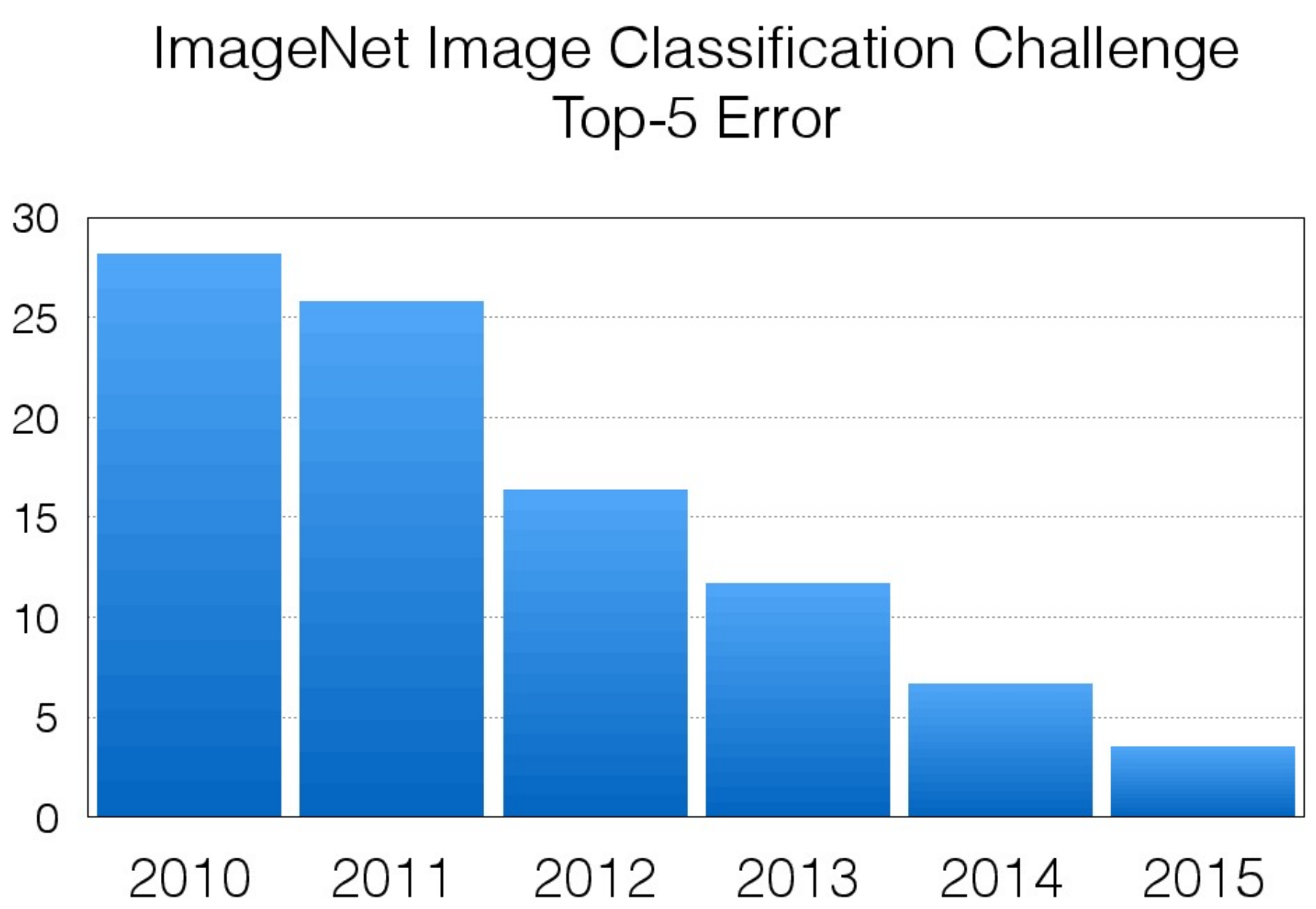
Image Captioning with CNNs and RNNs

CIS, LMU München
SS 2023

Prof. Dr. Alexander Fraser, CIS

Images

- Approaches for image classification and related problems have been dramatically changed by deep learning
- Current explosion of new work and dramatically different problems being addressed
- First let's look at accuracies on the ImageNet task (next slide)
- Then let's take a brief look at image captioning, as a prototypical text/image task

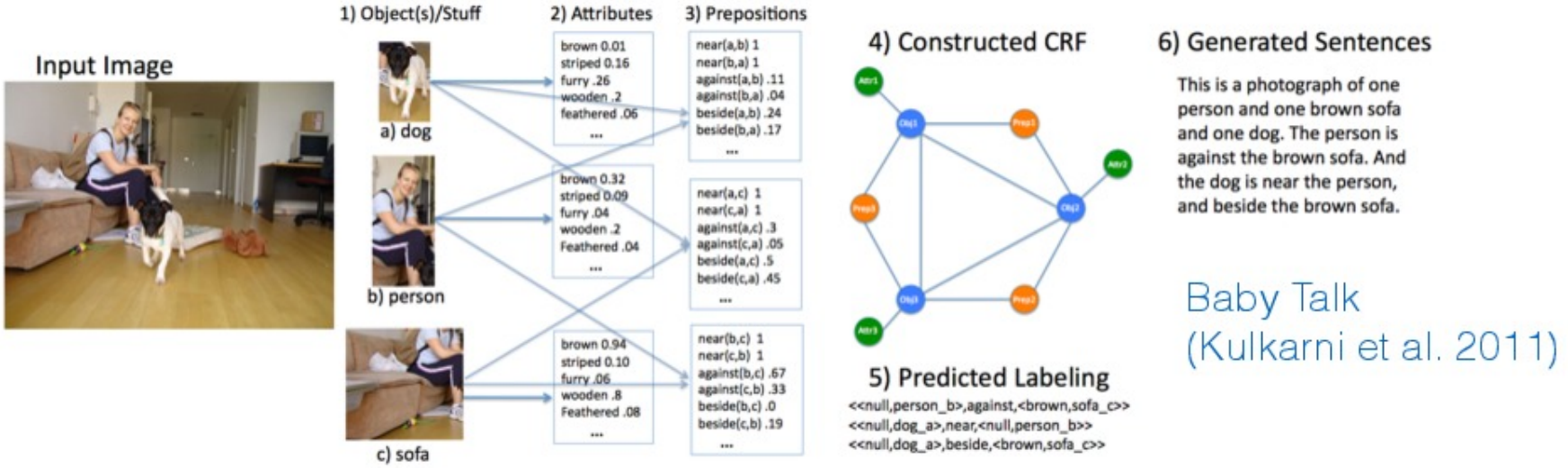


Slide from Andrej Karpathy, results from ILSVRC, Russakovsky et al., 2015

From image classification to image captioning

- Image classification has gotten much better
- The basic approach is the same as training a linear model like perceptron
  - Check if we get the right answer
  - If yes, do nothing
  - If no, update the parameters to make the right answer more likely
- But how can we generate captions?

Generating Descriptions: Prior work



example template:

(noun) _____ in (noun) _____ is (verb) _____ in (noun) _____.

[Barnard '03] [Duygulu '02] [Frome '13] [Gupta & Mannem '12] [Elliott & Keller '13] [Yatskar '14] [Kiros '14] [Yao '10] [Yang '11] [Barbu '12] [Mitchell '12]

Slide from Andrej Karpathy

Core Challenge

how can we predict sentences?

???

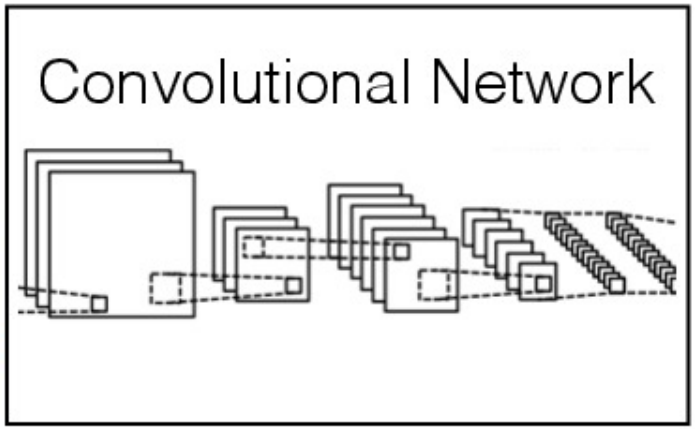
"A dog jumping over a hurdle"

differentiable function

Slide from Andrej Karpathy

Core Challenge

how can we predict sentences?



???

"A dog jumping over a hurdle"

differentiable function

Slide from Andrej Karpathy

Core Challenge

how can we predict sentences?

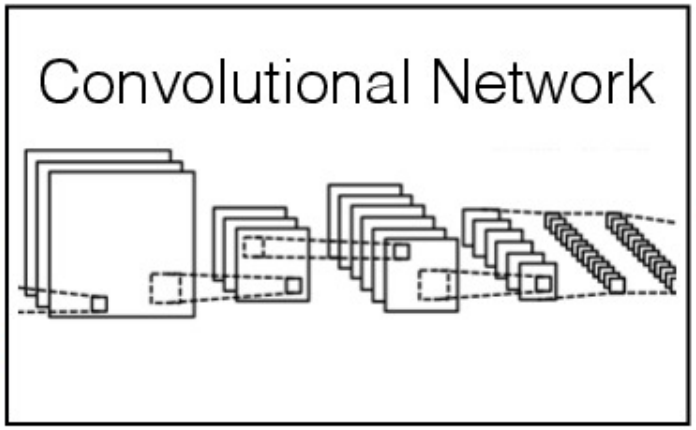


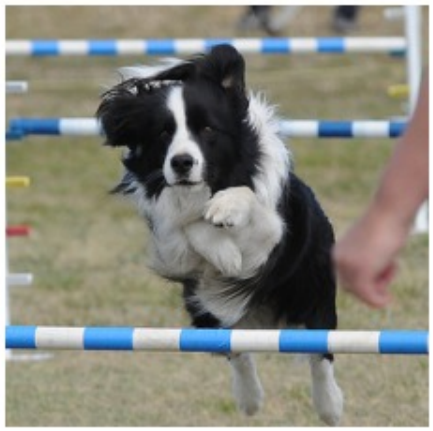
image classification

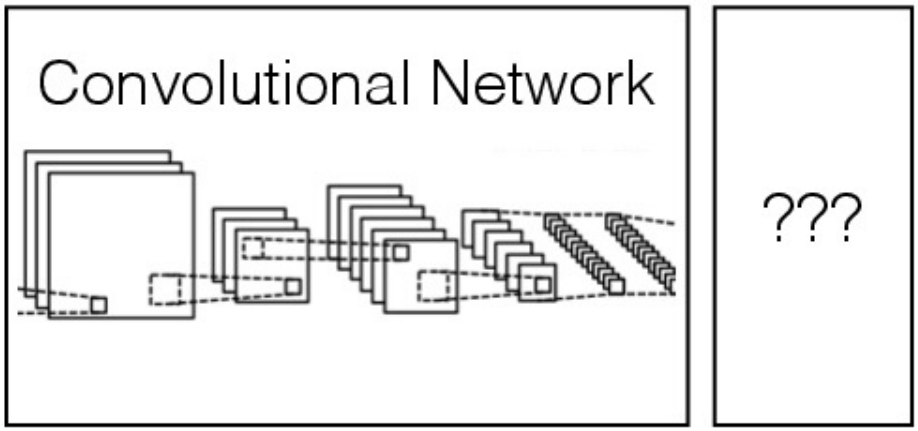
dog

differentiable function

Slide from Andrej Karpathy

Core Challenge





a

dog

jumping

over

a

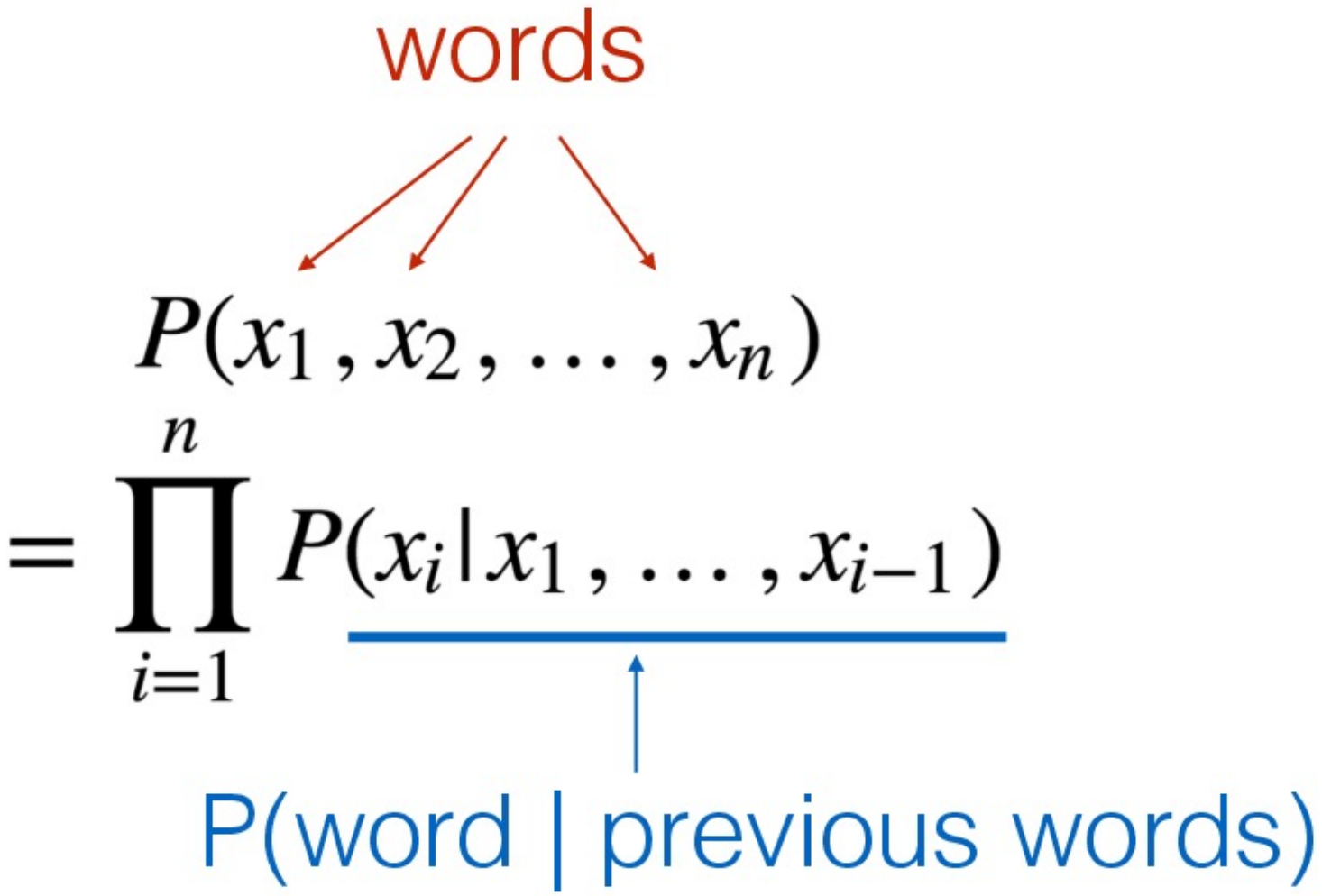
hurdle

<end>

sentences have variable number of words
=> output not fixed size!

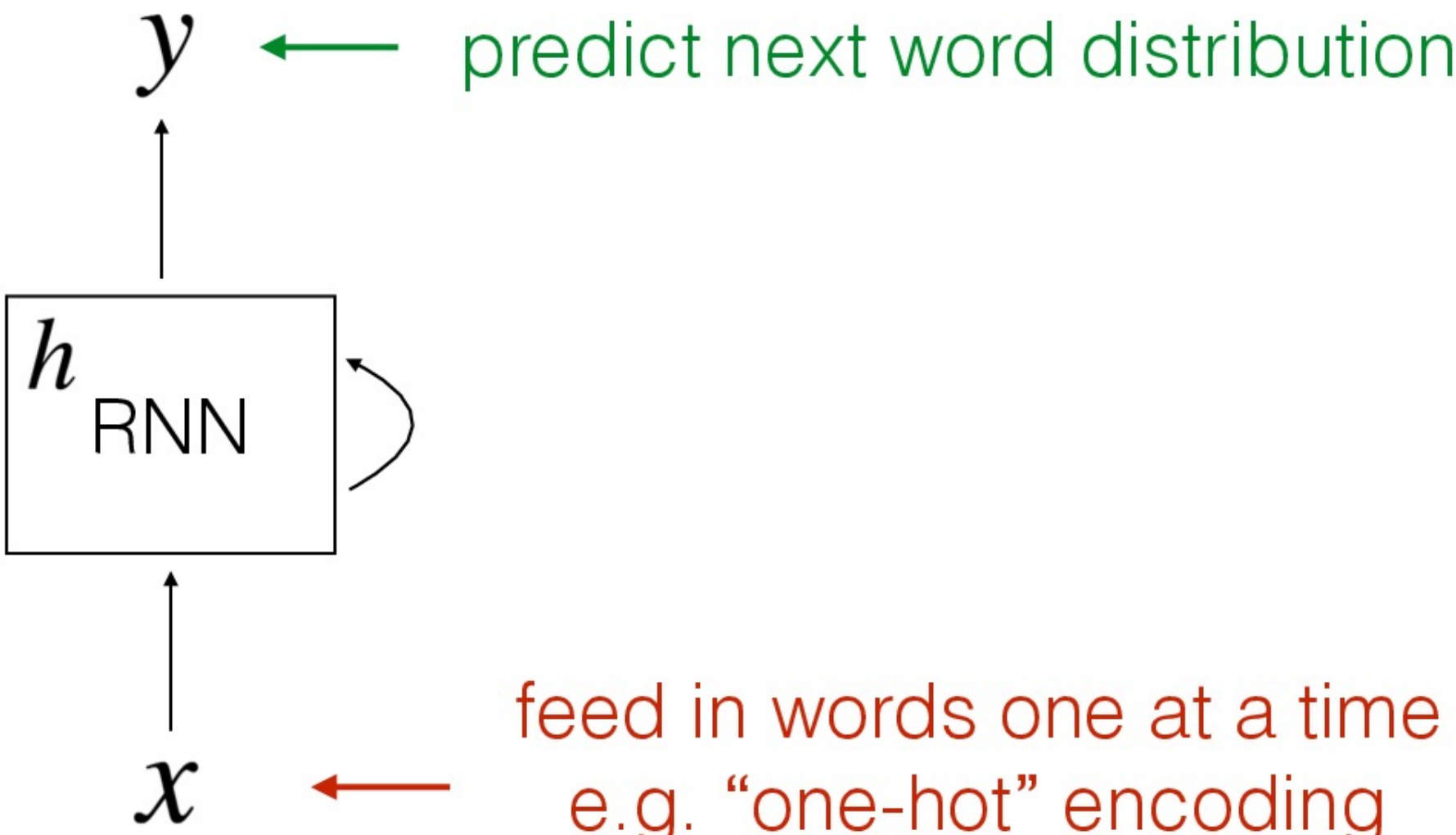
Slide from Andrej Karpathy

Language Model



Slide from Andrej Karpathy

Recurrent Neural Network Language Model



Slide from Andrej Karpathy

Image Classification

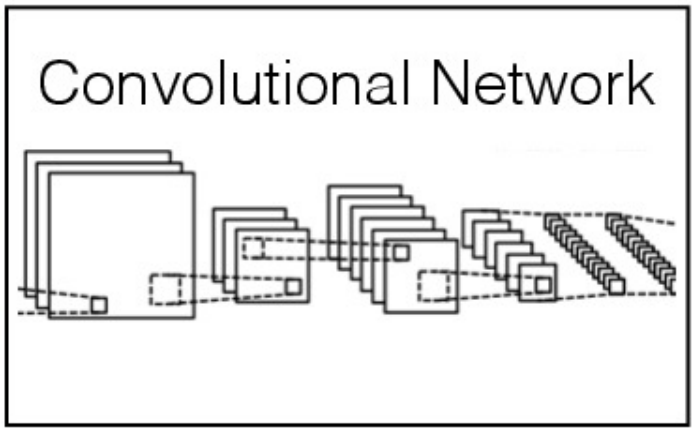
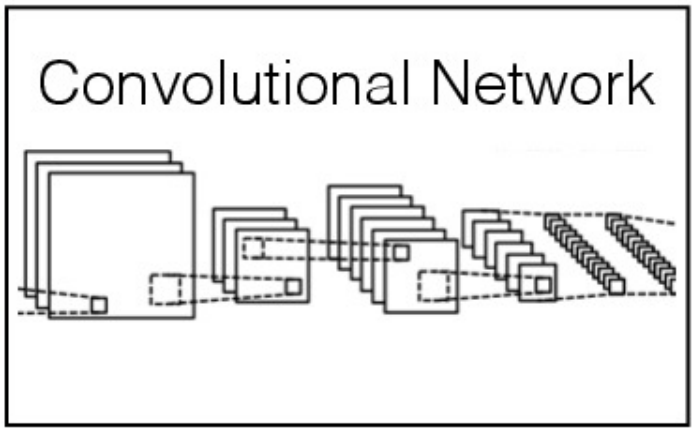


image classification

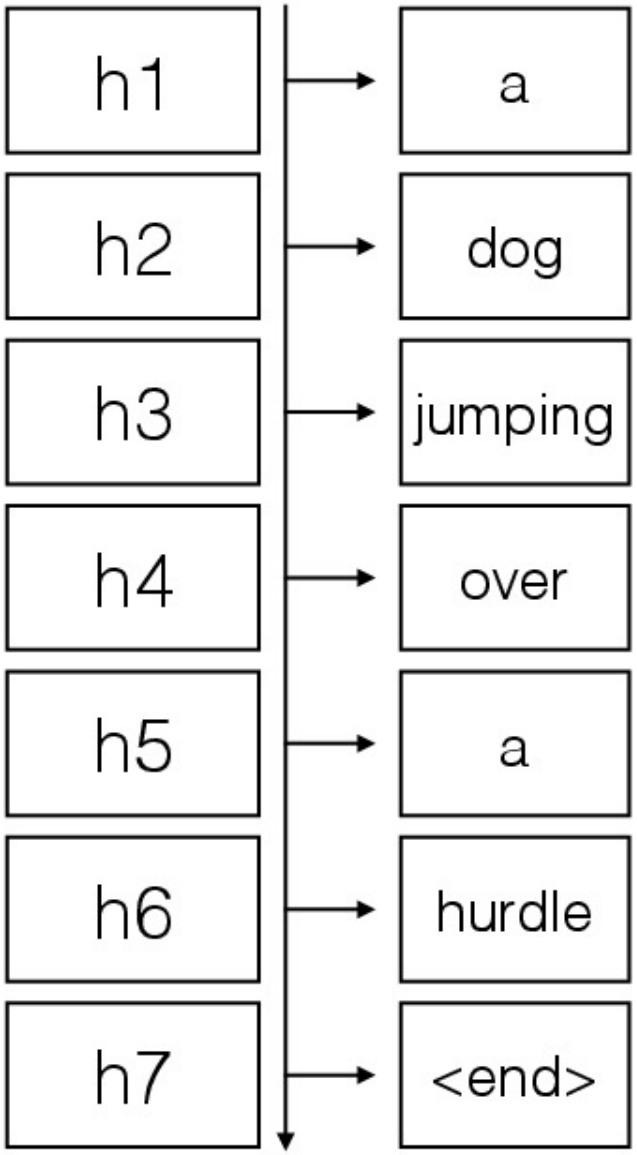
dog

Slide from Andrej Karpathy

Image Captioning

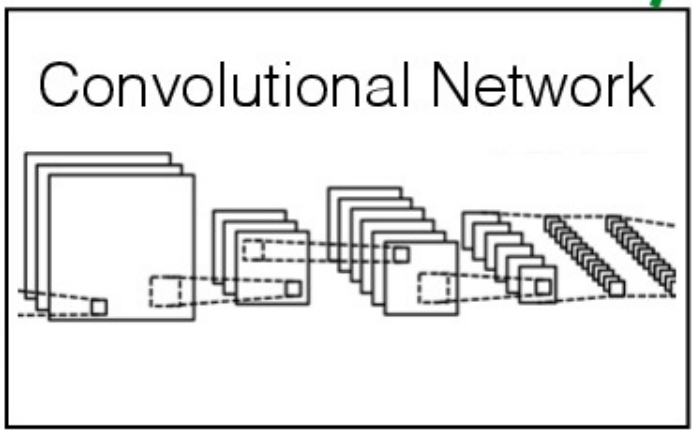


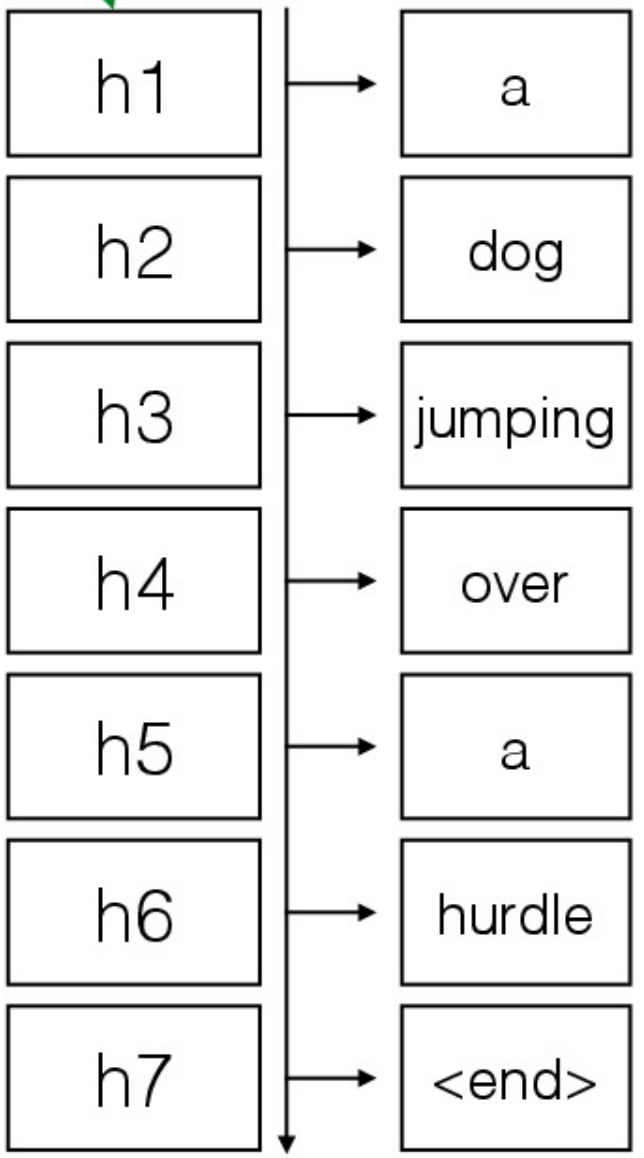
Q: how do we condition the generative process on the image information?



Slide from Andrej Karpathy

Image Captioning





$$h_0 = v$$

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$$

$$y_t = W_{hy} h_t$$

Slide from Andrej Karpathy

Image Sentence Datasets

1. A woman and her dog watch the cameraman in their living with wooden floors.
2. A woman sitting on the couch while a black faced dog runs across the floor.
3. A woman wearing a backpack sits on a couch while a small dog runs on the hardwood floor next to her.
4. A women sitting on a sofa while a small Jack Russell walks towards the camera.
5. White and black small dog walks toward the camera while woman sits on couch, desk and computer seen in the background as well as a pillow, teddy bear and moggie toy on the wood floor.

1. A man in a cowboy hat check approaches a small red sports car.
2. The back and left side of a red Ferrari and two men admiring it.
3. The sporty car is admired by passer by.
4. Two men next to a red sports car in a parking lot.
5. Two men stand beside a red sports car.

[1] **Pascal 1K**: 1,000 images (5 sentences per image)
[2] **Flickr8K**: 8,000 images
[3] **Flickr30K**: 30,000 images
[4] **MSCOCO**: 115,000 images

[1] Rashtchian et al., 2010
[2] Hodosh et al., 2013
[3] Young et al., 2014
[4] Lin et al., 2015

Slide from Andrej Karpathy

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

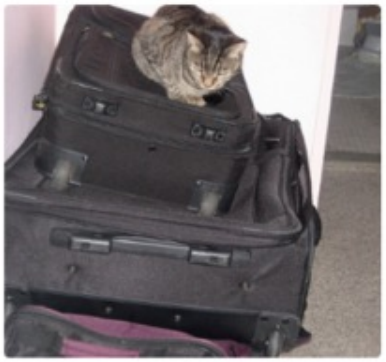
"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

"little girl is eating piece of cake."

"baseball player is throwing ball in game."

"woman is holding bunch of bananas."



"black cat is sitting on top of suitcase."

Slide from Andrej Karpathy

Example Error

"a woman in a bikini is jumping over a hurdle."

Slide modified from Andrej Karpathy

Limitations

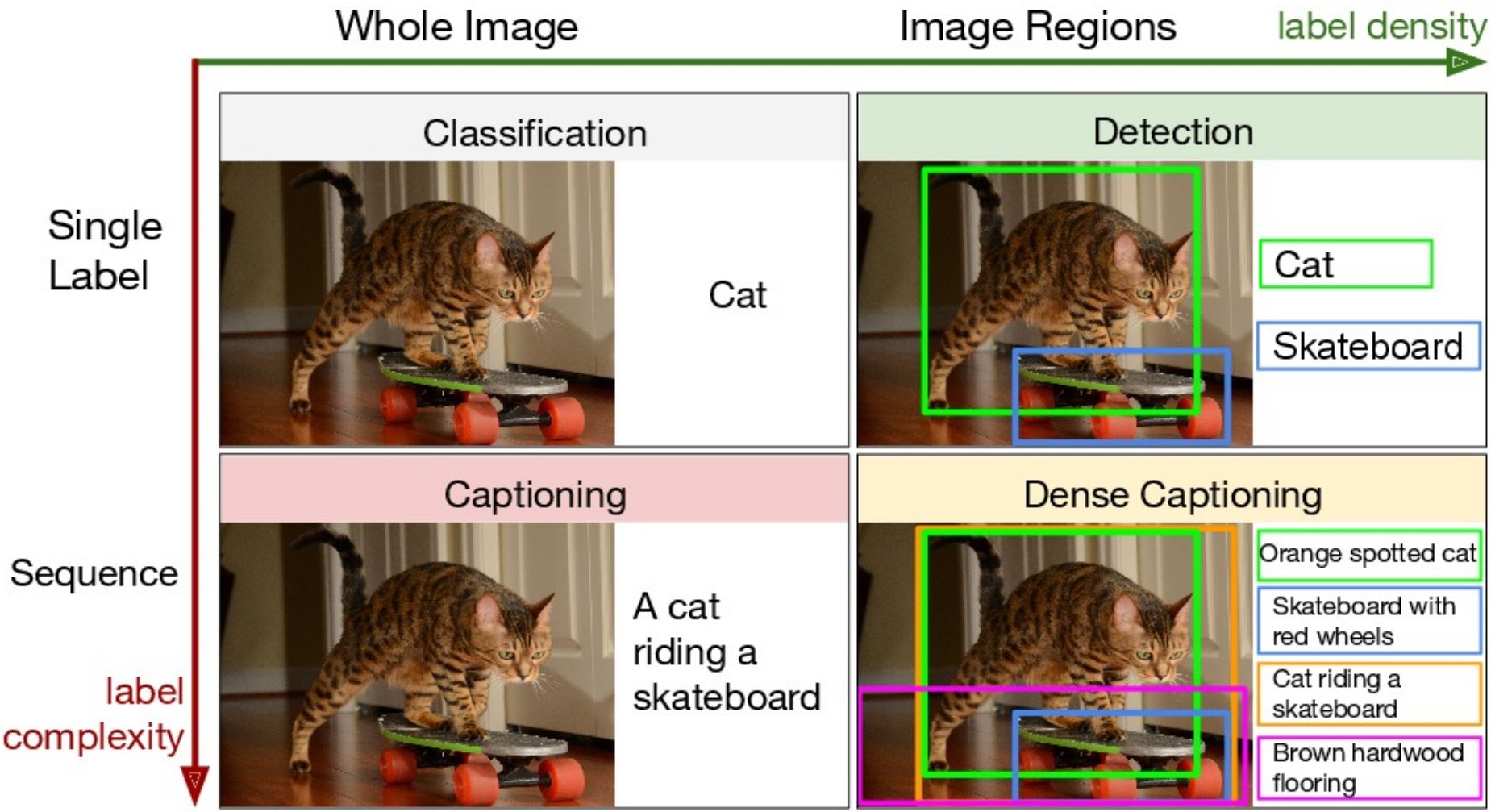
"A group of people in an office."

Slide from Andrej Karpathy

Can go even further...

- Deep learning enabled addressing image caption generation in a much more natural way
  - Also, cross-fertilization of ideas with machine translation (!)
  - Framework is actually very similar to neural machine translation
- Deep learning also enables solving new problems
  - For instance, there is now work on breaking images down into regions (next slide)

Dense Captioning



Slide from Andrej Karpathy

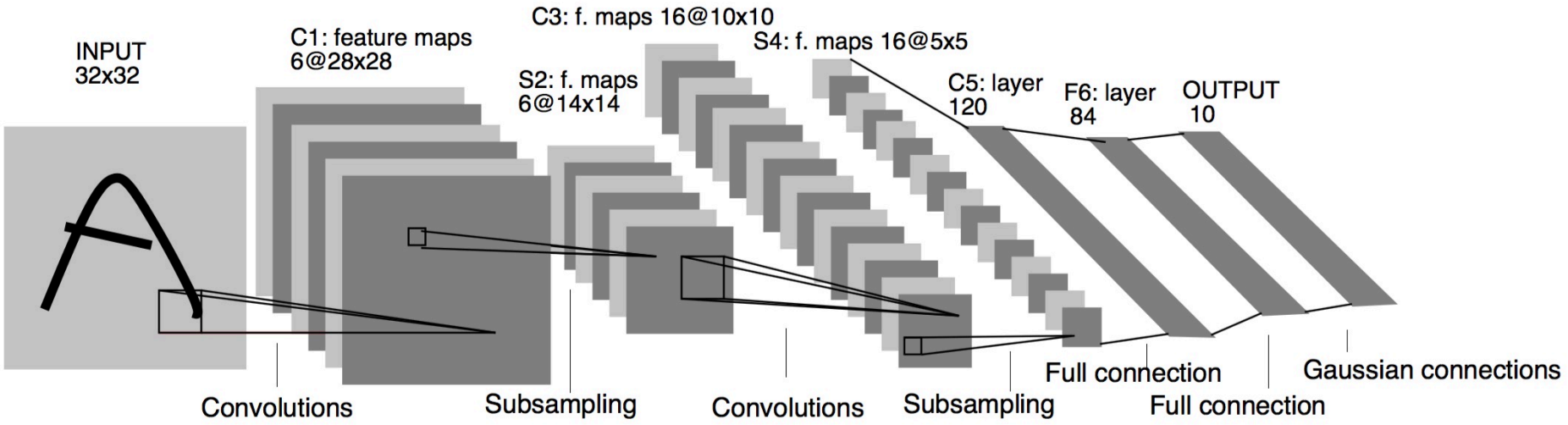
Slides

- The slides on image captioning are from Andrej Karpathy (former PhD student of Fei-Fei Li), now at OpenAI

- Questions?

- Thank you for your attention!

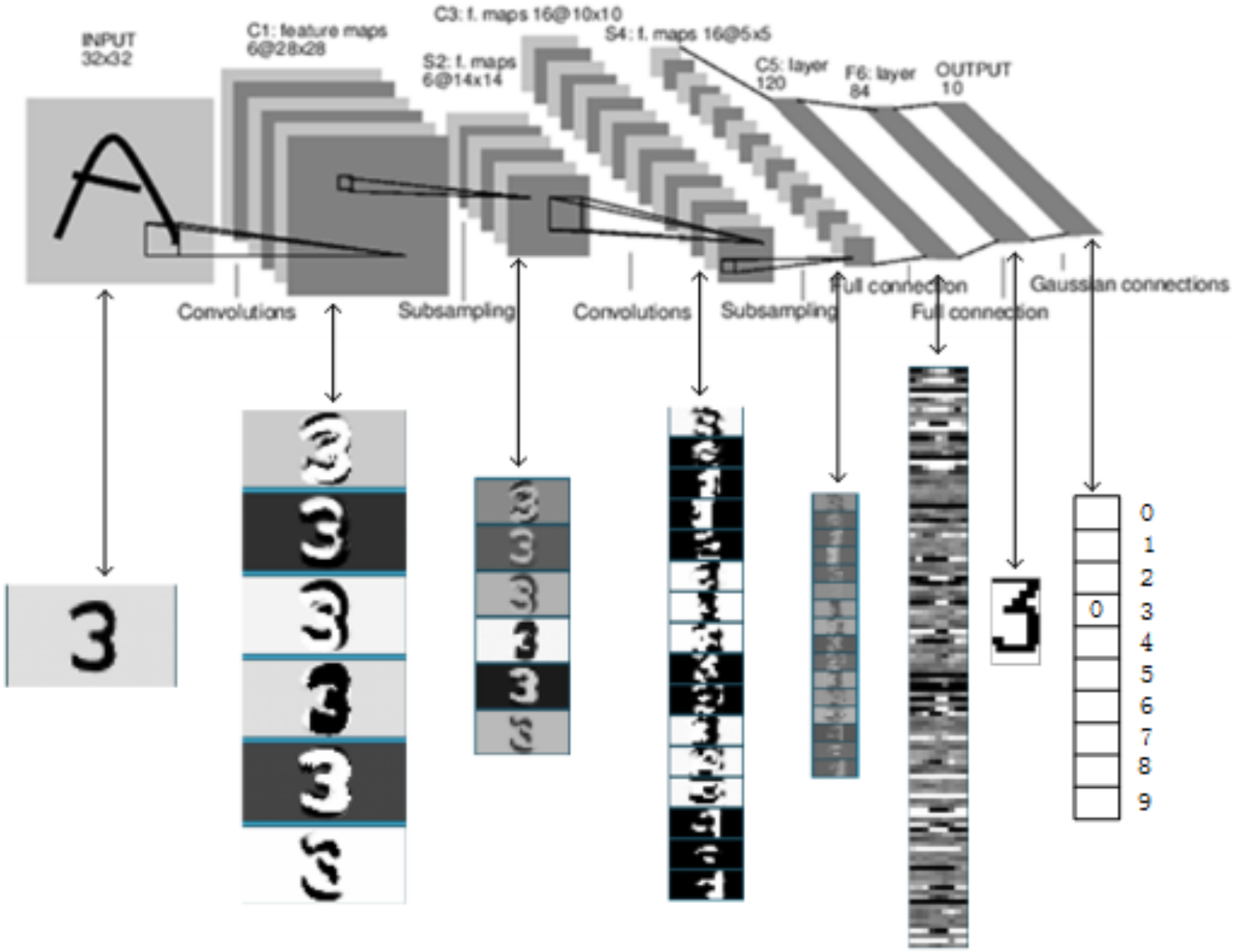
LeNet-5



- convolutional neural network use sequence of 3 layers: convolution, pooling, non-linearity –> This may be the key feature of Deep Learning for images since this paper!
- use convolution to extract spatial features
- subsample using spatial average of maps
- non-linearity in the form of tanh or sigmoids
- multi-layer neural network (MLP) as final classifier
- sparse connection matrix between layers to avoid large computational cost

(Graphic from Yann LeCun, Text from Culurciello et al.)

LeNet-5 recognizing "3"



(Graphic from Yann LeCun (and world4jason??))