CMPT 983

Grounded Natural Language Understanding

January 18, 2021

Review of deep learning models and word embeddings

Today

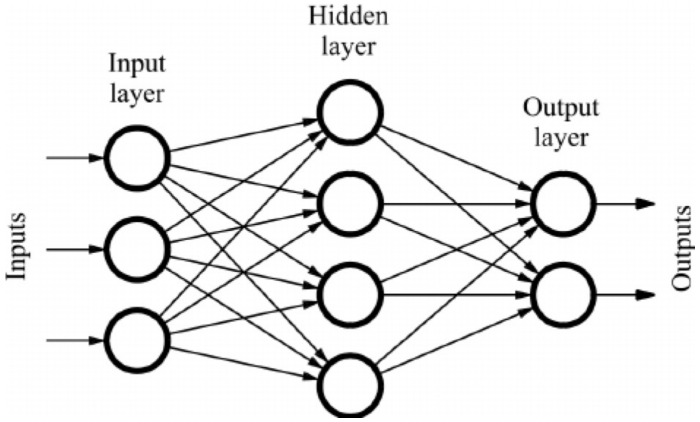
- Review of basic deep learning building blocks
  - CNNs
  - RNNs

- Word embeddings

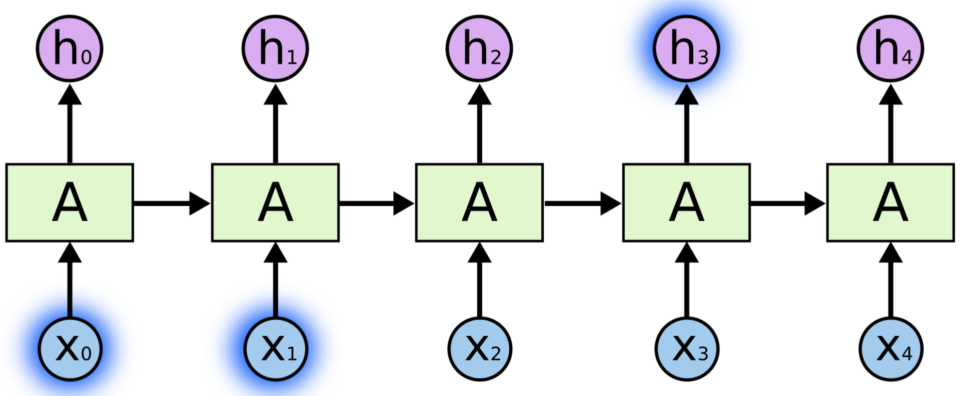
Deep learning models

Neural network architectures

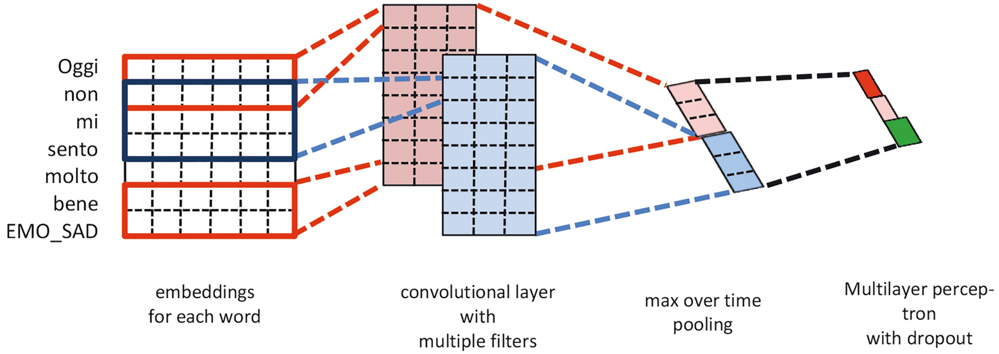
Feed-forward NNs



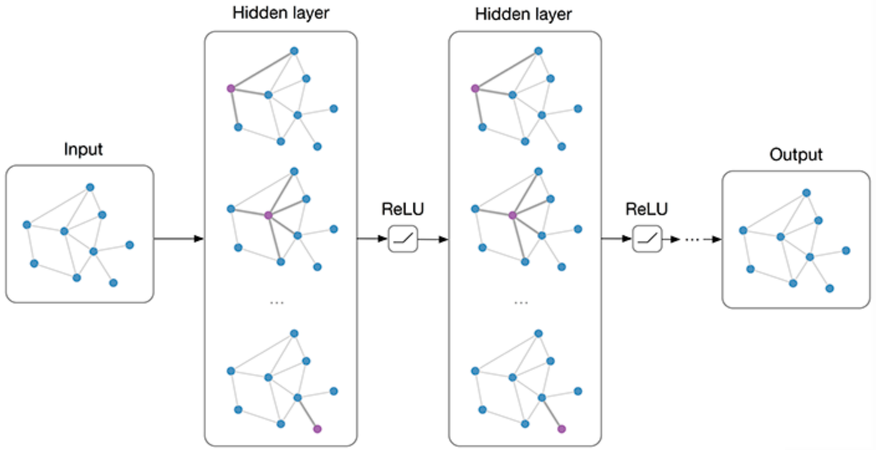
Recurrent networks (RNNs)



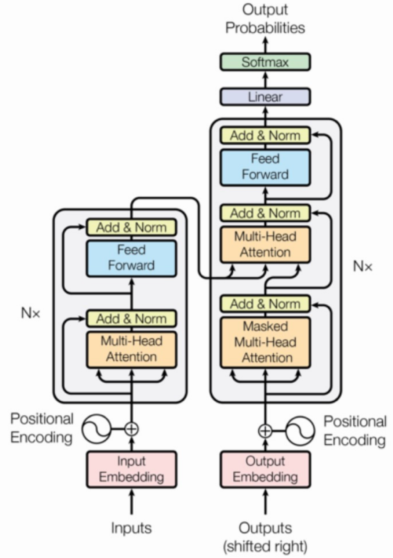
Convolution networks (CNNs)



Graph NNs



Transformers



- All network architectures can be used to model **images**, **text**, **3D representations**, etc.

- Traditionally:
  - CNNs for images – scale/translation invariance
  - RNNs for sequences (text)
  - Transformers were introduced for machine translation
    - Now used for images and 3D shapes as well

Modelling Images

Modelling Images

I

Image

Image Encoder

V

Useful Visual Feature

Modelling Images

Convolutional Neural Networks

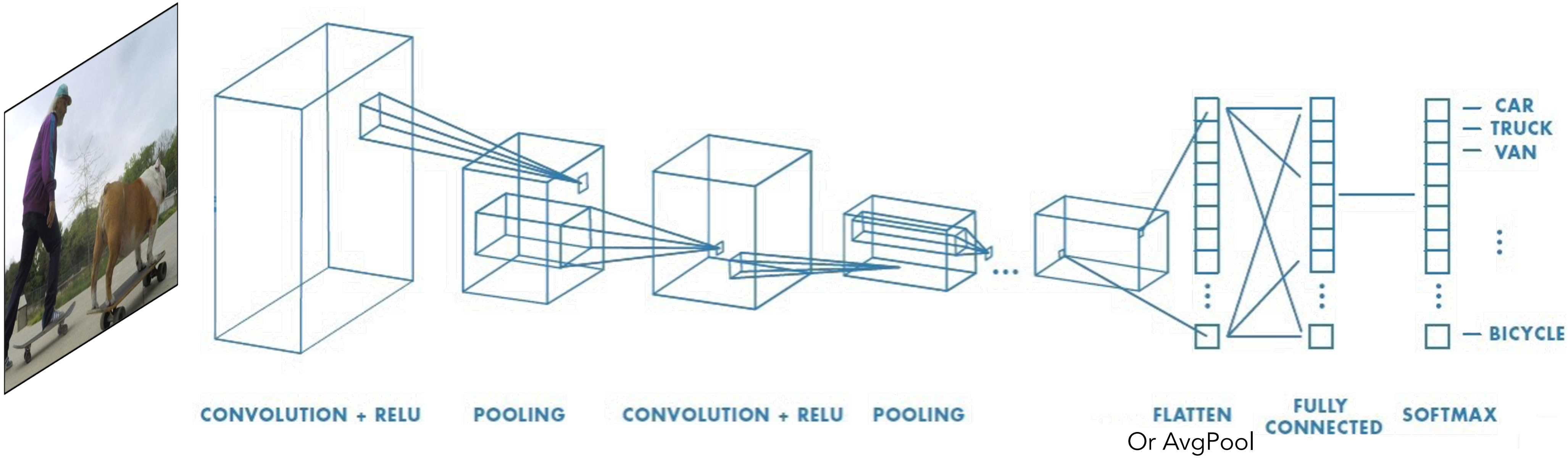
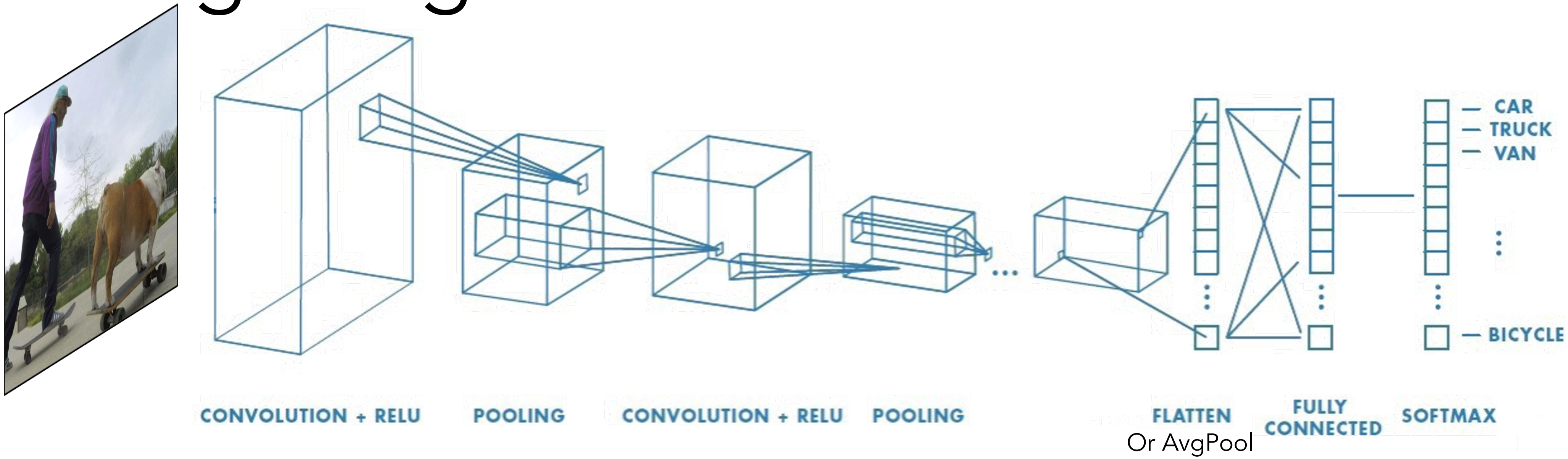


Image Credit: MathWorks

Modelling Images



Modelling Images

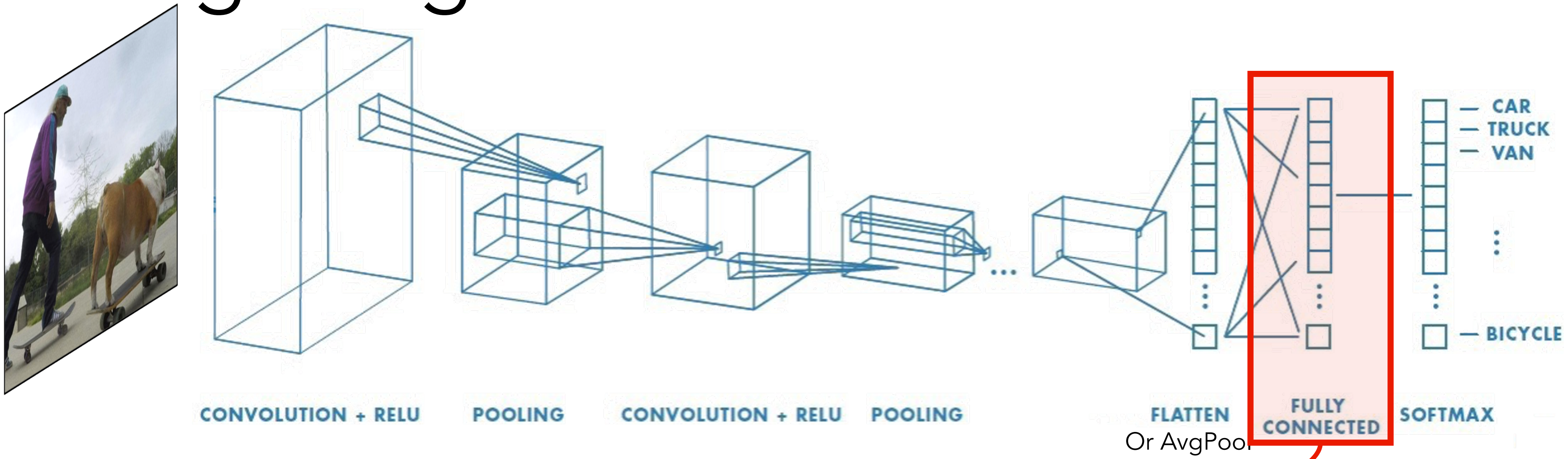
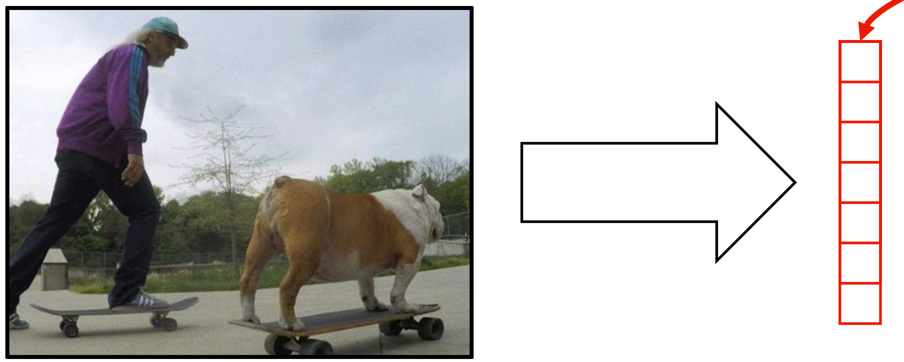


Image Level Feature

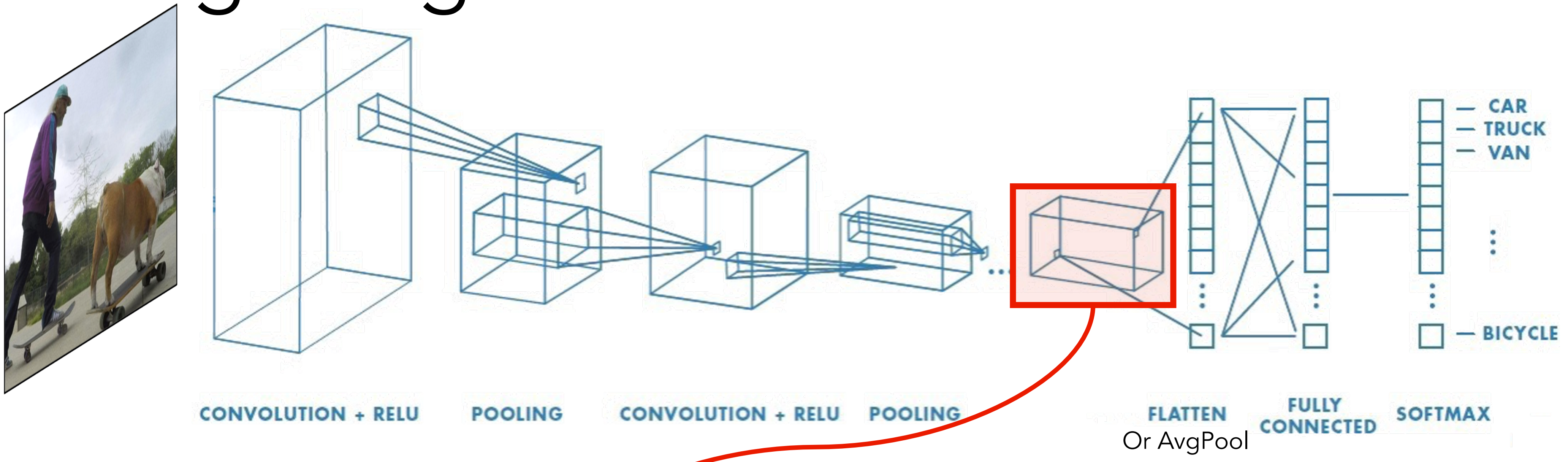


$$V \in \mathbb{R}^{1 \times d}$$

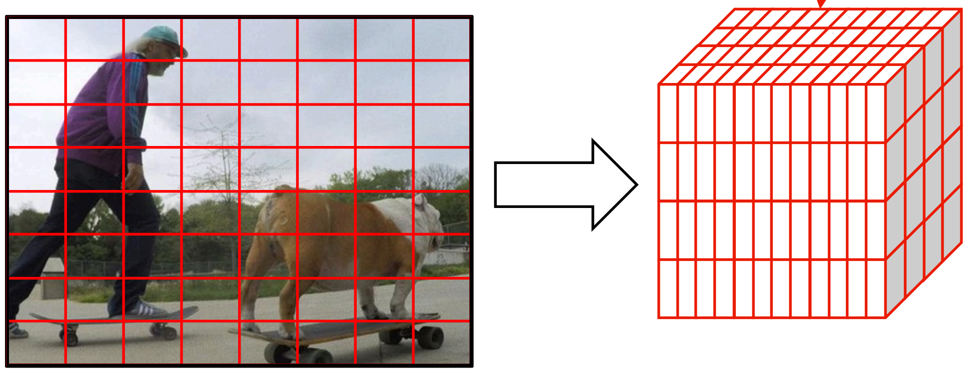
- No spatial information
- Highly compressed

Slide credit: Stefan Lee

Modelling Images



Spatial Image Features

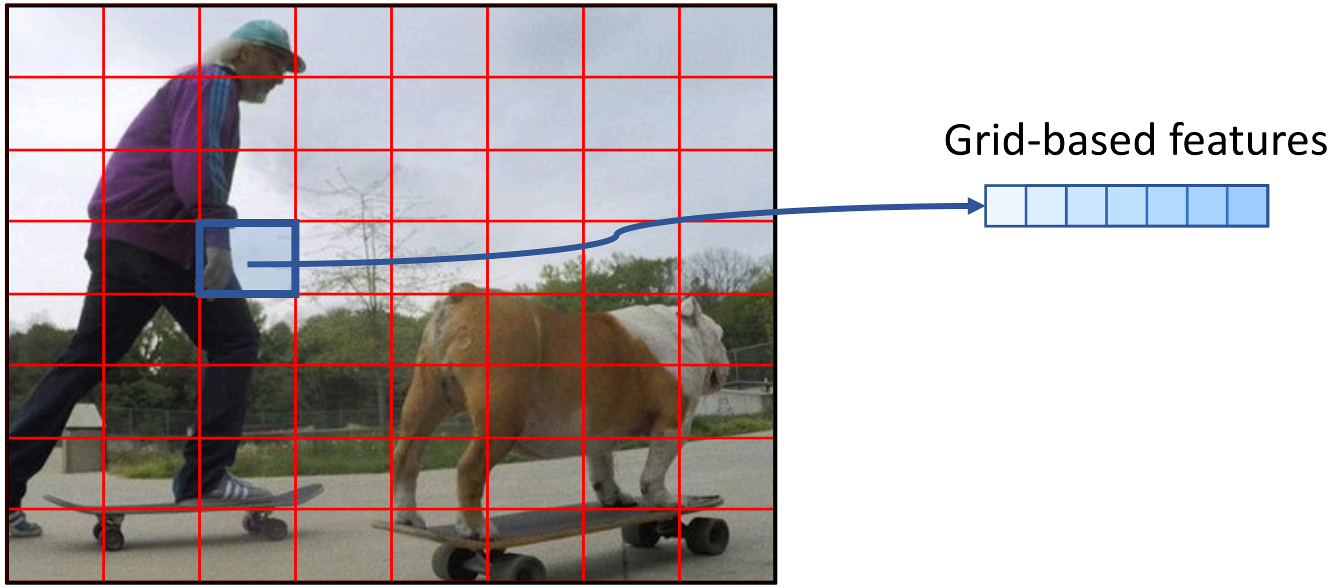


$V \in \mathbb{R}^{w \times h \times d}$

- Feature vector per grid cell
- Captures some spatial info
- Uniform grid…

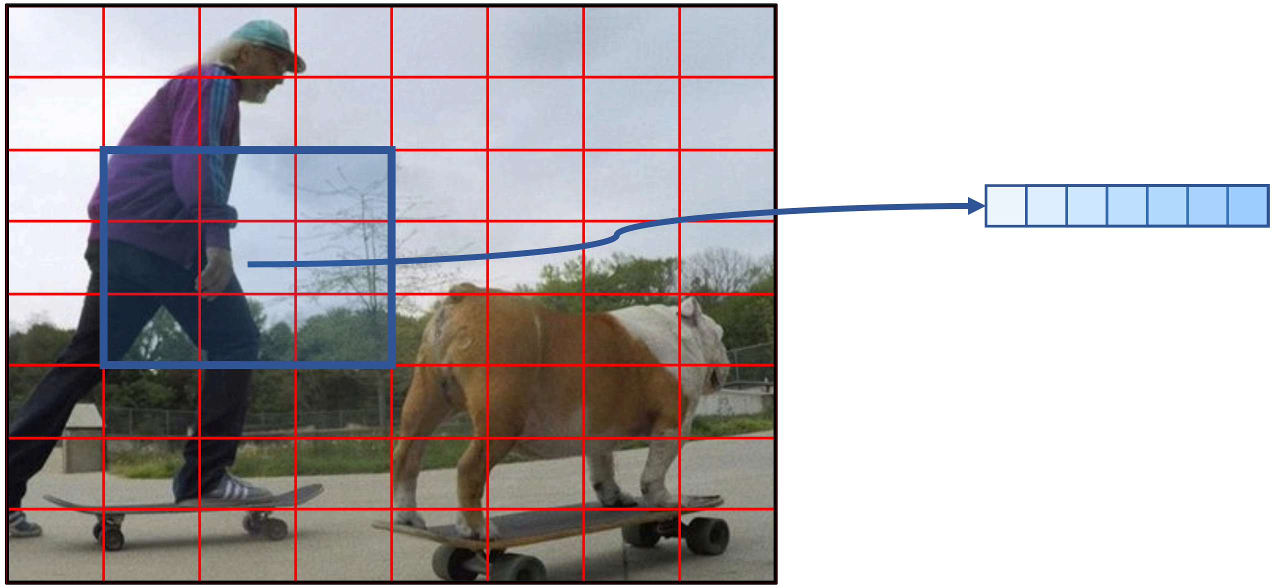
Slide credit: Stefan Lee

Modelling Images



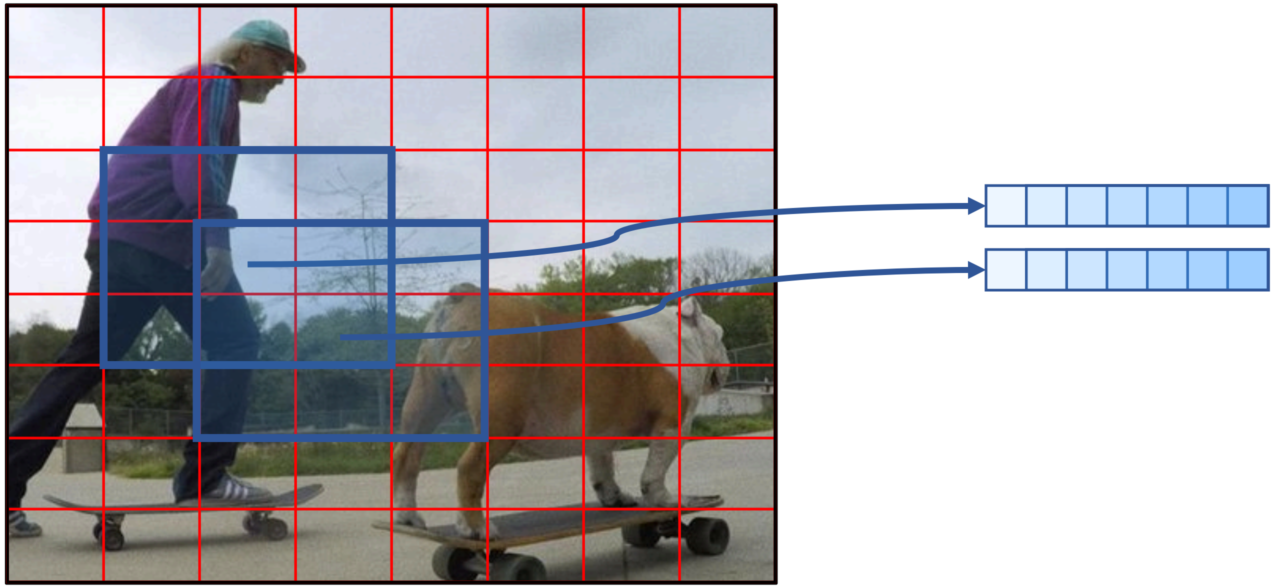
Slide credit: Stefan Lee

Modelling Images



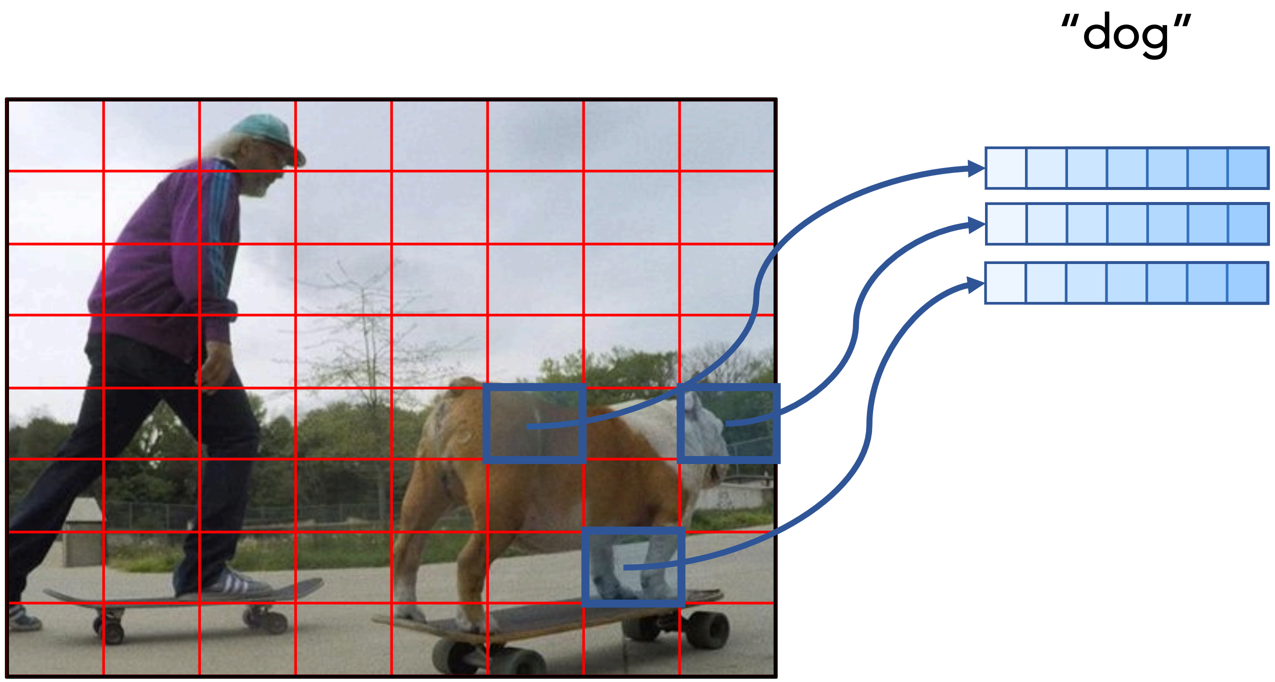
*Considering receptive field it is actually much more like

Modelling Images



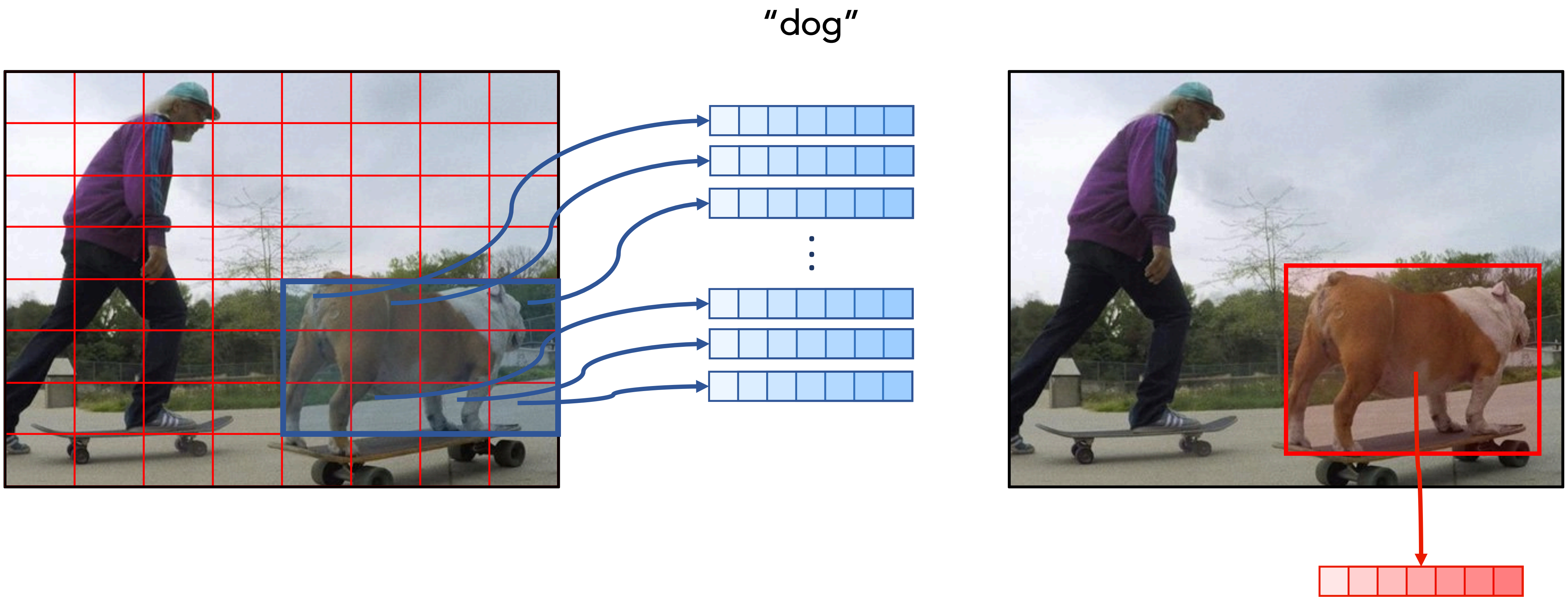
*Considering receptive field it is actually much more like

Modelling Images



Slide credit: Stefan Lee

Modelling Images



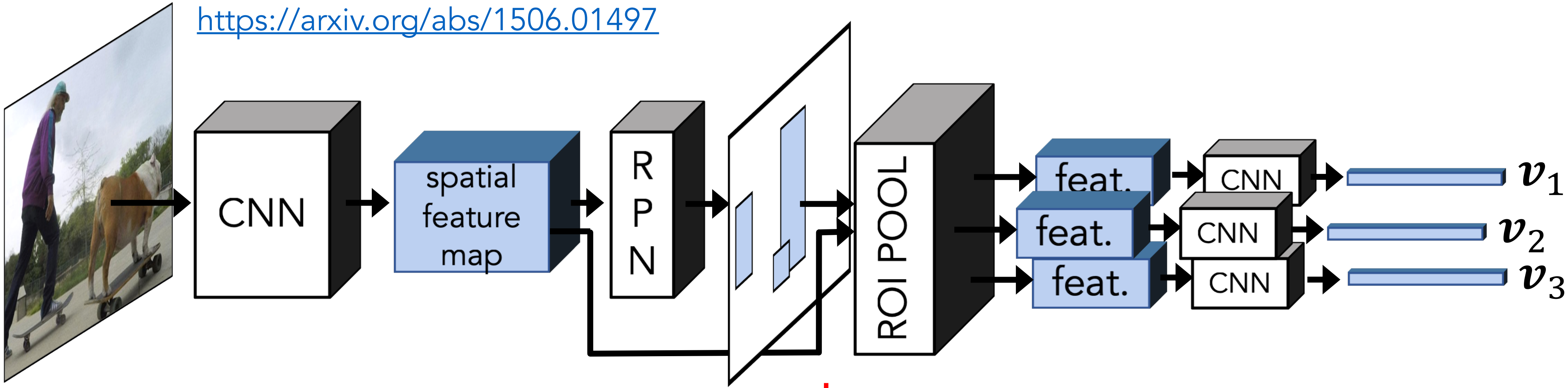
Slide credit: Stefan Lee

Modelling Images

Idea: Switch to object detection models as the backbone for image representation

- Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering arxiv.org/abs/1707.07998

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
https://arxiv.org/abs/1506.01497



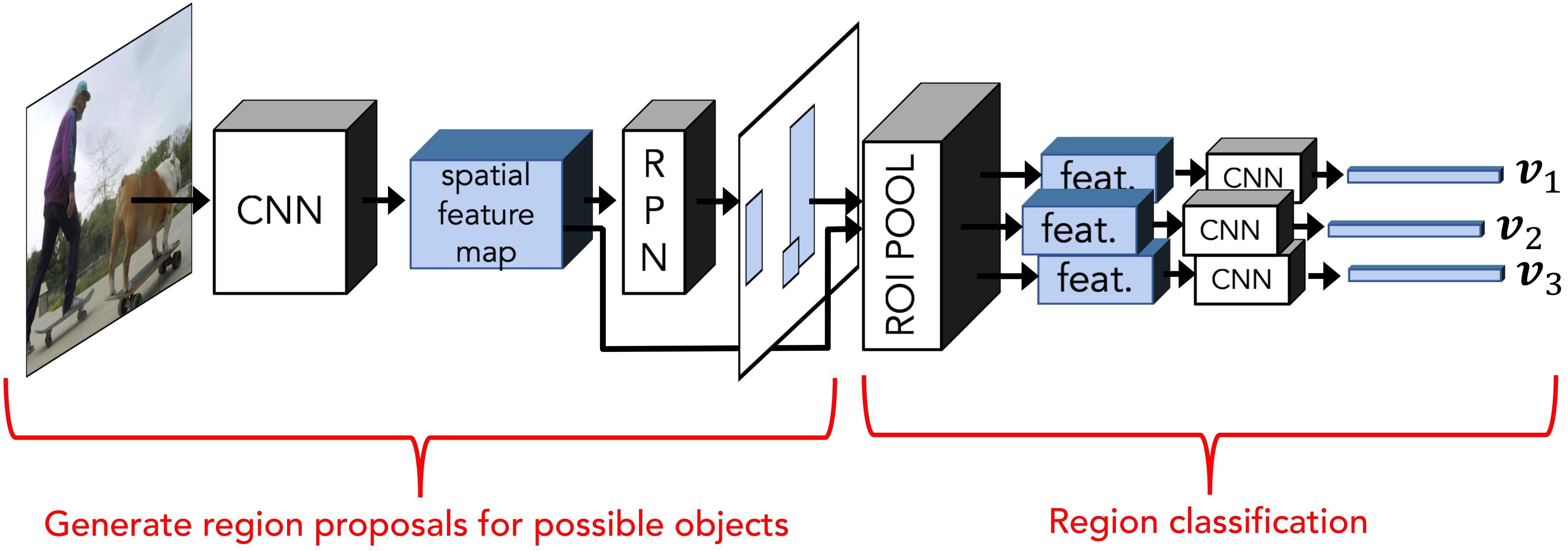
Generate region proposals for possible objects

Image Credit: Peter Anderson Slide credit: Stefan Lee

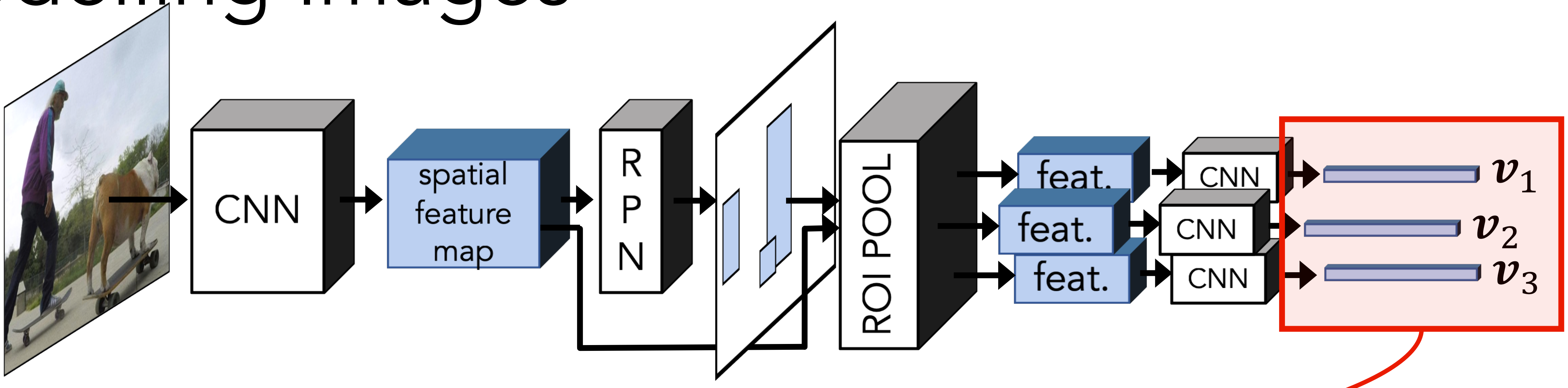
Modelling Images

Idea: Switch to object detection models as the backbone for image representation

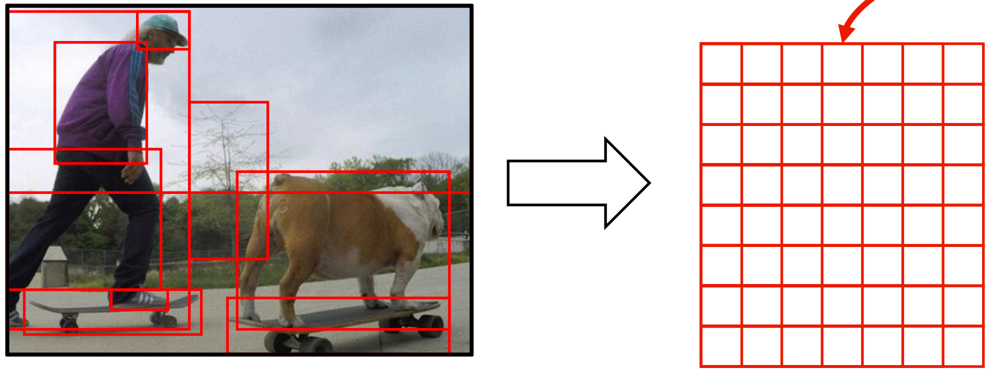
- Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering arxiv.org/abs/1707.07998



Modelling Images



Object-Centric Image Features



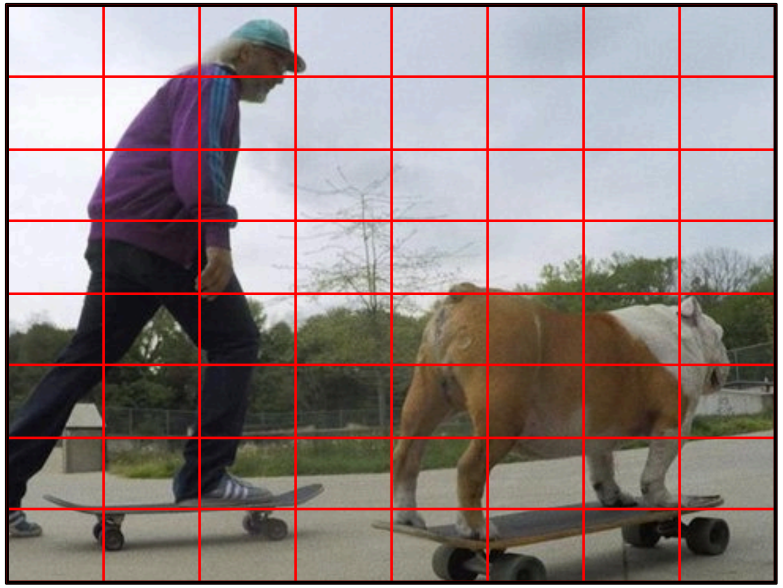
$$V \in \mathbb{R}^{k \times d}$$

- Feature vector per bounding box
- Spatial features can be added
- Object-centric

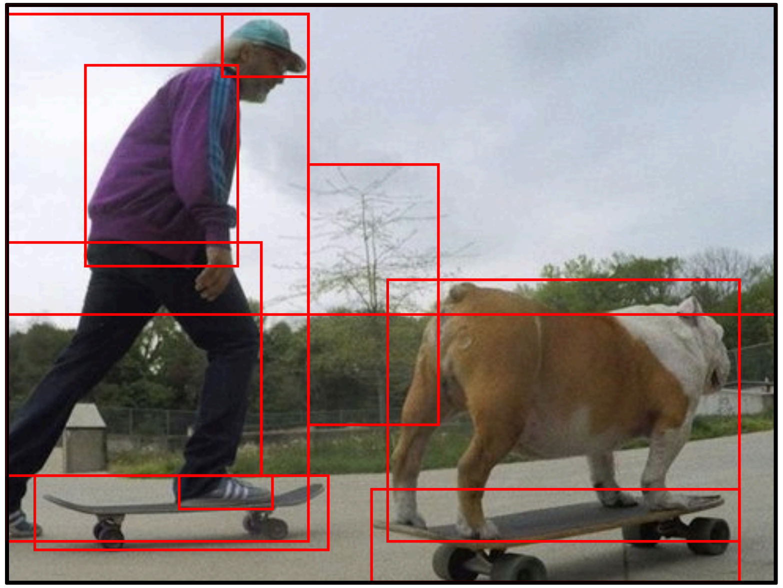
Modelling Images

Image Level Features

Spatial / Conv Features



Detection Features



ResNet 101
Trained on ImageNet

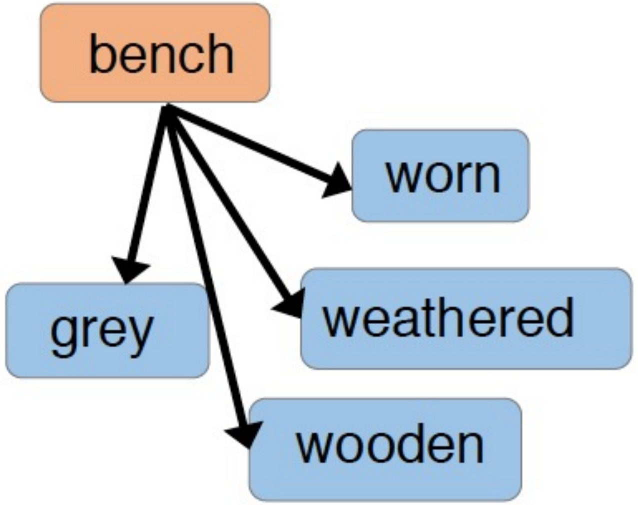
FasterRCNN - ResNet 101
Trained on Visual Genome

These are almost never fine-tuned for downstream tasks in vision-and-language.

Slide credit: Stefan Lee

Modelling Images: Pretraining

ResNet 101 Pre-training on ImageNet
- 1000 object classes (many fine-grained)



Faster R-CNN Pre-training on Visual Genome
- 1600 object classes
- 400 attribute classes

Slide Credit: Peter Anderson

Modelling Sequences

Modelling Sequences

Recurrent Neural Networks

- Ideal for processing sequential data containing possibly long-term dependencies.
- Various implementations (e.g. simple RNN, LSTM, GRU) expose the same API

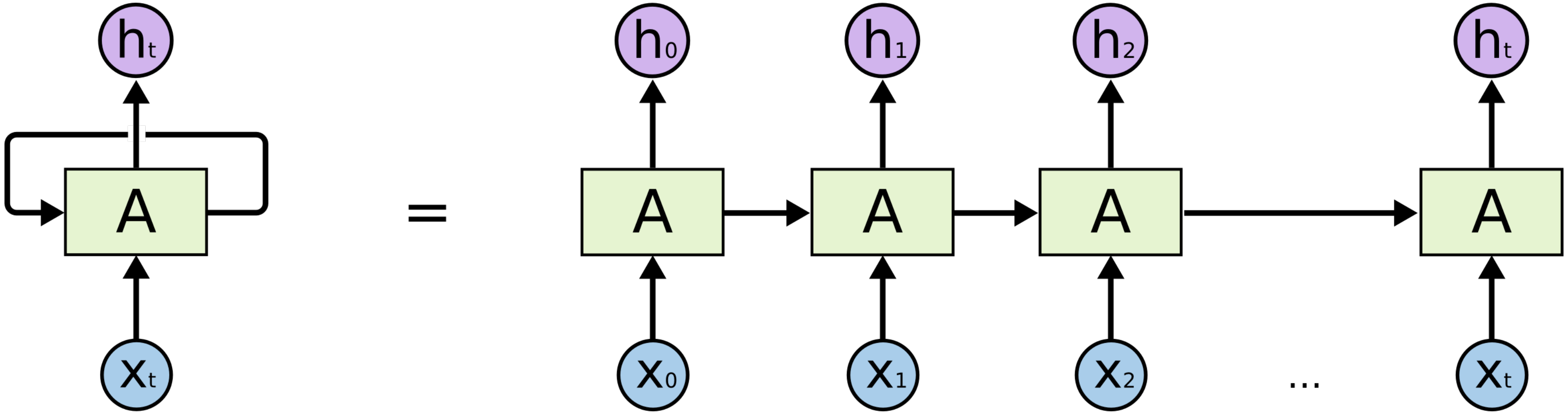


Image Credit: Christopher Olah

Modelling Sequences

Recurrent Neural Networks

- Ideal for processing sequential data containing possibly long-term dependencies.
- Various implementations (e.g. simple RNN, LSTM, GRU) expose the same API

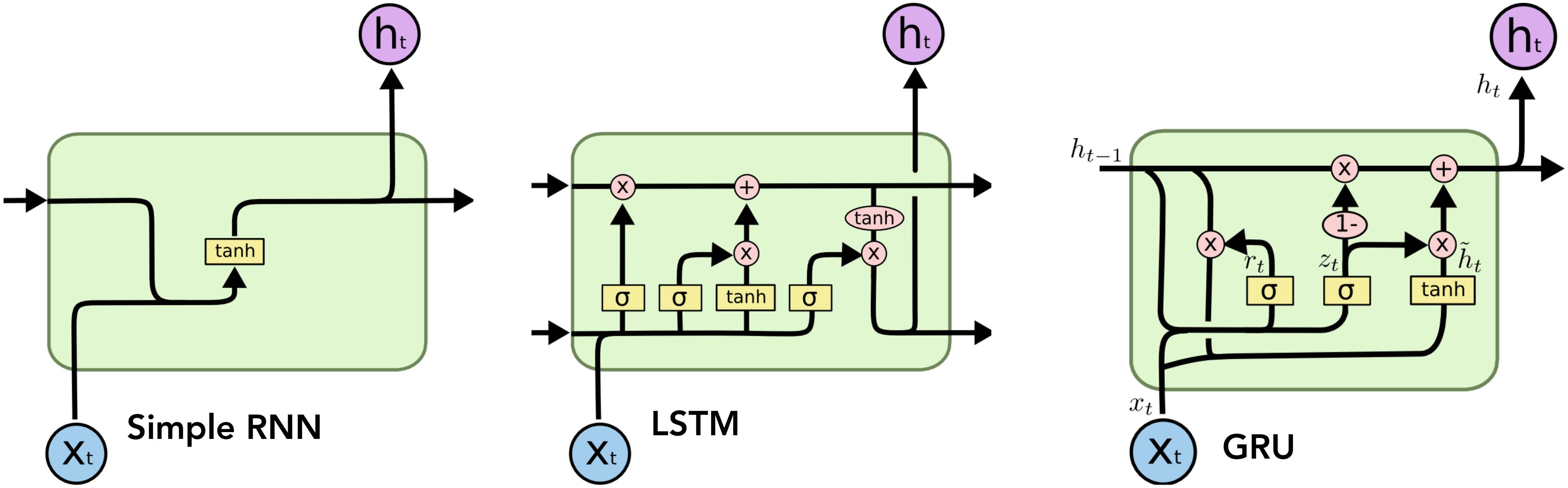
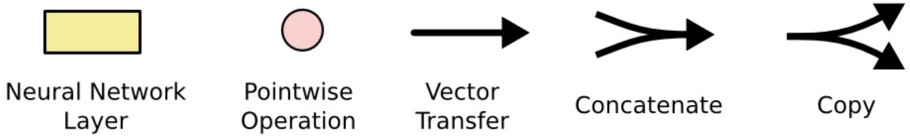


Image Credit: Christopher Olah

Slide credit: Stefan Lee



Modelling Sequences

Recurrent Neural Networks

- Ideal for processing sequential data containing possibly long-term dependencies.
- Various implementations (e.g. simple RNN, LSTM, GRU) expose the same API

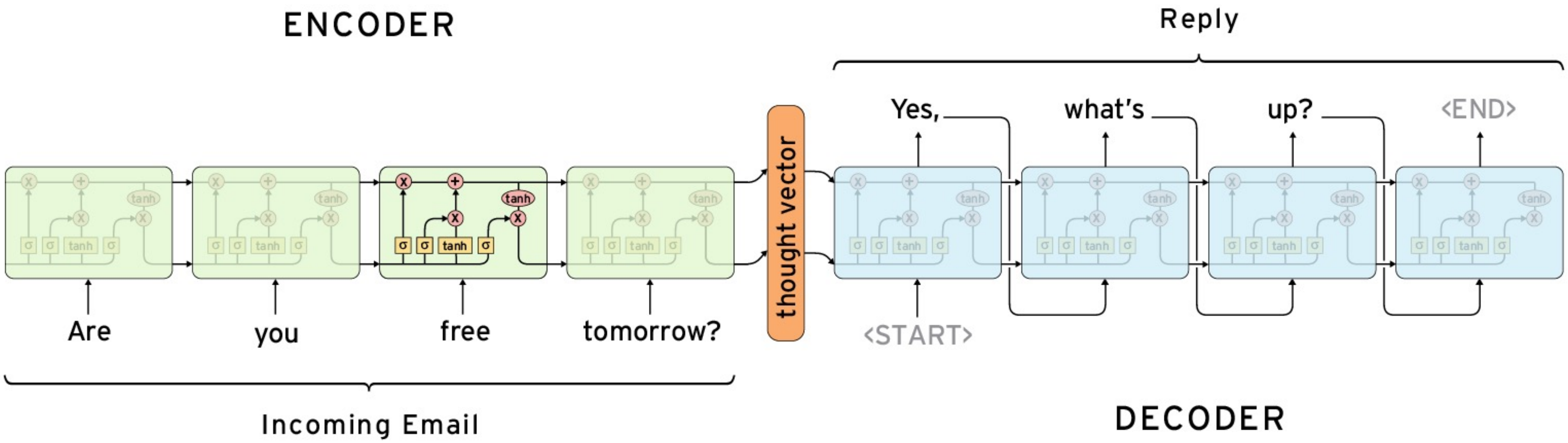


Image Credit: Christopher Olah

Multi-layer (stacked) RNNs

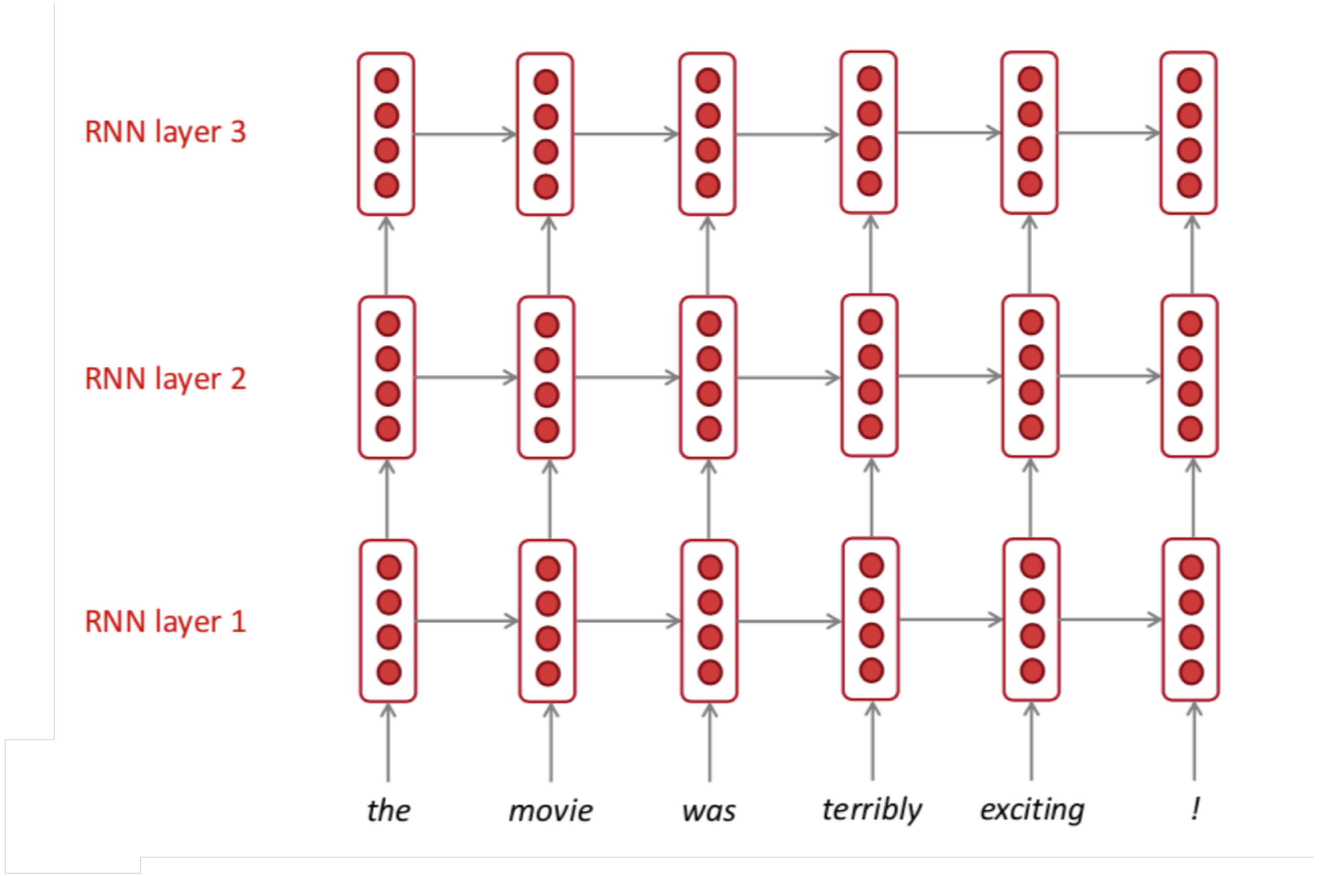


Image Credit: Abigail See

The hidden states from RNN layer i are the inputs to RNN layer $i + 1$

In practice, using 2 to 4 layers is common (usually better than 1 layer)

Transformer-based networks can be up to 24 layers with lots of skip-connections.

Bidirectional RNNs

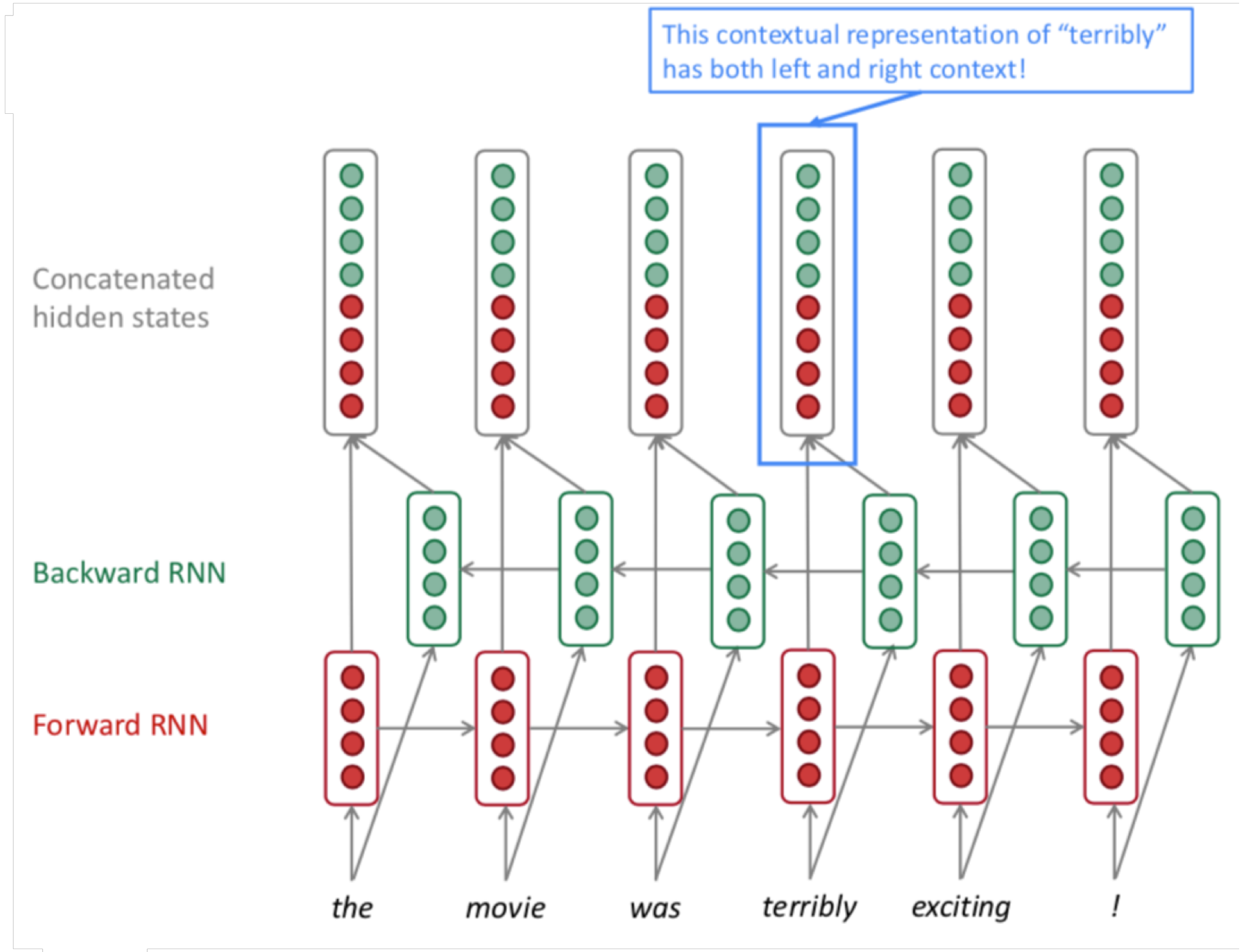


Image Credit: Abigail See

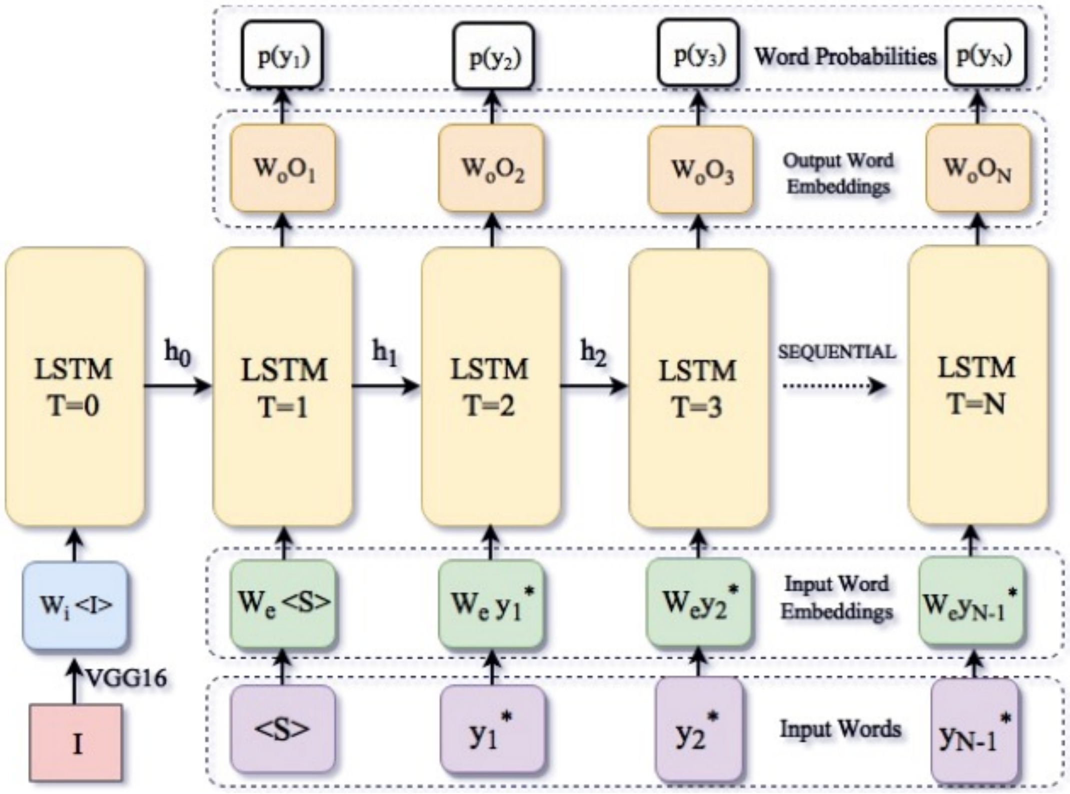
Incorporate information from both directions

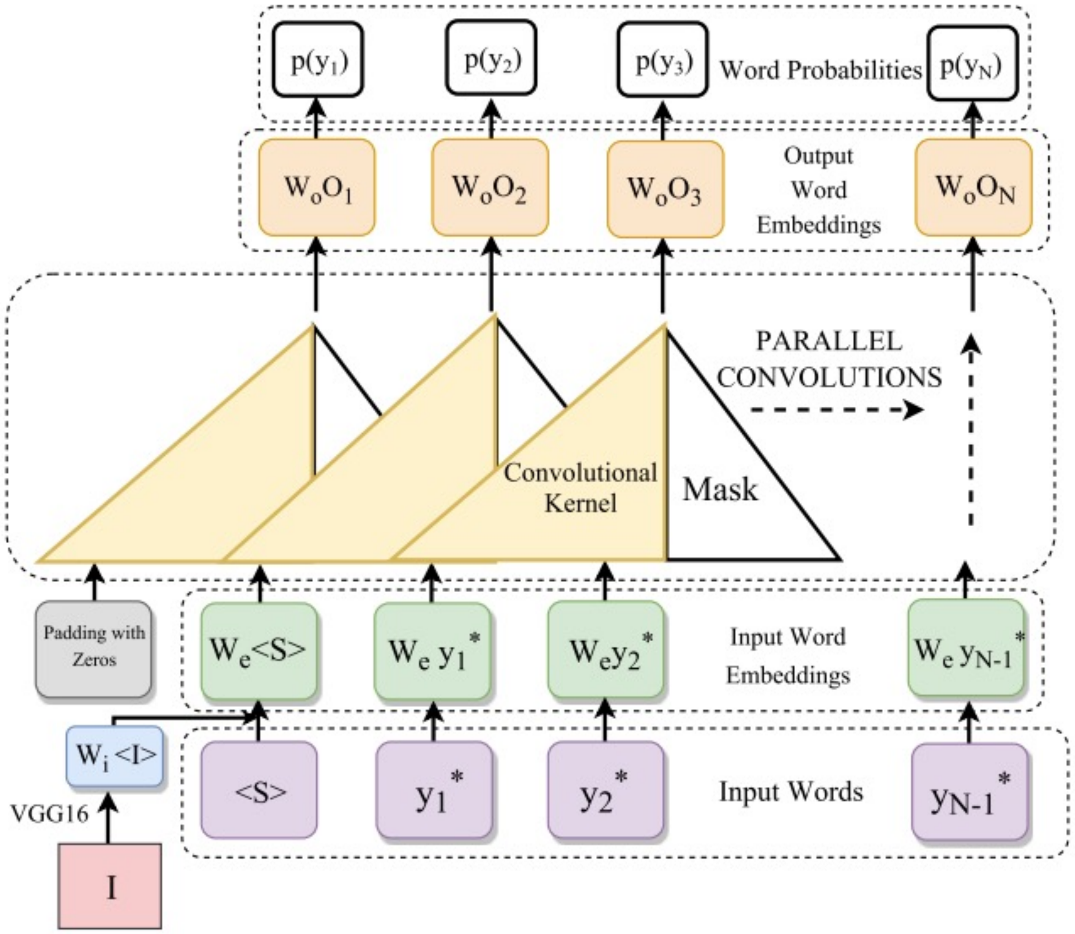
Useful in encoder

Modelling Sequences

CNNs as a fixed-time horizon alternative:

- Parallel computation!
- Tricky encoding.



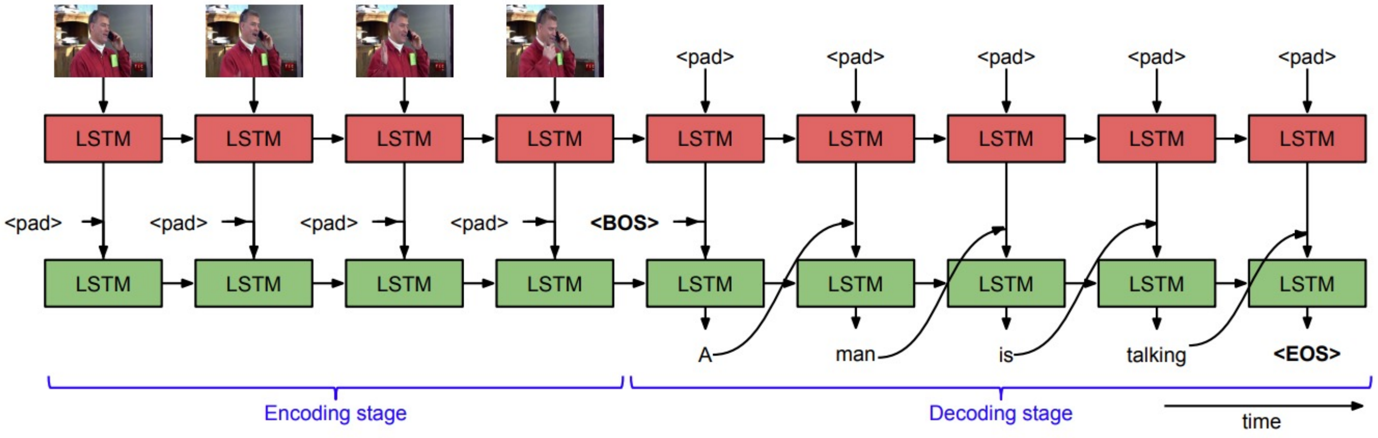


Aneja et al. CVPR 2018

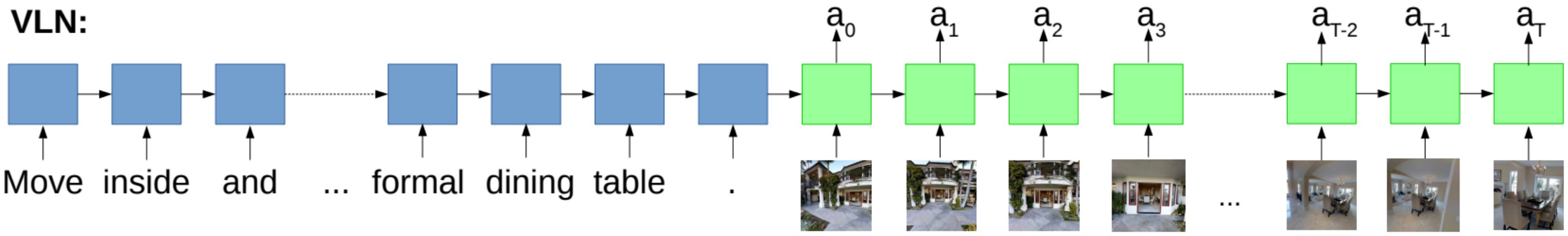
Slide credit: Stefan Lee

Multimodal seq2seq models

- Video captioning (video frames to text)

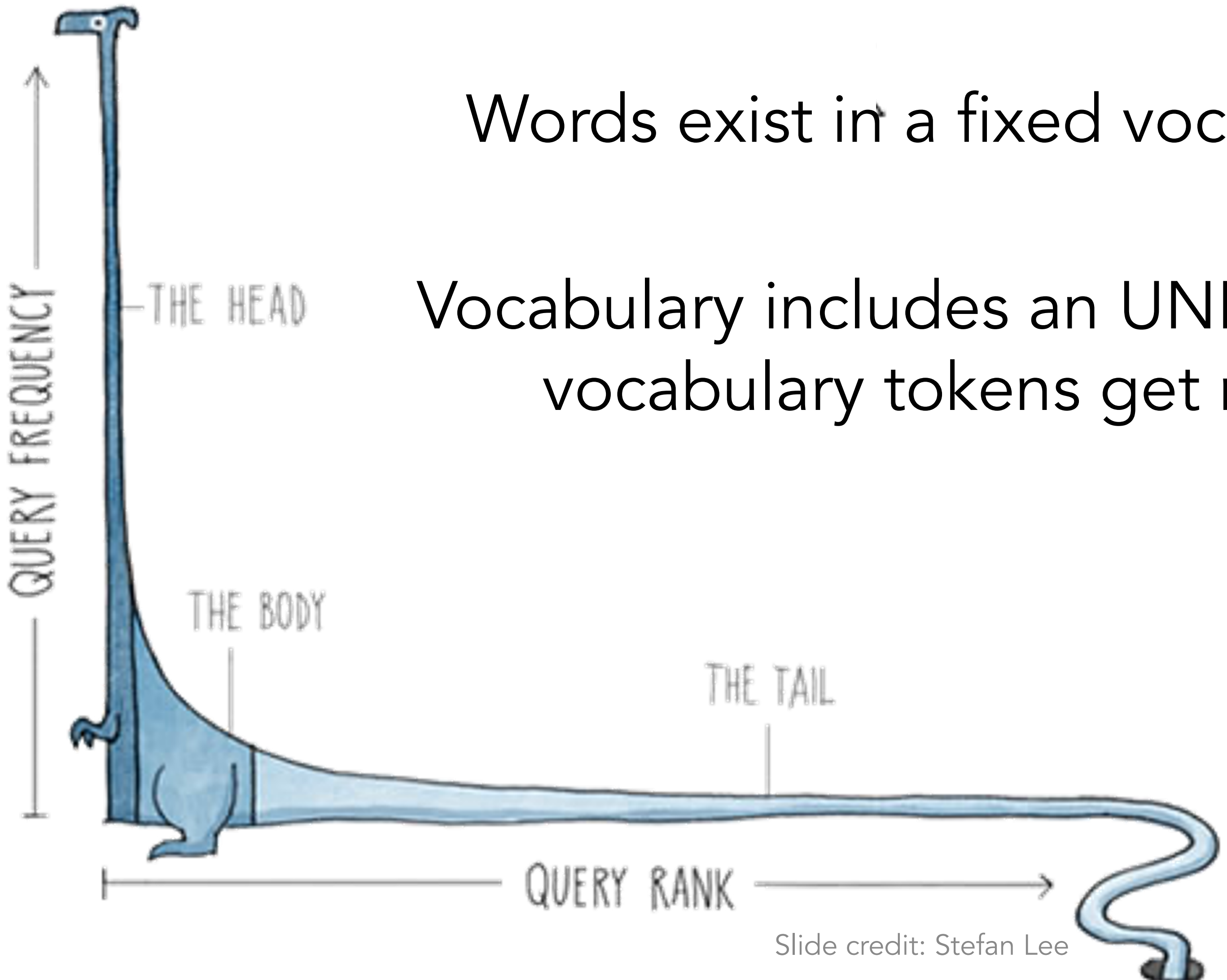


- Embodied AI (text + frames to actions)



Some Notes on Representing Text

Words and Vocabularies



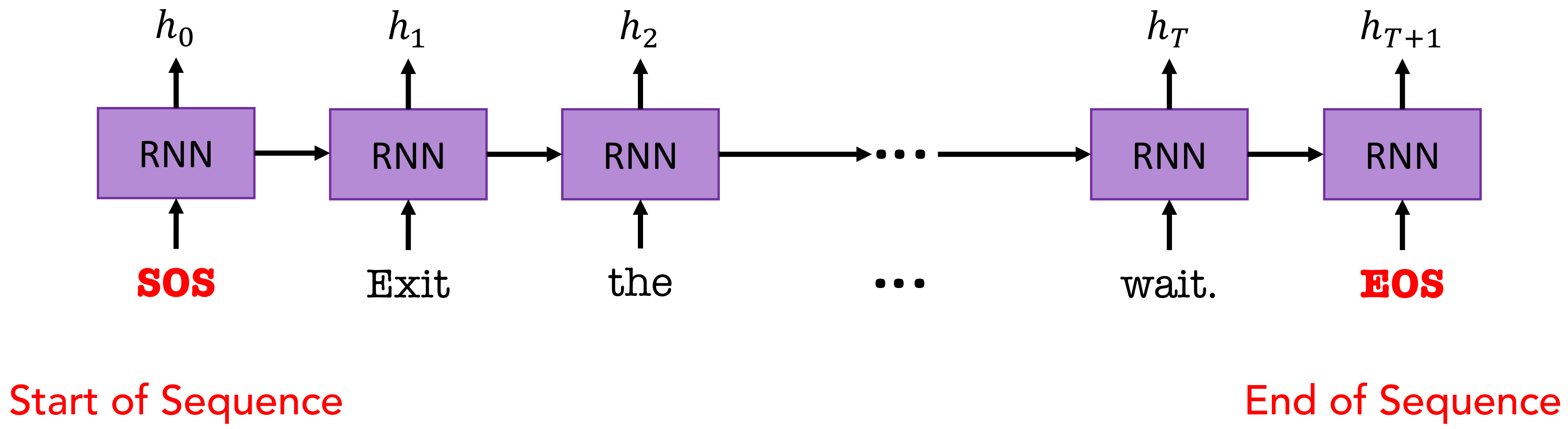
Words exist in a fixed vocabulary, i.e. $w \in V$

Vocabulary includes an UNK token – any out of vocabulary tokens get mapped to this.

Slide credit: Stefan Lee

Some Notes on Representing Text

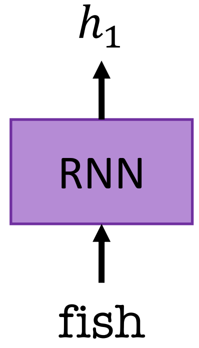
Quirks of Common Practice



Some Notes on Representing Text

What is actually input to represent tokens?

- One-hot vector → learned representation
  - For $V = \{cat, dog, fish\}$, $w_{fish} = [0\ 0\ 1]$.



$$
\begin{array}{cc}
\text{fish} & \mathbf{W} \\
[0\ 0\ 1] & \begin{bmatrix} 0.2 & 1 & 1.5 & 0.8 & -0.2 & 1.2 \\ -1.3 & 2 & -2 & 1.2 & 0.56 & 0.1 \\ 0.13 & 0.2 & 0.95 & 0.2 & -1.3 & 0.5 \end{bmatrix}
\end{array}
$$

$$w_{fish} * W = [0.13\ \ 0.2\ \ 0.95\ \ 0.2\ \ -1.3\ \ 0.5]$$

Initialize to random vectors and learn the embeddings during training

Slide credit: Stefan Lee

Some Notes on Representing Text

What is actually input to represent tokens?

dog

cat

fish

- Use pretrained word embeddings
  - Word2Vec
  - GloVE

$$w_{fish} = GloVE(\text{“}fish\text{”})$$

- Can do a mix of these
  - initialize learned embeddings with pretrained values

Next time

- Multimodal representations