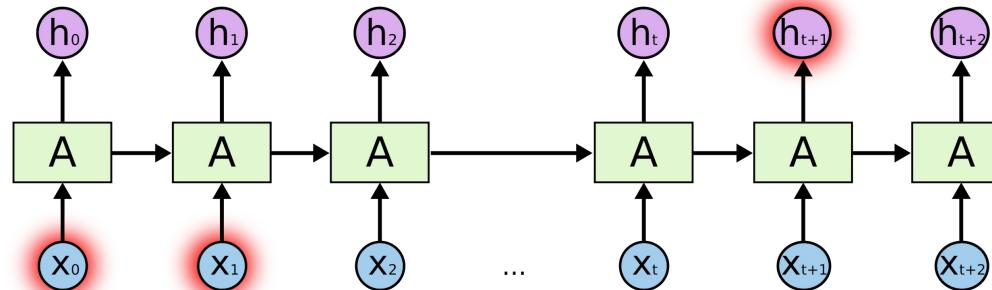
```
class RNN_Decoder(tf.keras.Model):
    def __init__(self, embedding_dim, units, vocab_size):
        super(RNN_Decoder, self).__init__()
        self.units = units

        self.embedding = tf.keras.layers.Embedding(vocab_size, embedding_dim)
        self.gru = tf.keras.layers.GRU(self.units,
                                      return_sequences=True,
                                      return_state=True,
                                      recurrent_initializer='glorot_uniform')
        self.fc1 = tf.keras.layers.Dense(self.units)
        self.fc2 = tf.keras.layers.Dense(vocab_size)

        self.attention = BahdanauAttention(self.units)
```

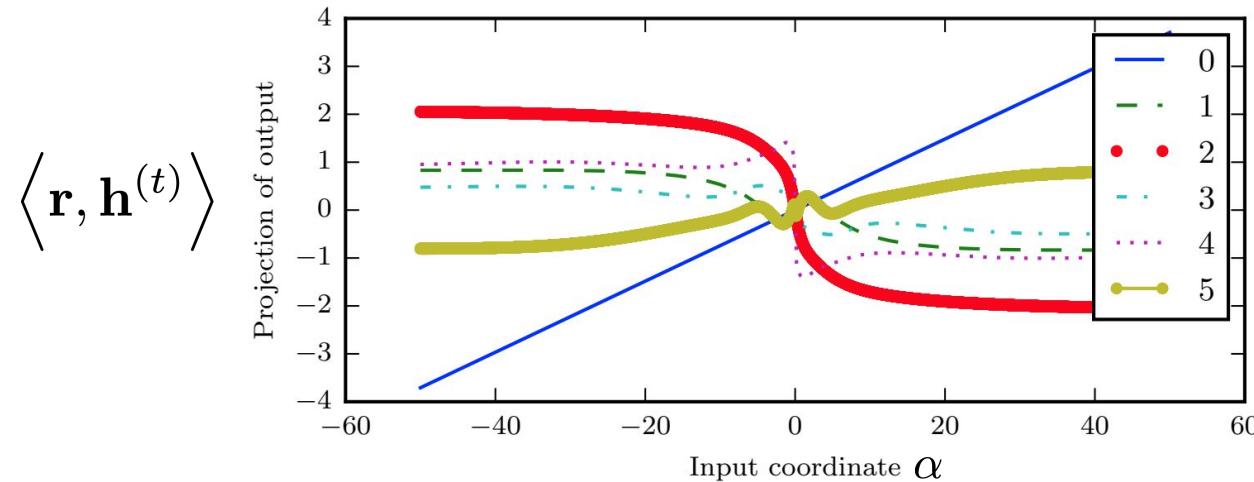
# Difficulty of training RNNs

- Hard to retain long term dependencies in the state vector



# Difficulty of training RNNs

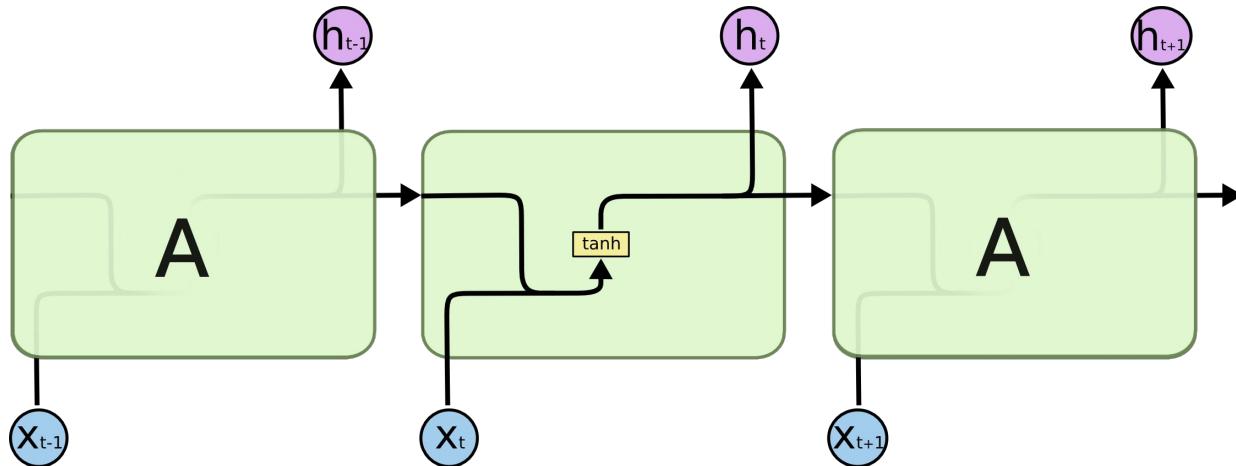
- Vanishing/exploding gradients



$$\mathbf{h}^{(t)} = \tanh(W\mathbf{h}^{(t-1)} + w_0)$$

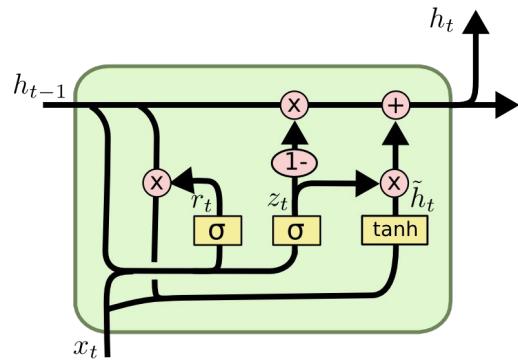
$$\mathbf{h}^{(0)} = \alpha \bar{\mathbf{r}}$$

# Simple RNN unit

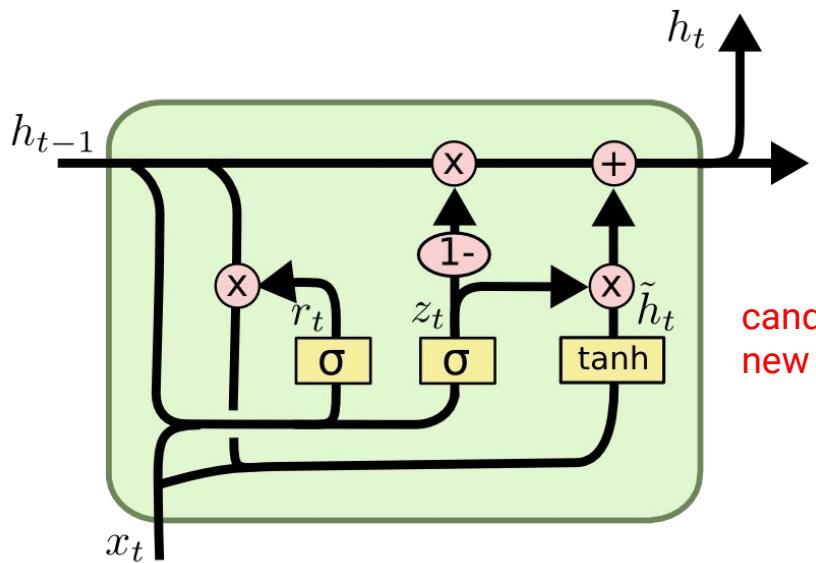


$$\mathbf{h}^{(t)} = \tanh(W\mathbf{h}^{(t-1)} + U\mathbf{x}^{(t)} + w_0)$$

# Gated Recurrent Unit (GRU)



# Gated Recurrent Unit (GRU)



$$z_t = \sigma (W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma (W_r \cdot [h_{t-1}, x_t])$$

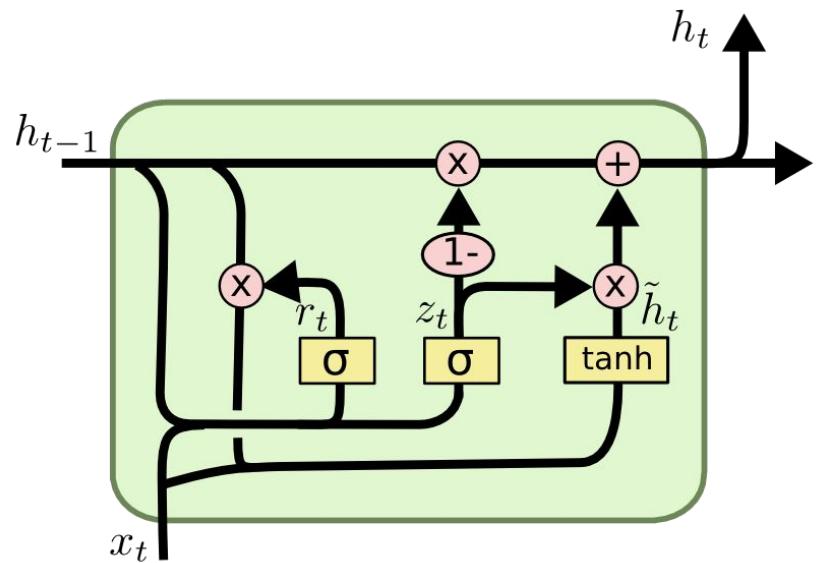
candidate  
new state

$$\tilde{h}_t = \tanh (W \cdot [r_t * h_{t-1}, x_t])$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

recall the skip connections in the  
ResNet!

# Gated Recurrent Unit (GRU)



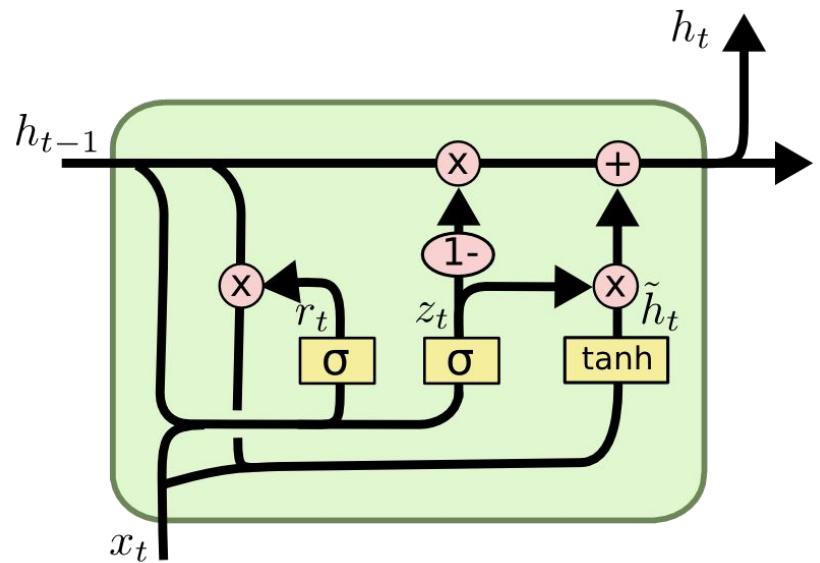
$$z_t = \sigma (W_z \cdot [h_{t-1}, x_t]) \quad \text{update gate}$$

$$r_t = \sigma (W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh (W \cdot [r_t * h_{t-1}, x_t])$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

# Gated Recurrent Unit (GRU)



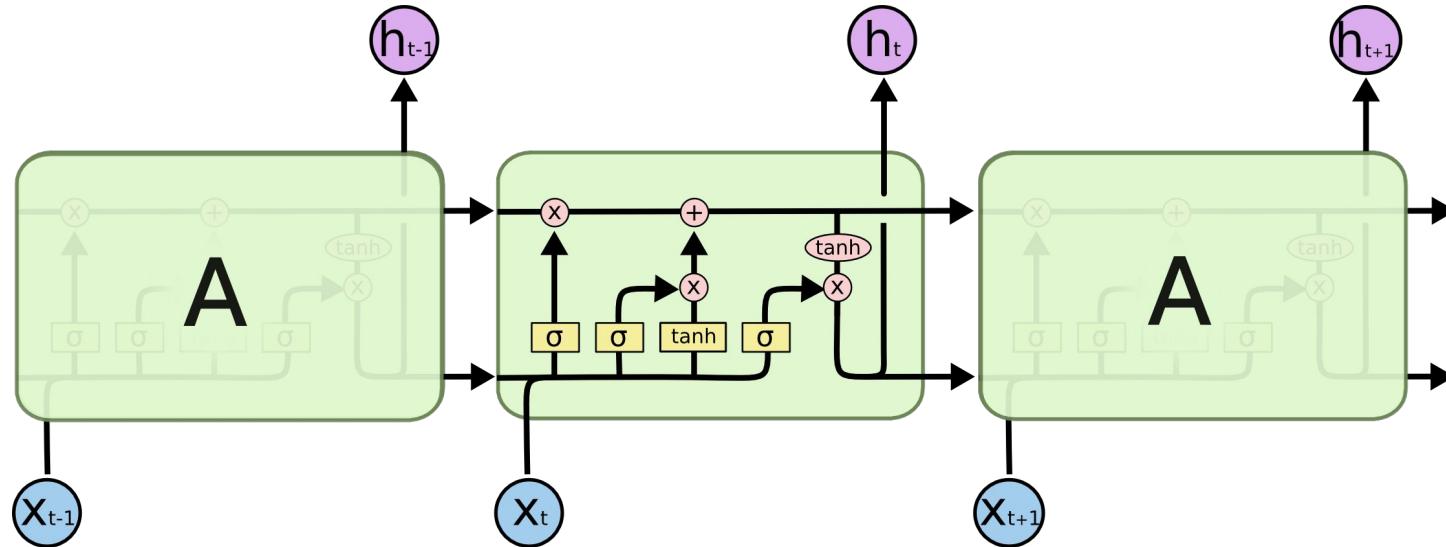
$$z_t = \sigma (W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma (W_r \cdot [h_{t-1}, x_t]) \text{ reset gate}$$

$$\tilde{h}_t = \tanh (W \cdot [r_t * h_{t-1}, x_t])$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

# Long Short-term Memory (LSTM)



[Hochreiter&Schmidhuber, Long short-term Memory, Neural Computation, 1997](#)

# Visualizing the hidden state of an RNN

- Character level RNN with LSTM units trained on a large corpora of text (wikipedia)
  - output is a single character
  - RNN has to learn to form words, punctuation and mark-up
- <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

# Visualizing RNN state

activation of one of the dimensions of  $\mathbf{h}^{(t)}$ , green means very excited and blue not excited

# Visualizing RNN state

http://www.ynetnews.com/ ] English-language website of Israel's largest English-language news site of Israel is sing d : xne.waea..awatoa. s &ntiaca-sardeelh oan t bisanfanreif ' aatd mw-2 ♦ piiisoesssis./ern.c](dceen epeasaaiki ieel edh,irthraonse , cose dr. <.ahb-nptwt.xi gh/ma) Tvdryzi couedlsu:tha-oo tu,stuif lvepery stp, tcoa2drulwoclensr] p. llvaod,,eyt c-n dm-oibuvb] bb imsulta tlybn

gest newspaper ' [[ Yedioth Ahronoth ] ] ' ' Hebrew-language period et aawsaperso' [[ Tel t i (feanemti) ] ] ' ' [ terrewsleanguage:arosodi ir scoe ena itThAoainnh Srmuw] ey s [ 'ineia'siwdde'hsolrifr: us..setlgor s.asat Careeg' aClrisz] ie': :, #: T Aaaaat Baseeil o'ianfvl - tuaevrtid, tBAmSusyut]] Asaoigs] ], . . . sMBolous: Toua-n: d woapnu a, d, iiuiticp.] (ISvHvtusuiDnoegano ., ] : { CCuibohCybksls:r-epcnts

icals: ' ' \* ' [[ Globes ] ] ' [ http://www.globes.co.il/ ] business data cal : ' ' \* ' [ Taabah ] ' ' ([ http://www.buobal.comun/sA -ytiness aet s tl' [ hAeovelt sahad : xge. waoir. rtoae. el. iT &ai eg eooy tt' ' ' &[ &&mCoerone': :, i' odw., : niiisaue. eni /omlcC. (eftgir iiu a'n:, C: &: #\*: afDrusu] l, . omel p<, dha; deuoot/ihncsifs, urhost, tun nk i <]: &11s TGuitrsi, : bacmr-xtprob-gresislerlnafad] losptad, ifrm

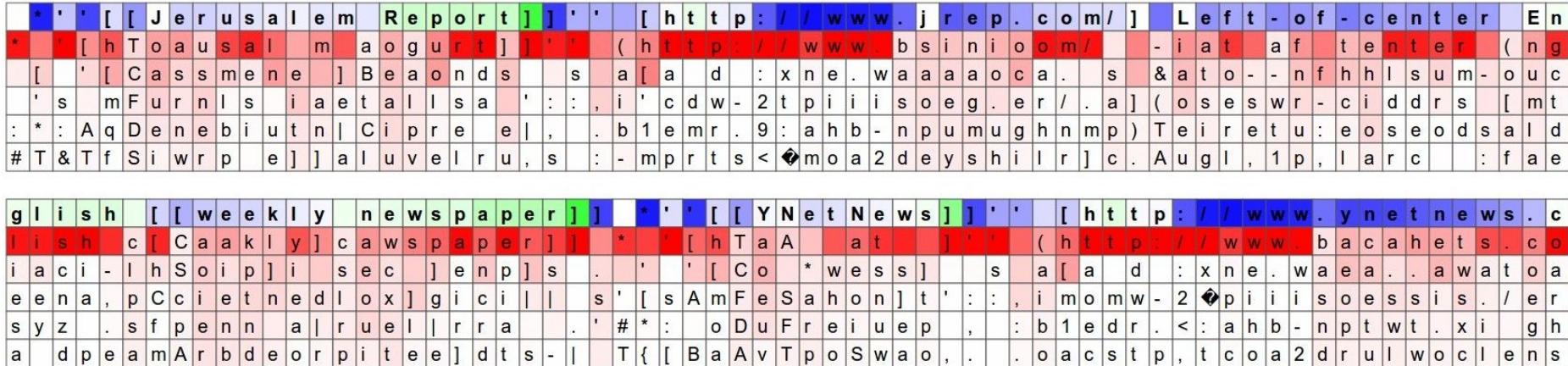
ily \* ' [[ Haaretz | Ha'aretz ] ] ' [ http://www.haaretz.co.il/ ] Relativ ly \* ' [[ Terrdn Ferantah ] ] ' ' ([ http://www.bonmdst.comun/s -esateoi re ' hAilnntteHalsrcnol' sahad : xne. waamrt dheoh. ol. c &opinive ki. : \* sCOSanlt hitim'li] e : , imcdw-2 ♦ phiiserdit. ina/cmfi. (afl cana ds-! [t BTCommgd] ] Won a ae, : baerr. <taib-dulcnnclarne] liceysto nds#&: GI Duvccsaosucltel] zl, : o'omt], : eoa2nivfsrooeiunala) uvvro

# Visualizing RNN state

The image shows a 4x100 grid of colored squares, where each row represents a different piece of text and each column represents a time step. The colors range from red (low activation) to green (high activation). The text pieces are:

- Line 1: [[Jerusalem Report]] [[http://www.jrep.com/]] Left-of-center English  
[ [hToausalmaogurt]] [[http://www.bsinioom/]] -iat af tenter (ng  
[ [Cassmene] Beaonds s a[d : xne. waaaaoca. s &ato- -nfhhls sum- ouc  
's mFurnls iaetal ls a': :, i'cdw- 2t piiiisoeg. er/. a] (oseswr- ciddrs [ mt  
: \* : AqDenebiutn| Cipre el, . b1emr. 9:ahb- npumughnmp) Teiretu: eoseodsald  
#T&TfSiwrpe] jaluvvelru, s : -mprt s< moa2deyshilr] c. Augl, 1p, larc :fae
- Line 2: glish [[weekly newspaper]] [[YNet News]] [[http://www.ynetnews.co  
lish c [Caakly] cawspaper]] [[hTaA at]] [[http://www.bacahets.co  
iaci - IhSoip] i sec ] enp] s . ' ' [Co \*wess] s a[a d : xne. waea.. awatoa  
ee na, pCciet nedlox] gicill s' [sAmFeSahon] t' :, imomw- 2 piiiisoessis. /er  
syz . sf penn alruellrra . ' #\* : oDuFreiuep , : b1edr. <: ahb- nptwt. xi gh  
a dpeamArbdeorpitee] dt s- | T{ [BaAvTp oSwao, . . oacstp, tcoa2drulwoclens
- Line 3: om] English-language website of Israel's largest newspaper ' [[Yed  
m/ -xglish linguis gesairsite of tsraelis singet aawspaperso' [[Tel  
. s &ntiaca- sardeelh oan t bisan fanreif ' aat dir scoe ena iTTAoai  
n. c] (dceen epesaaiki ieel edh, irthraonse , coseus. setlgor s. asat Care  
/ ma) Tvdryzi couedlsu: tha- oo tu, stuif lvepery - tuaevrtid, tBAmSusy  
r] p. llvaod, eytc-n dm-oibuv s] bb imsulta tlybna, d, iiuiticp.] (ISvHvtu
- Line 4: ioth Ahronoth] [[ Hebrew-language periodicals: [[Globes] ]  
t i( feanemti] [[errewsle ngue: arosodical : [[Taaba] ]  
nnh Srmuw] ey s [ 'ineia' siwdde' hsolrifr: stl'  
eg' aClri sszz] ie' ::, #: TAaaaat Baseeil o'ianfvl tt' ' & [&&mCoer one' ::  
ut] Asaoigs] ], . . : sMBolous: Toua- n: d woapnuau'n:, C: & : #\* : afDrusu] l,  
suiednoegano ., ] : { CCui bohe Cybksls: r-epcnts nk i <] : & 11s T Guitrsi,

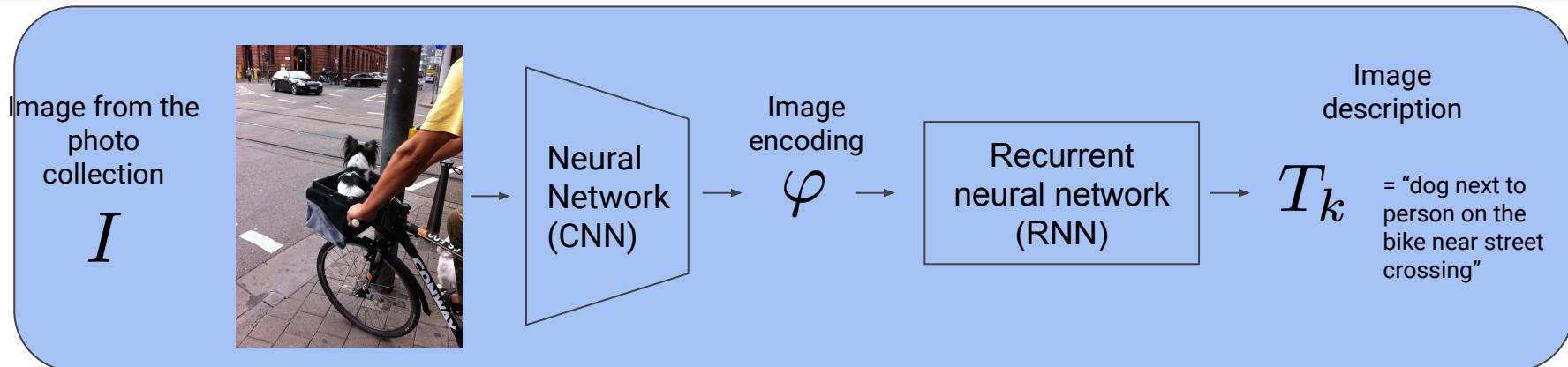
# Visualizing RNN state



# Visualizing RNN state

The figure displays a grid of colored squares representing the hidden state of an RNN processing a sequence of words. The columns represent time steps, and the rows represent different words or categories. The colors indicate which category each word belongs to, such as 'The Guardian' (green), 'Israel Insider' (green), 'http://www.israelinsider.com' (blue), 'The Soir dan' (red), 'Tn lael hde' (red), 'http://www.bmsacl.ng' (blue), 'GhAoioMrltrin' (red), 'Cs if iDatsi' (red), 'ahad :xne.wsnoonsitet' (red), 'BmSroiCannt' (red), 'amDmaeit' (red), 'tans':.i'ol.' (red), 'gi;asce..l..e' (red), 'GeuyLusInsca' (red), 'CTAoonfe' (red), 'c s|,,o1ecw-x' (red), 'po bigck.rtaado' (red), 'DTCaarGeadanol' (red), 'STOSTtul' (red), 'ceo]' (red), 'i.s]' (red), 'bmpdr,<tmhd-nelliefsssa' (red). The grid shows how the RNN's state transitions between different semantic categories over the sequence of words.

# Roadmap



Define similarity function. Order images according to similarity to the query.

$Q$

Query: "person walking with a dog on the beach"

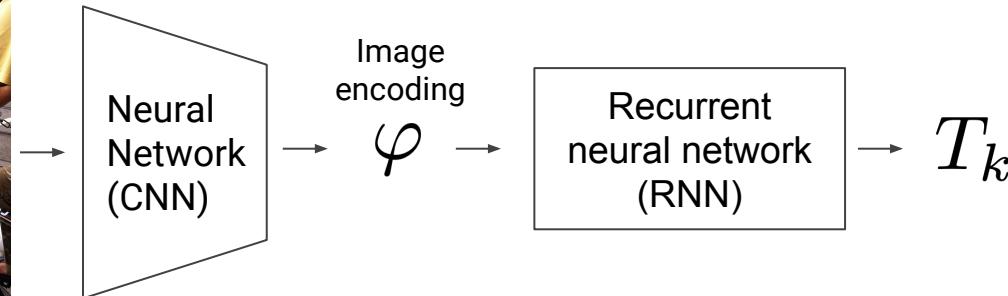


$$\text{sim}(Q, T_1) > \text{sim}(Q, T_2)$$

# Image captioning model

Image from the  
photo collection

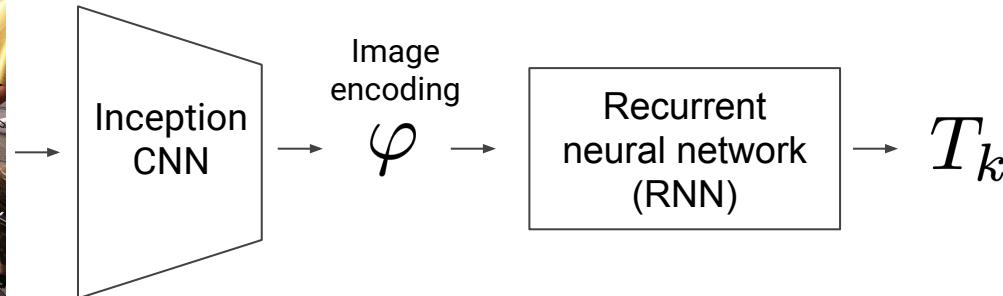
$I$



# Image captioning model

Image from the  
photo collection

$I$



# Image captioning model

Image from the  
photo collection

$I$

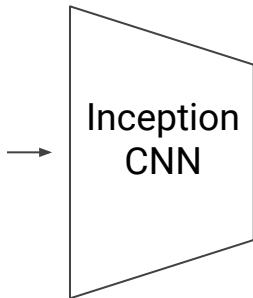
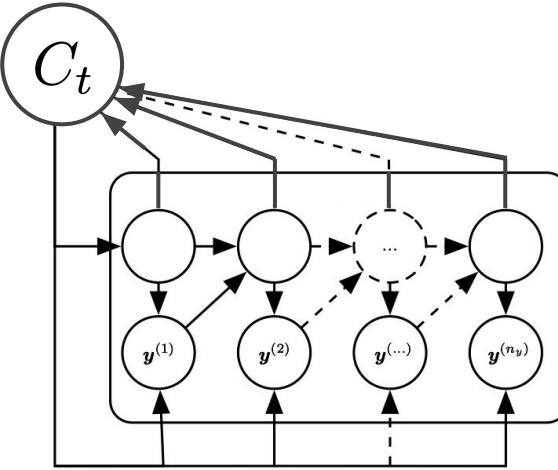


Image  
encoding  
 $\varphi$



# Image captioning model

Image from the  
photo collection

$I$

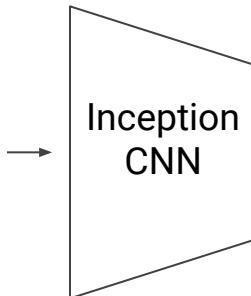
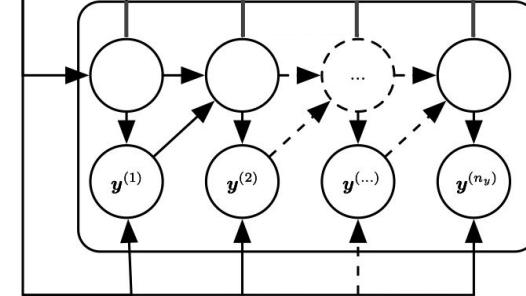


Image  
encoding

$\varphi$

$C_t$

attention model



# Image captioning model

Image from the  
photo collection

$I$

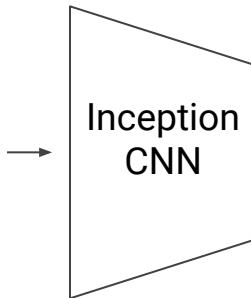
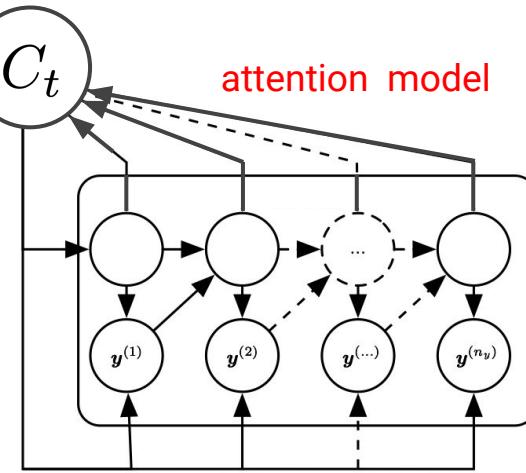


Image  
encoding  
 $\varphi$

context vector

$C_t$

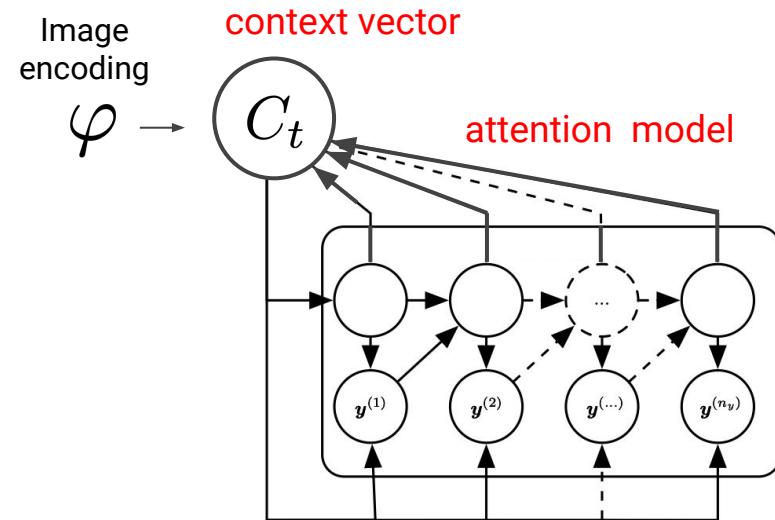
attention model



# Attention model

$$e_{tj} = v_a^T \tanh(W_a \mathbf{h}^{(t-1)} + U_a \varphi_j)$$

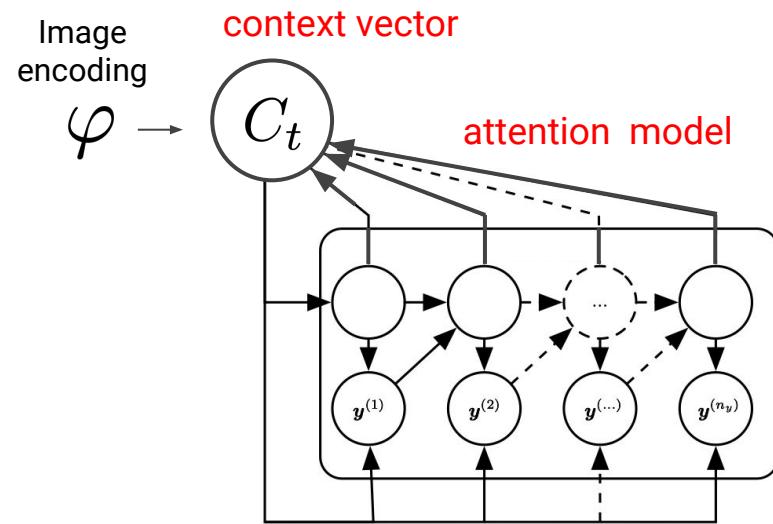
image features at location j  
↑



# Attention model

$$e_{tj} = v_a^T \tanh(W_a \mathbf{h}^{(t-1)} + U_a \varphi_j)$$

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_k \exp(e_{tk})}$$



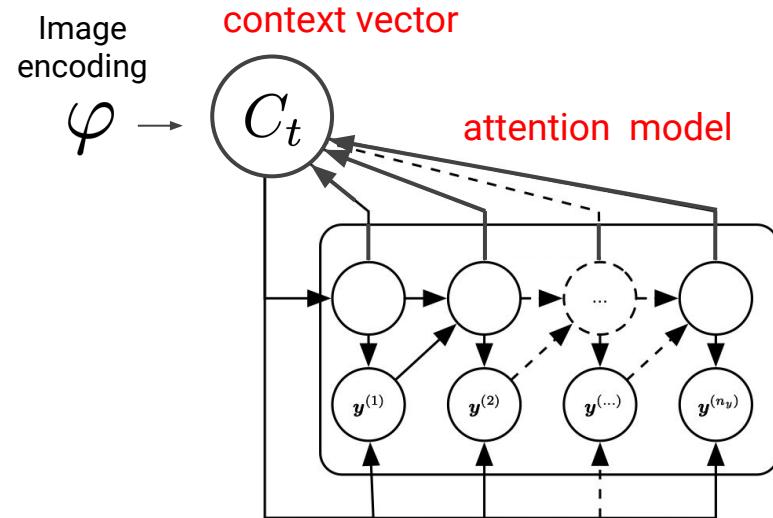
# Attention model

$$e_{tj} = v_a^T \tanh(W_a \mathbf{h}^{(t-1)} + U_a \varphi_j)$$

image features at location j

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_k \exp(e_{tk})}$$

$$\mathbf{C}_t = \sum_j \alpha_{tj} \varphi_j$$



Bahdanau et al, Neural Machine Translation by Jointly Learning to Align and Translate, ICLR'15

# Recurrent neural networks

Image captioning model from:

[https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/text/image\\_captioning.ipynb](https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/text/image_captioning.ipynb)

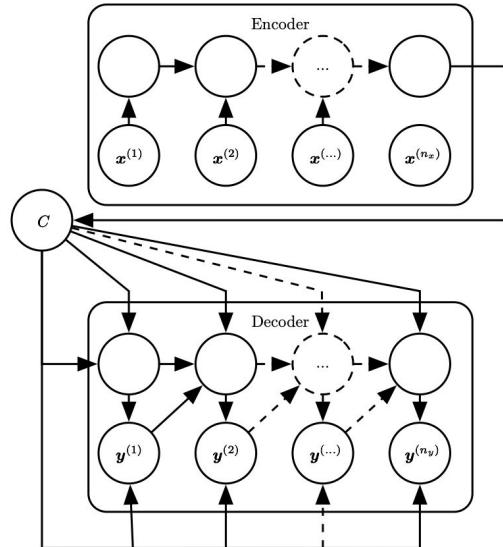
```
class RNN_Decoder(tf.keras.Model):
    def __init__(self, embedding_dim, units, vocab_size):
        super(RNN_Decoder, self).__init__()
        self.units = units

        self.embedding = tf.keras.layers.Embedding(vocab_size, embedding_dim)
        self.gru = tf.keras.layers.GRU(self.units,
                                      return_sequences=True,
                                      return_state=True,
                                      recurrent_initializer='glorot_uniform')
        self.fc1 = tf.keras.layers.Dense(self.units)
        self.fc2 = tf.keras.layers.Dense(vocab_size)

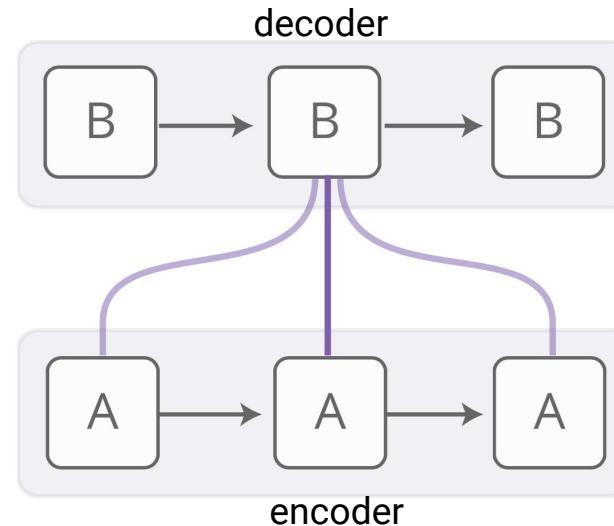
        self.attention = BahdanauAttention(self.units)
```

Bahdanau et al, [Neural Machine Translation by Jointly Learning to Align and Translate](#), ICLR'15

# Machine translation with attention



Machine translation as a  
sequence-to-sequence model



Machine translation with  
attention

# Machine translation with attention

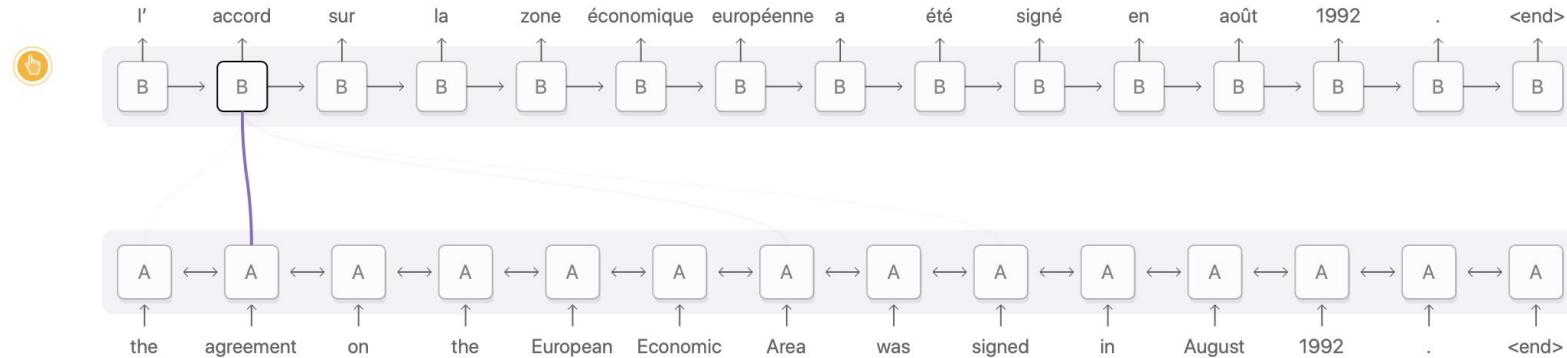


Diagram derived from Fig. 3 of Bahdanau, et al. 2014

<https://distill.pub/2016/augmented-rnns/#attentional-interfaces>

# Machine translation with attention

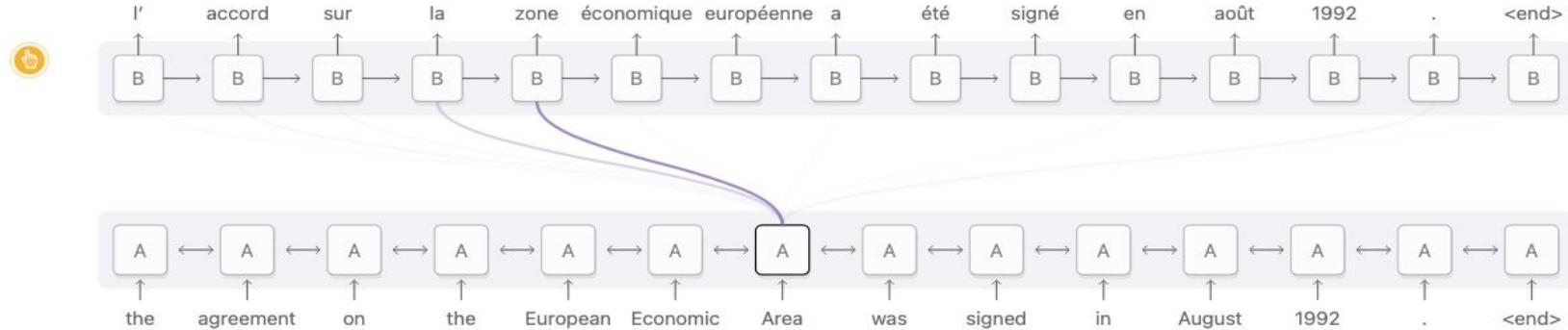


Diagram derived from Fig. 3 of [Bahdanau, et al. 2014](#)

<https://distill.pub/2016/augmented-rnns/#attentional-interfaces>

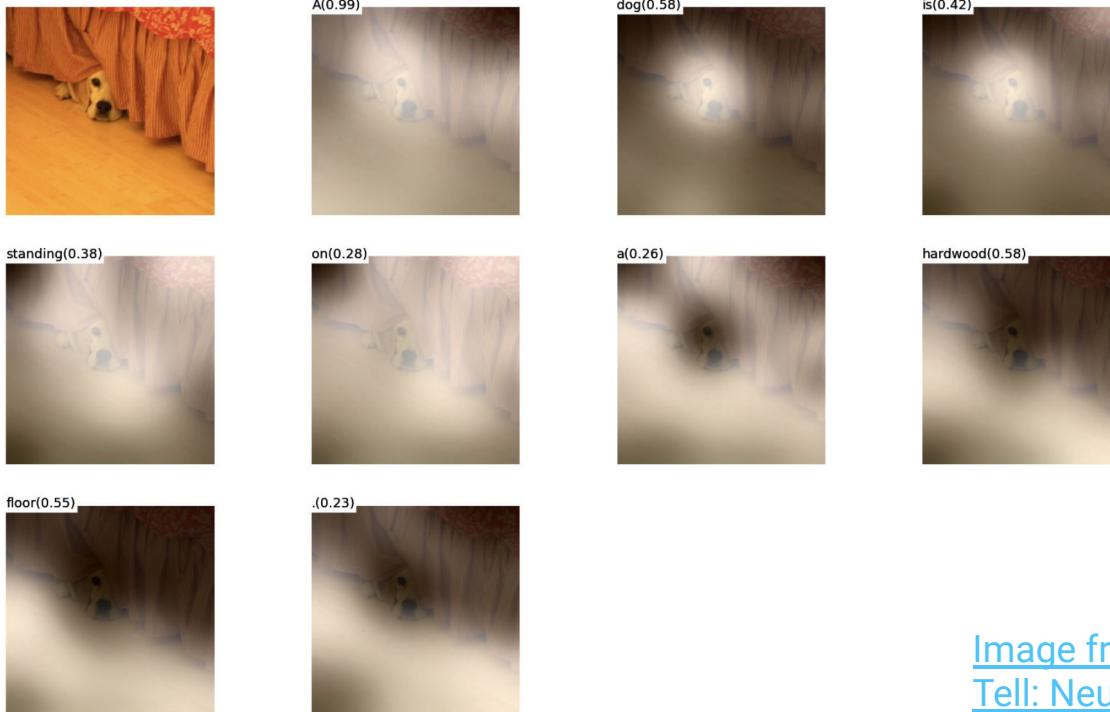
# Image captioning with attention



[Image from “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”](#)

(b) A woman is throwing a frisbee in a park.

# Image captioning with attention



(b) A dog is standing on a hardwood floor.

[Image from “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”](#)

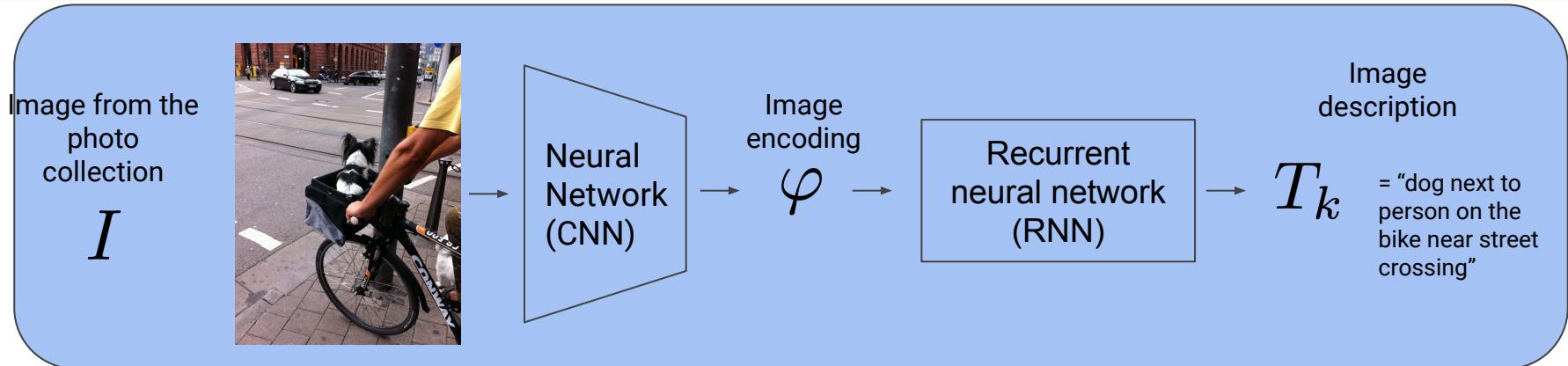
# Image captioning with attention



(b) A woman holding a clock in her hand.

[Image from “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”](#)

# Congratulations! We did it!



$Q$

Query: "person walking with a dog on the beach"

Define similarity function. Order images according to similarity to the query.



$$\text{sim}(Q, T_1) > \text{sim}(Q, T_2)$$

# Thank you for your attention!

# Take a quiz!