Visual attention [with and for] deep neural nets

June 16, 2016

Adobe CTL Vision and Learning Reading Group

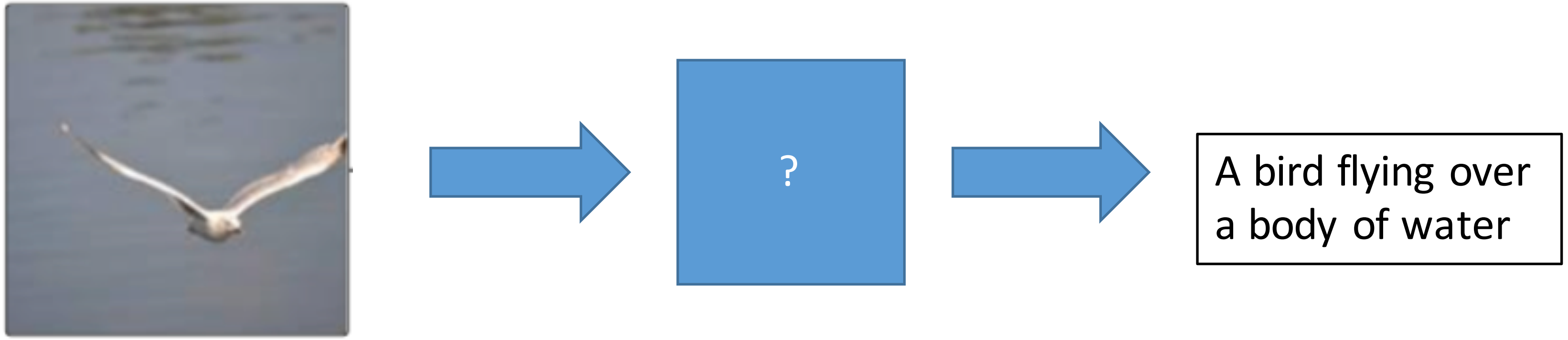
Zoya Bylinskii

Visual attention
for deep neural nets **[captioning]**

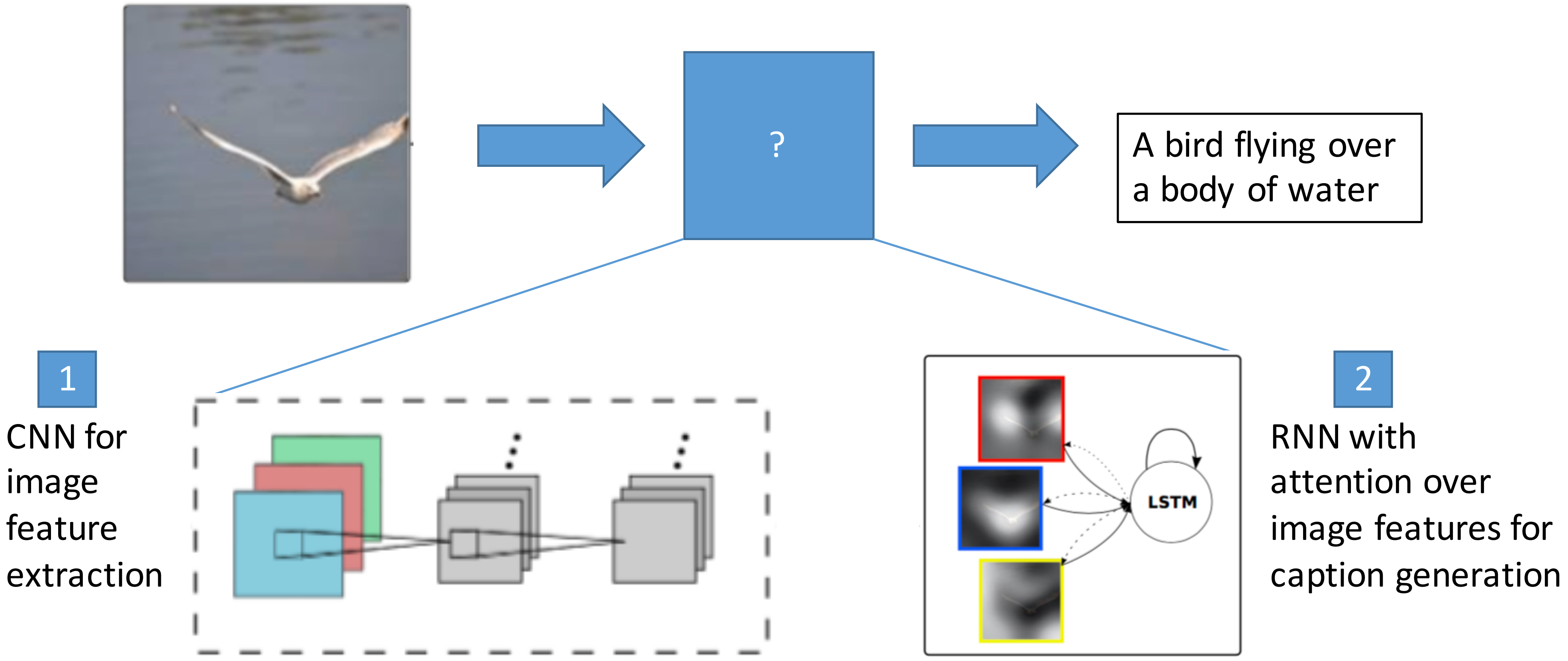
Paper discussion:

“Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”
K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio
[ICML 2015]

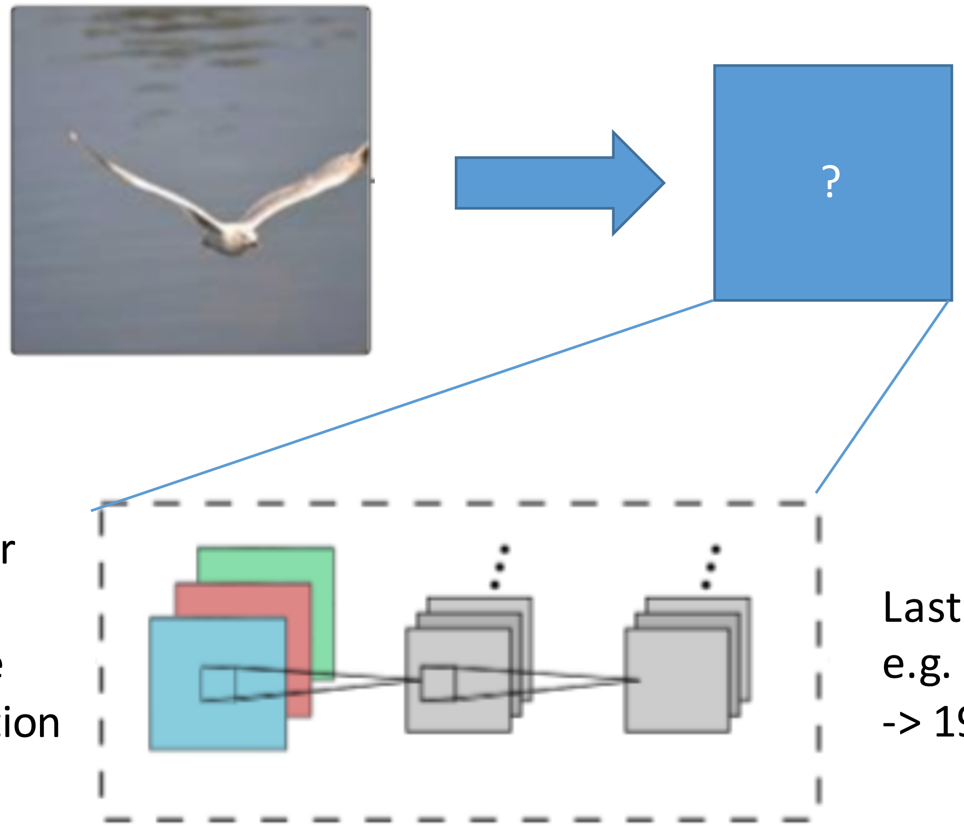
Visual Attention for Image Captioning



Visual Attention for Image Captioning



Visual Attention for Image Captioning

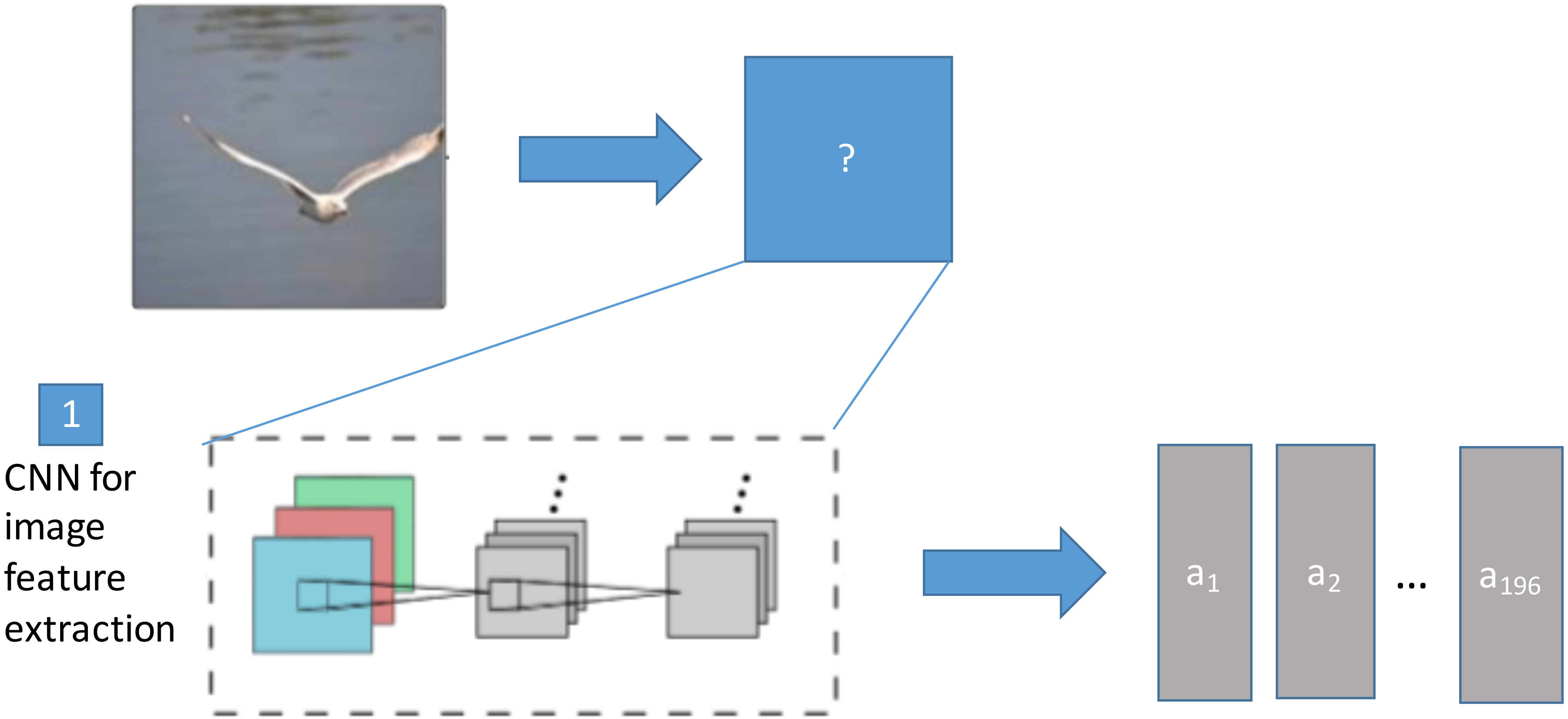


1

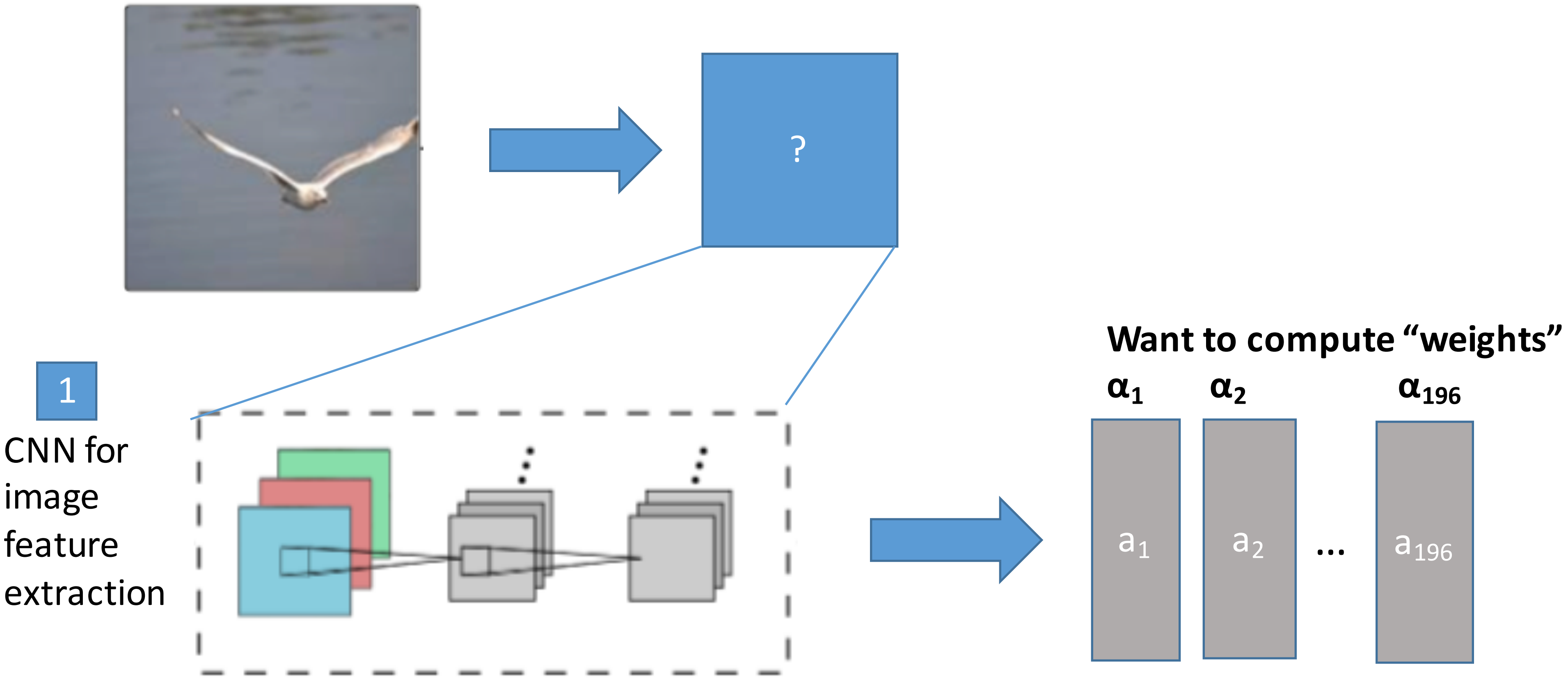
CNN for image feature extraction

Last convolutional layer before max pooling
e.g. VGGnet: 14x14x512 feature map
-> 196 image feature vectors of dimension 512

Visual Attention for Image Captioning



Visual Attention for Image Captioning



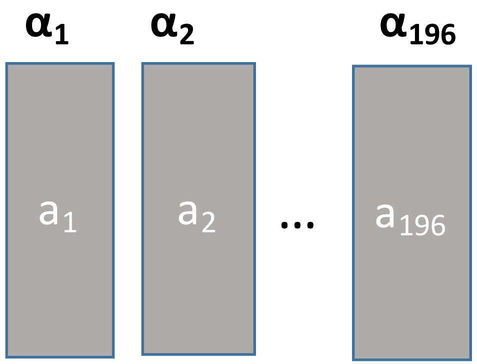
Visual Attention for Image Captioning

- The positive weight: $\alpha_{ti} = \exp(e_{ti}) / \sum_k \exp(e_{tk})$.
- Attention model[2]:

$$e_{ti} = f_{att}(a_i, h_{t-1}) = \begin{cases} a_i^\top W_a h_{t-1} \\ V_a^\top \tanh(W_a[a; h_{t-1}]) \end{cases}$$

Equations on this slide courtesy of: http://people.ee.duke.edu/~lcarin/Yunchen9.25.2015.pdf

Want to compute "weights"



Visual Attention for Image Captioning

- The context vector: $\hat{z}_t = \phi(\{a_i\}, \{\alpha_{ti}\})$.
- The positive weight: $\alpha_{ti} = \exp(e_{ti}) / \sum_k \exp(e_{tk})$.
- Attention model[2]:

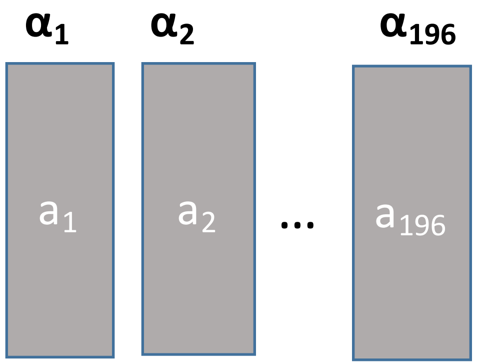
$$e_{ti} = f_{att}(a_i, h_{t-1}) = \begin{cases} a_i^\top W_a h_{t-1} \\ V_a^\top \tanh(W_a[a; h_{t-1}]) \end{cases}$$

- Deterministic "Soft" Attention: $\hat{z}_t = \sum_{i=1}^{L} \alpha_{ti} a_i$
- Stochastic "Hard" Attention: $\boldsymbol{s}_t$ is a one-hot vector:

$$p(s_{ti} = 1 | s_{j<t}, \boldsymbol{a}) = \alpha_{ti} \qquad \hat{z}_t = \sum_{i=1}^{L} s_{ti} a_i$$

Equations on this slide courtesy of: http://people.ee.duke.edu/~lcarin/Yunchen9.25.2015.pdf

Want to compute "weights"



Visual Attention for Image Captioning

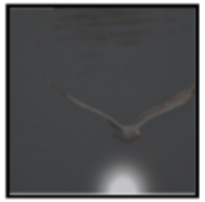
- The context vector: $\hat{z}_t = \phi(\{a_i\}, \{\alpha_{ti}\})$.

- Deterministic "Soft" Attention: $\hat{z}_t = \sum_{i=1}^{L} \alpha_{ti} a_i$
- Stochastic "Hard" Attention: $\boldsymbol{s}_t$ is a one-hot vector:

$$p(s_{ti} = 1 | s_{j<t}, \boldsymbol{a}) = \alpha_{ti} \qquad \hat{z}_t = \sum_{i=1}^{L} s_{ti} a_i$$

Relative importance of location *i* for producing next word

Probability that location *i* is right place to focus for producing next word

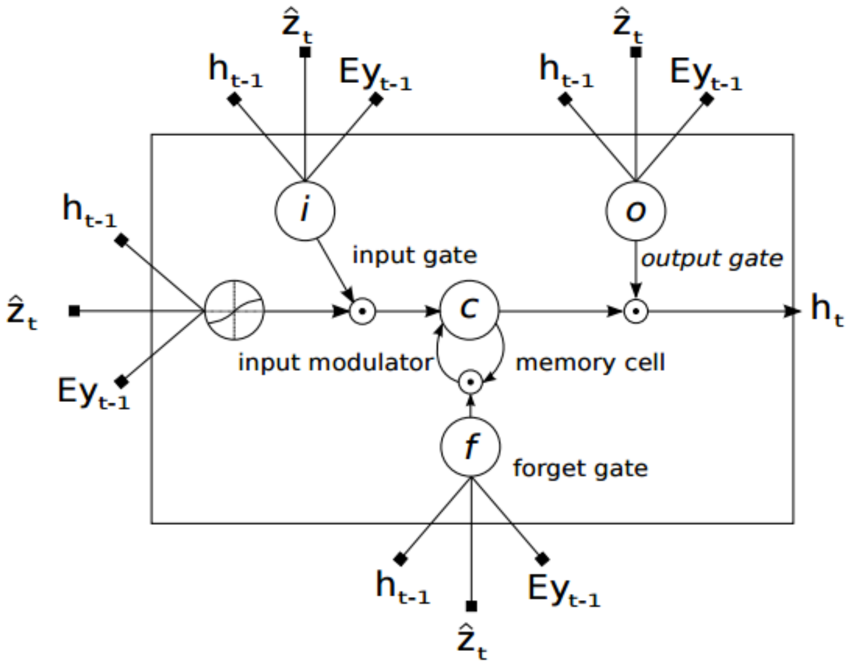


water

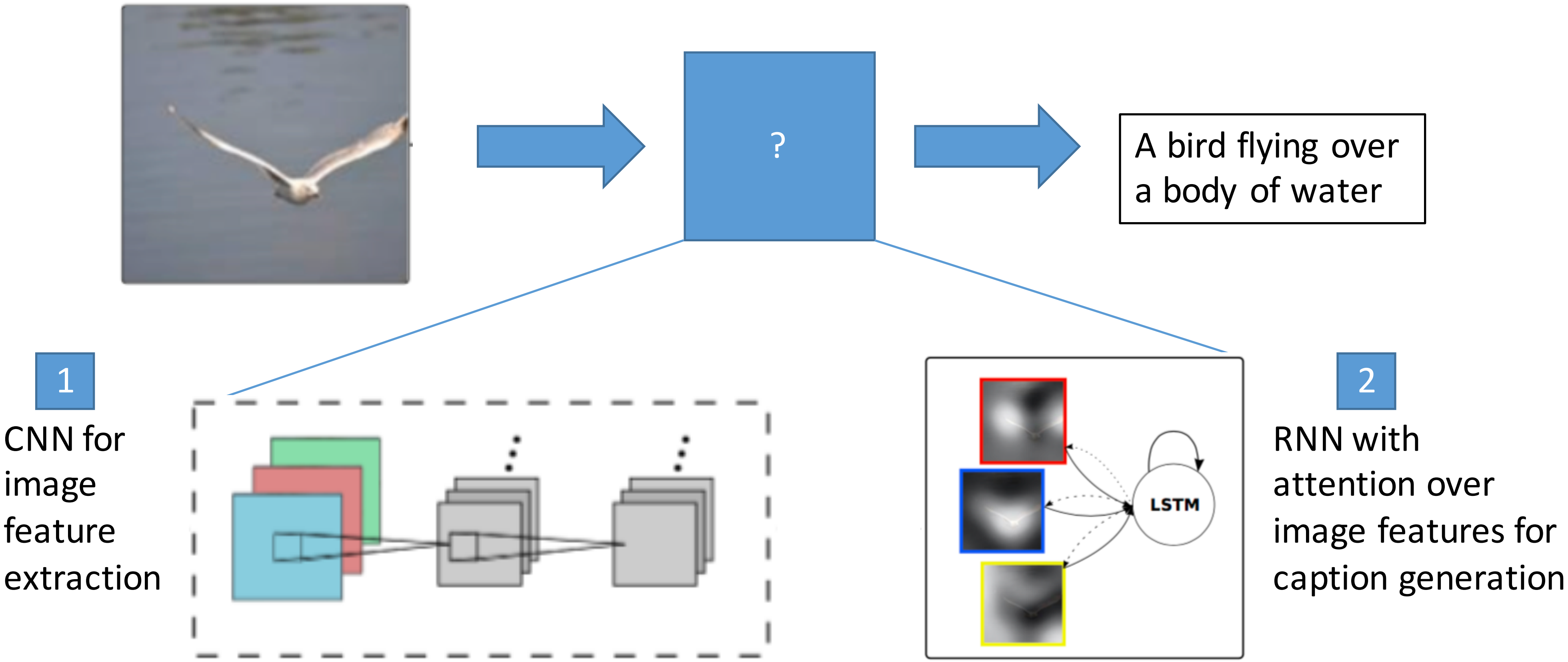
Equations on this slide courtesy of: http://people.ee.duke.edu/~lcarin/Yunchen9.25.2015.pdf

Visual Attention for Image Captioning

- LSTM network generates one word $\mathbf{y_t}$ at every time step conditioned on a context vector, the previous hidden state and the previously generated words
- The context vector $\mathbf{z_t}$ is a dynamic representation of the relevant part of the image input at time t

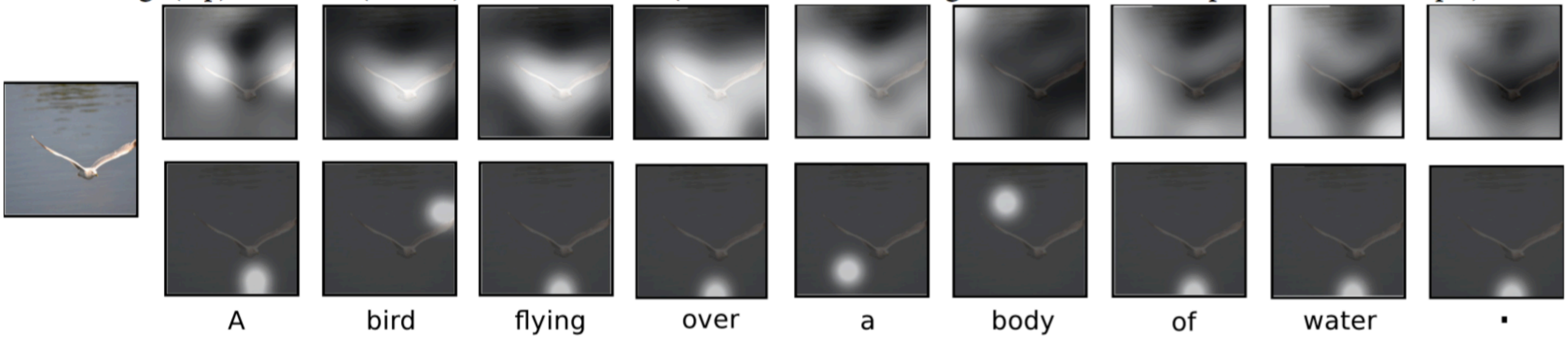


Visual Attention for Image Captioning



Visual Attention for Image Captioning

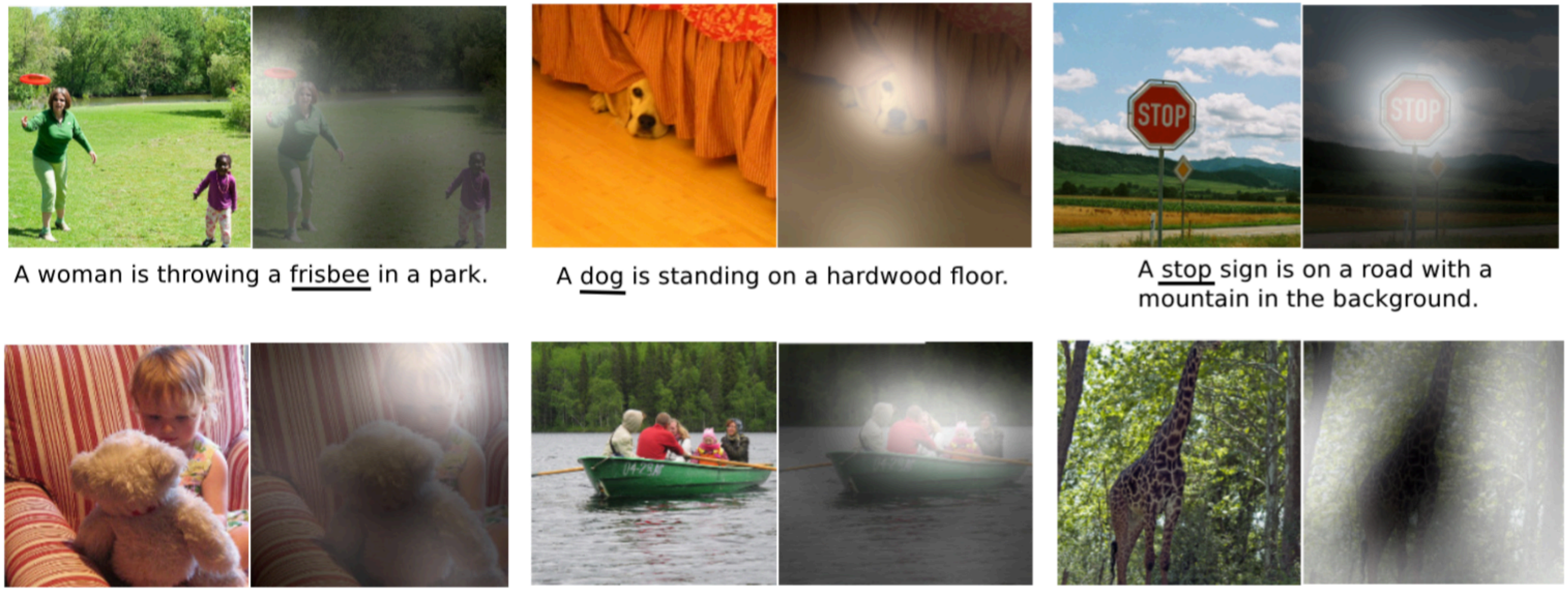
Figure 3. Visualization of the attention for each generated word. The rough visualizations obtained by upsampling the attention weights and smoothing. (top)“soft” and (bottom) “hard” attention (note that both models generated the same captions in this example).



“Show, Attend and Tell: Neural Image Caption Generation with Visual Attention” K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio [ICML 2015]

Visual Attention for Image Captioning

Figure 4. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



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.

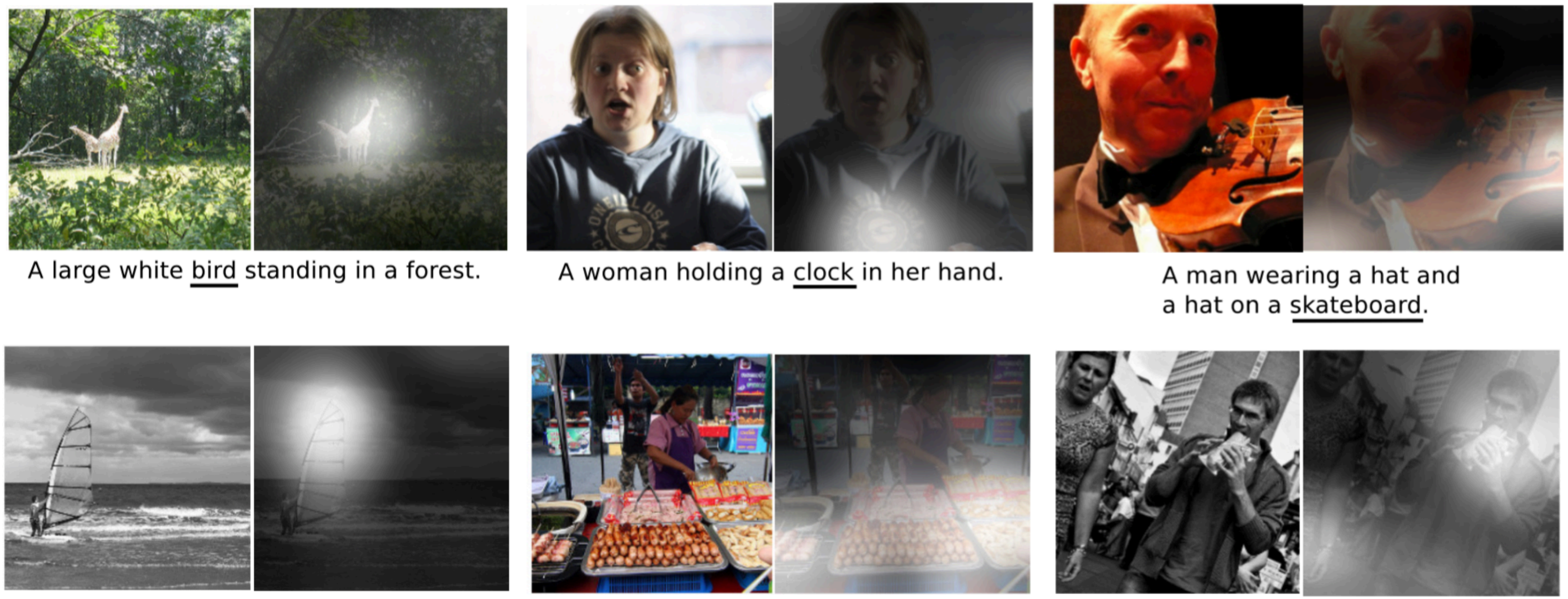
A group of people sitting on a boat in the water.

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

"Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio [ICML 2015]

Visual Attention for Image Captioning

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.

“Show, Attend and Tell: Neural Image Caption Generation with Visual Attention” K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio [ICML 2015]

Visual attention
for deep neural nets
[question answering]

Paper discussion:

“Stacked Attention Networks for Image Question Answering”, Z. Yang, X. He, J. Gao, L. Deng, A. Smola [arXiv, Jan 2016]

Visual Attention for Question Answering

Answer: Dogs

Question: What are sitting in the basket on a bicycle?

Visual Attention for Question Answering

Question: What are sitting in the basket on a bicycle?

?

Answer: Dogs

3

stacked attention model

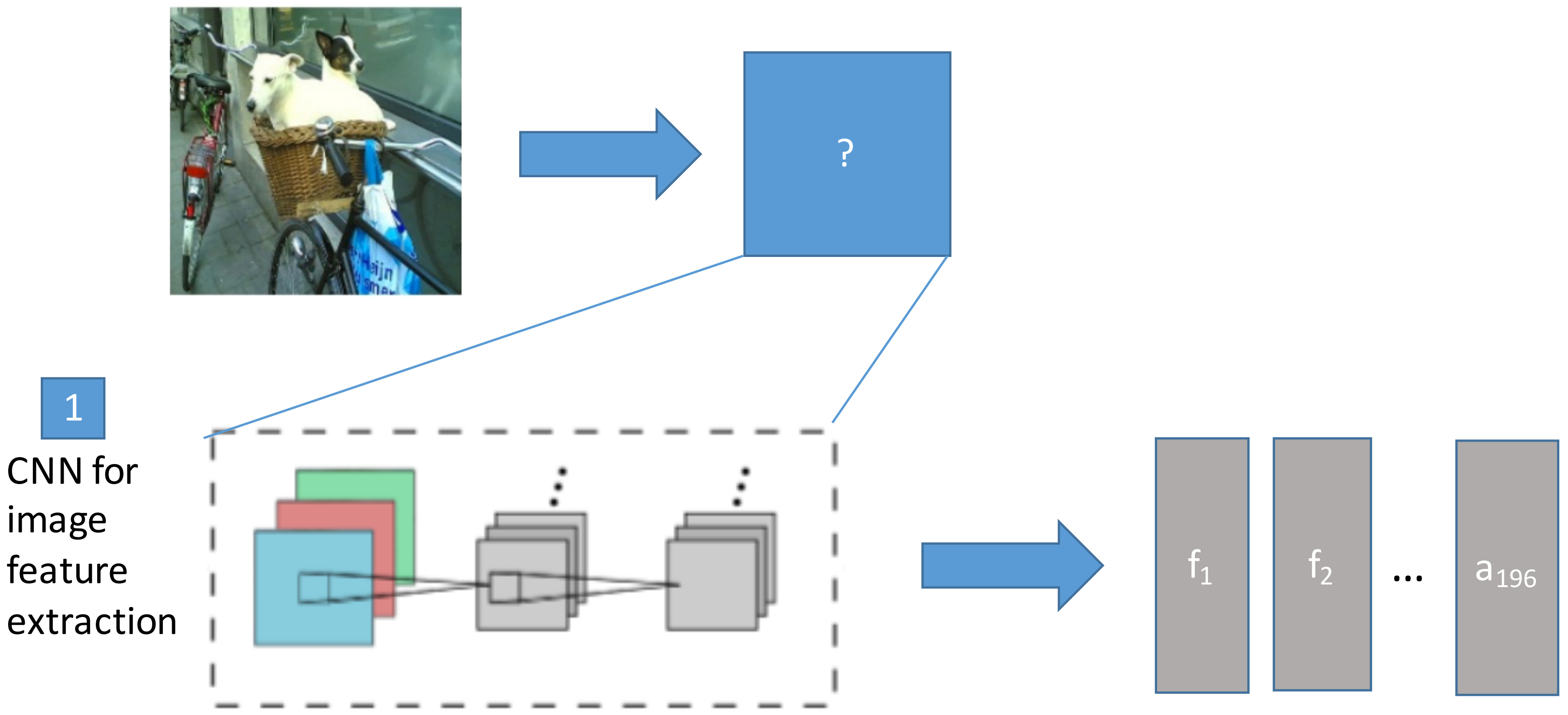
1

CNN for image feature extraction

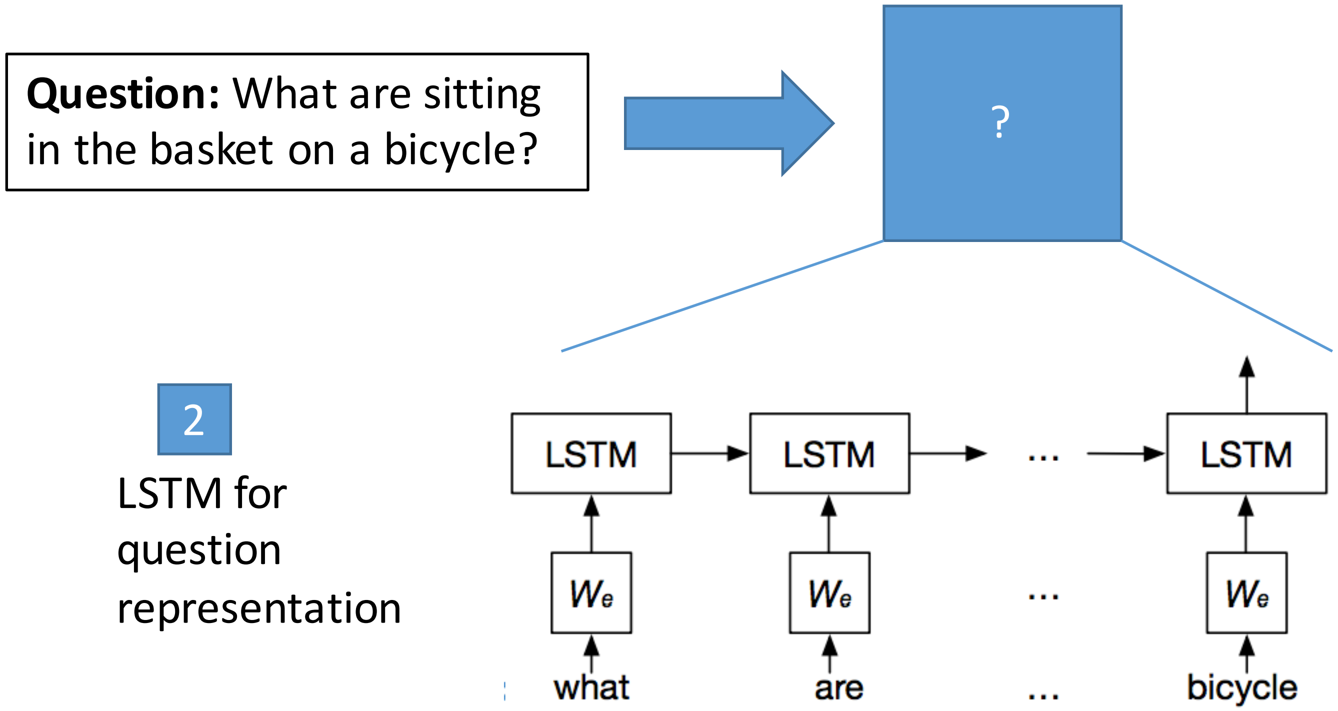
2

CNN or LSTM for question representation

Visual Attention for Question Answering

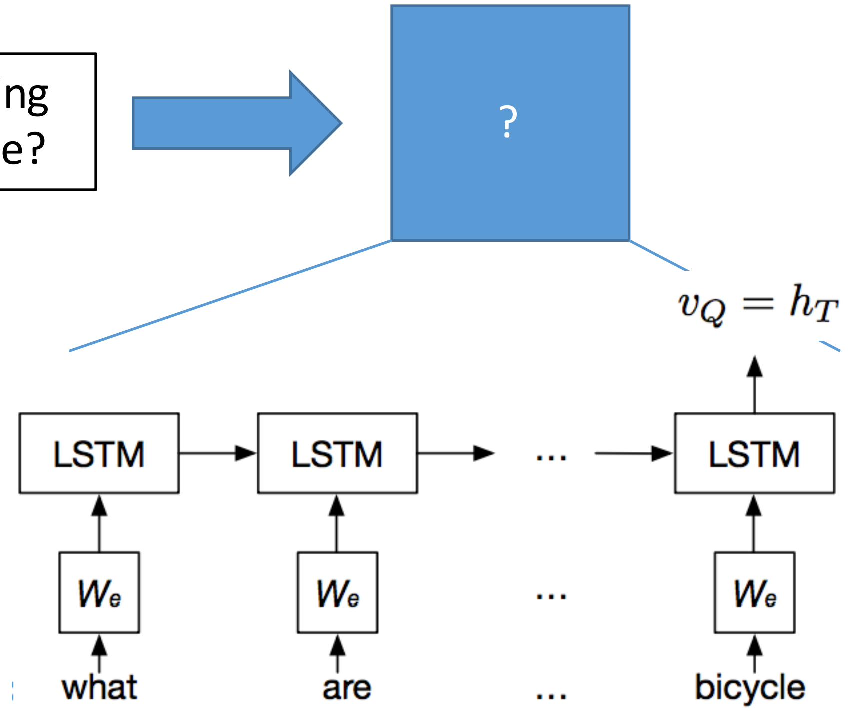


Visual Attention for Question Answering



Visual Attention for Question Answering

Question: What are sitting in the basket on a bicycle?



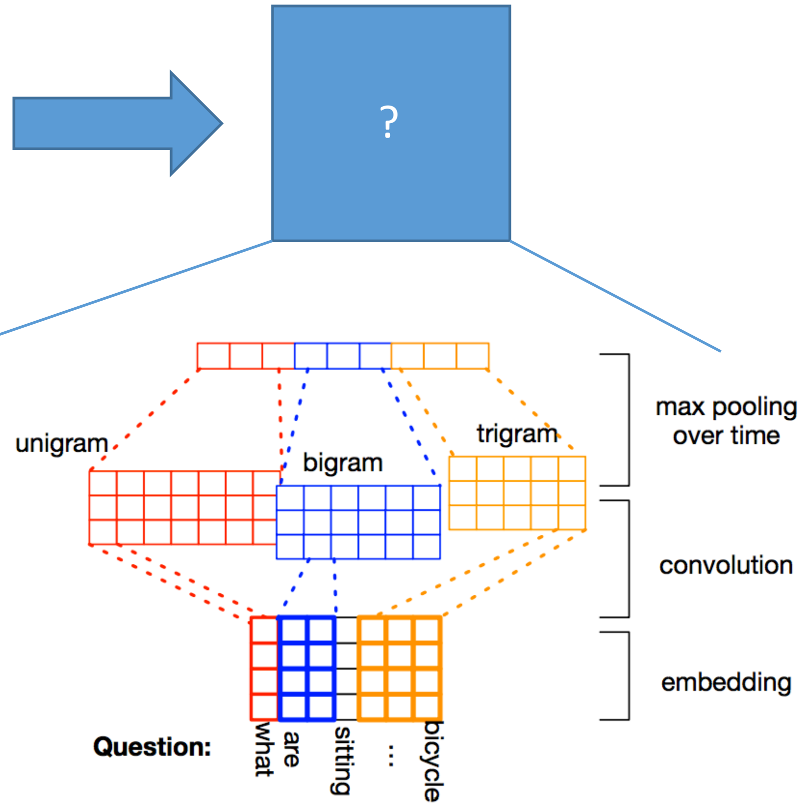
2

LSTM for question representation

$$
\begin{aligned}
x_t &= W_e q_t, t \in \{1, 2, ...T\}, \\
h_t &= \mathrm{LSTM}(x_t), t \in \{1, 2, ...T\}
\end{aligned}
$$

Visual Attention for Question Answering

Question: What are sitting in the basket on a bicycle?



2

CNN for question representation

$$v_Q = [\tilde{h}_1, \tilde{h}_2, \tilde{h}_3]$$

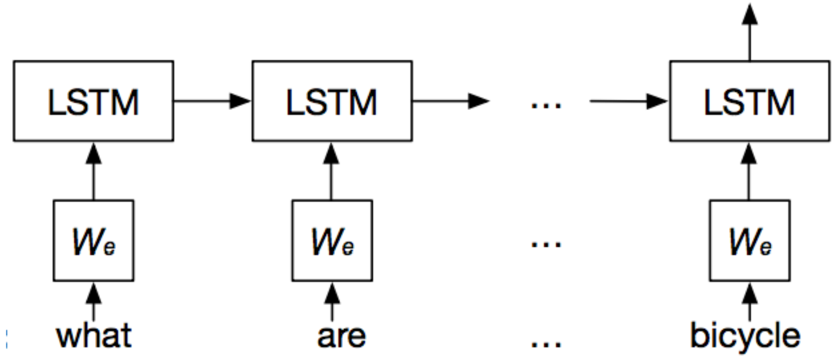
$$\tilde{h}_c = \max_t [h_{c,1}, h_{c,2}, ..., h_{c,T-c+1}]$$

$$h_{c,t} = \tanh(W_c x_{t:t+c-1} + b_c)$$

$$x_t = W_e q_t, t \in \{1, 2, ...T\},$$

Visual Attention for Question Answering

2a LSTM for question representation

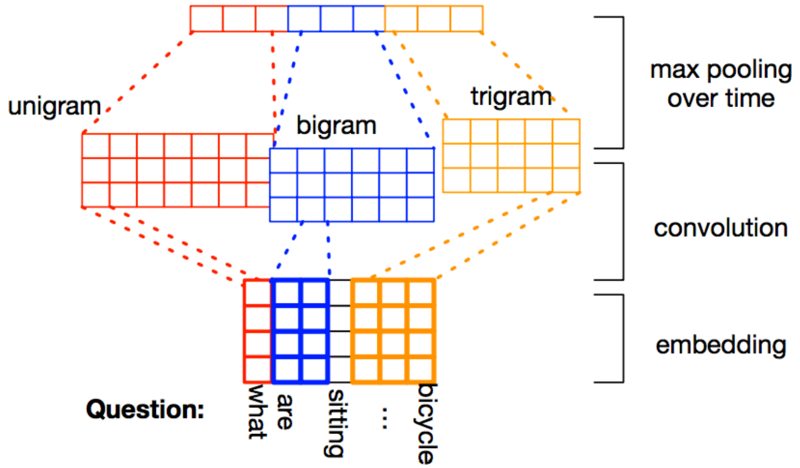


$$v_Q = h_T$$

$$h_t = \text{LSTM}(x_t), t \in \{1, 2, ...T\}$$

$$x_t = W_e q_t, t \in \{1, 2, ...T\},$$

2b CNN for question representation



$$v_Q = [\tilde{h}_1, \tilde{h}_2, \tilde{h}_3]$$

$$\tilde{h}_c = \max_t [h_{c,1}, h_{c,2}, ..., h_{c,T-c+1}]$$

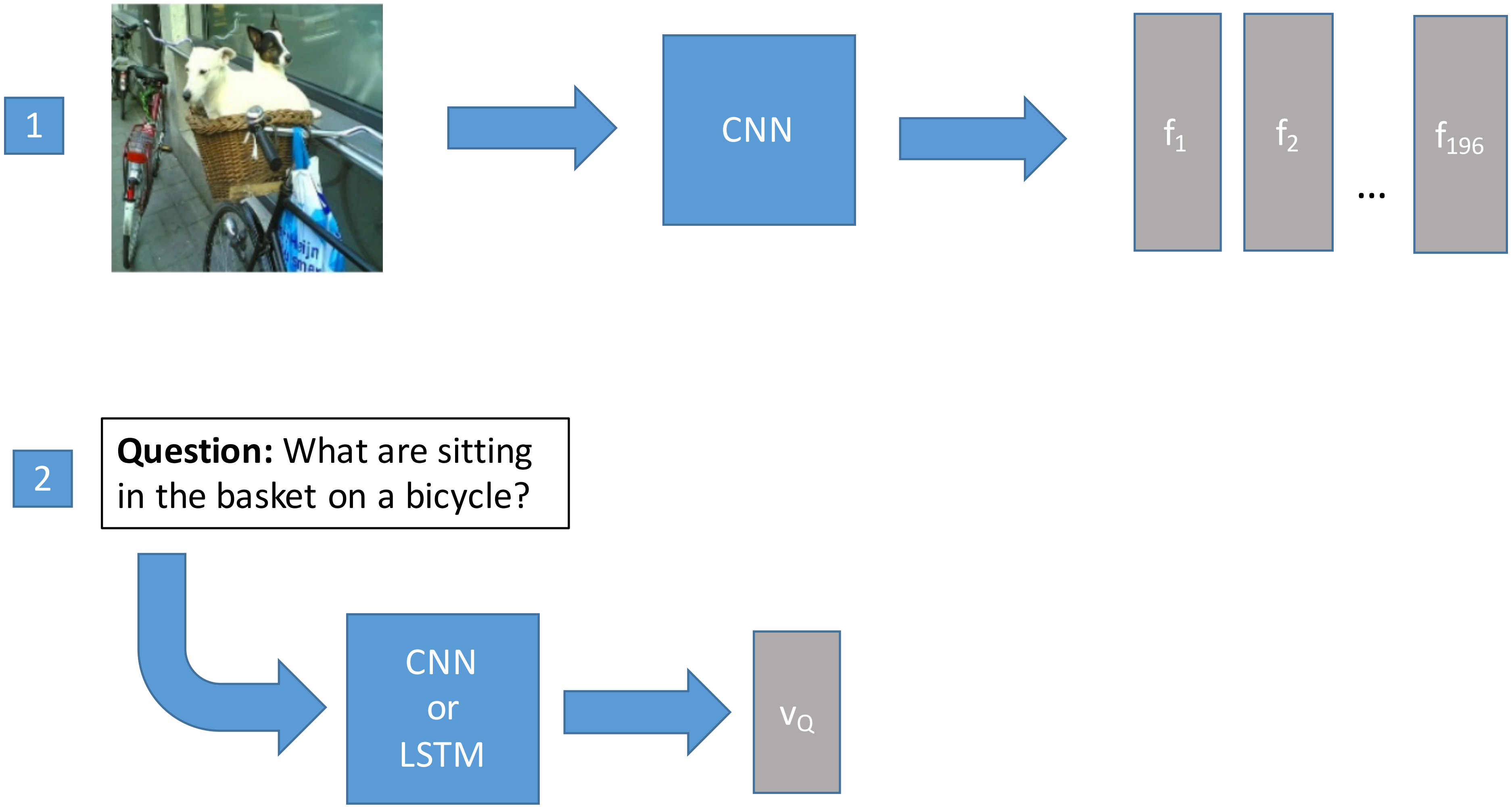
$$h_{c,t} = \tanh(W_c x_{t:t+c-1} + b_c)$$

$$x_t = W_e q_t, t \in \{1, 2, ...T\},$$

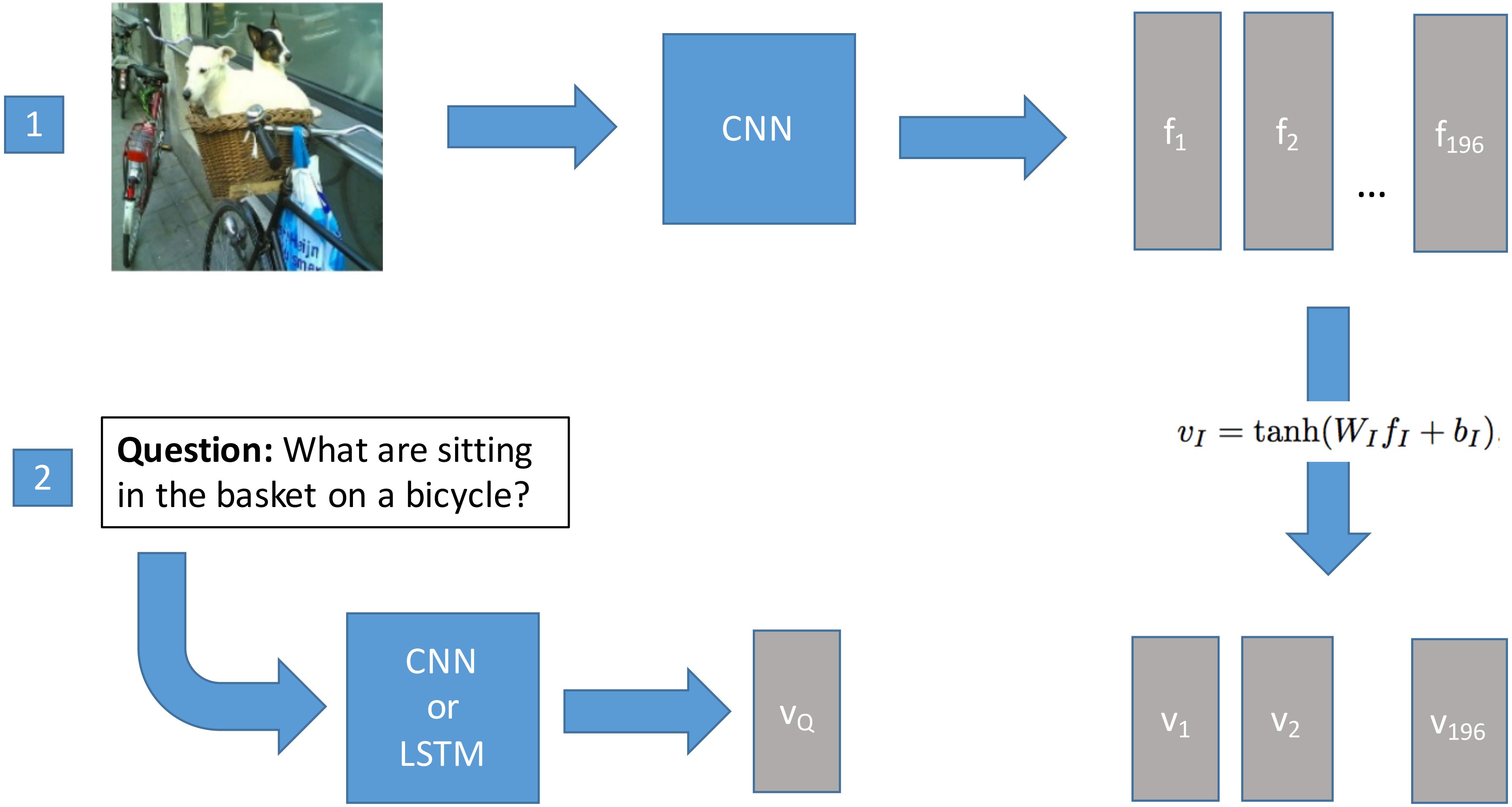
Questions for discussion

- When might we prefer one of these language representations over another?
  - (2a) an LSTM that builds up a representation over multiple time steps
    - Do we keep enough information from early on in the sentence? Should we favor latter parts of a sentence?
  - (2b) a CNN that statically combines word/sentence features at a few scales
    - To what extent does the N used for N-grams affect the resulting representation?

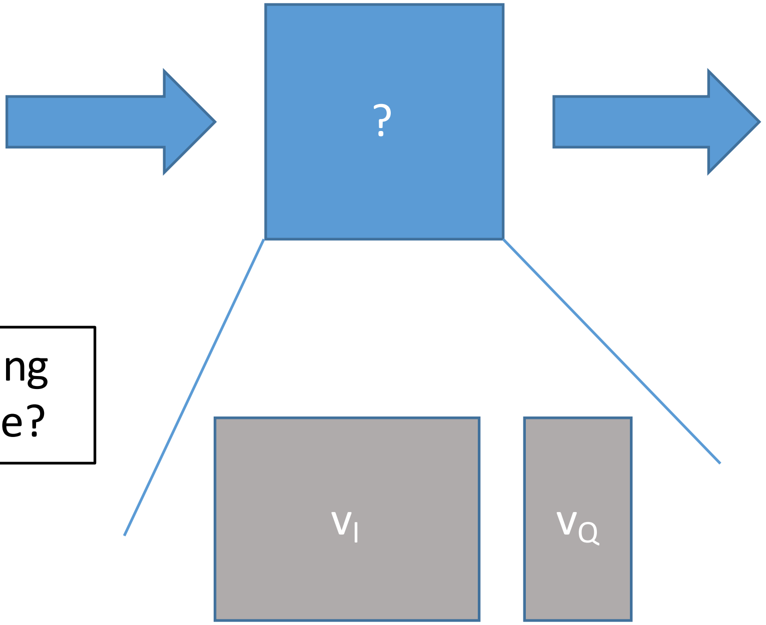
Visual Attention for Question Answering

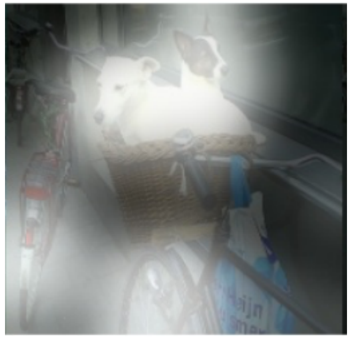


Visual Attention for Question Answering



Visual Attention for Question Answering

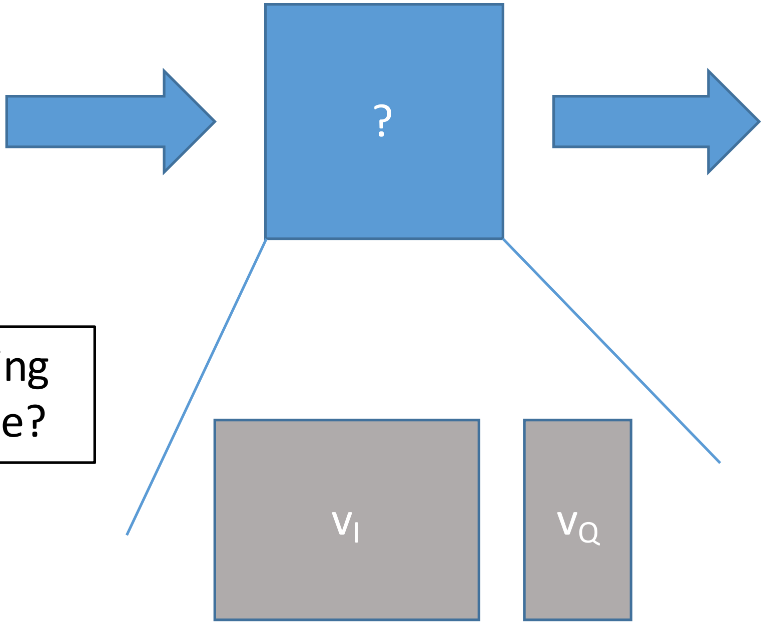


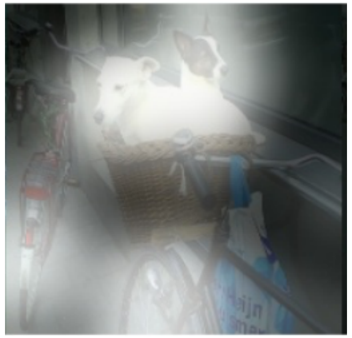


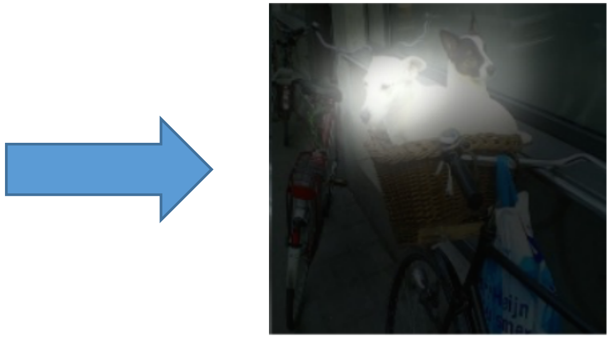
Question: What are sitting in the basket on a bicycle?

$$h_A = \tanh(W_{I,A}v_I \oplus (W_{Q,A}v_Q + b_A)),$$
$$p_I = \mathrm{softmax}(W_P h_A + b_P),$$

Visual Attention for Question Answering







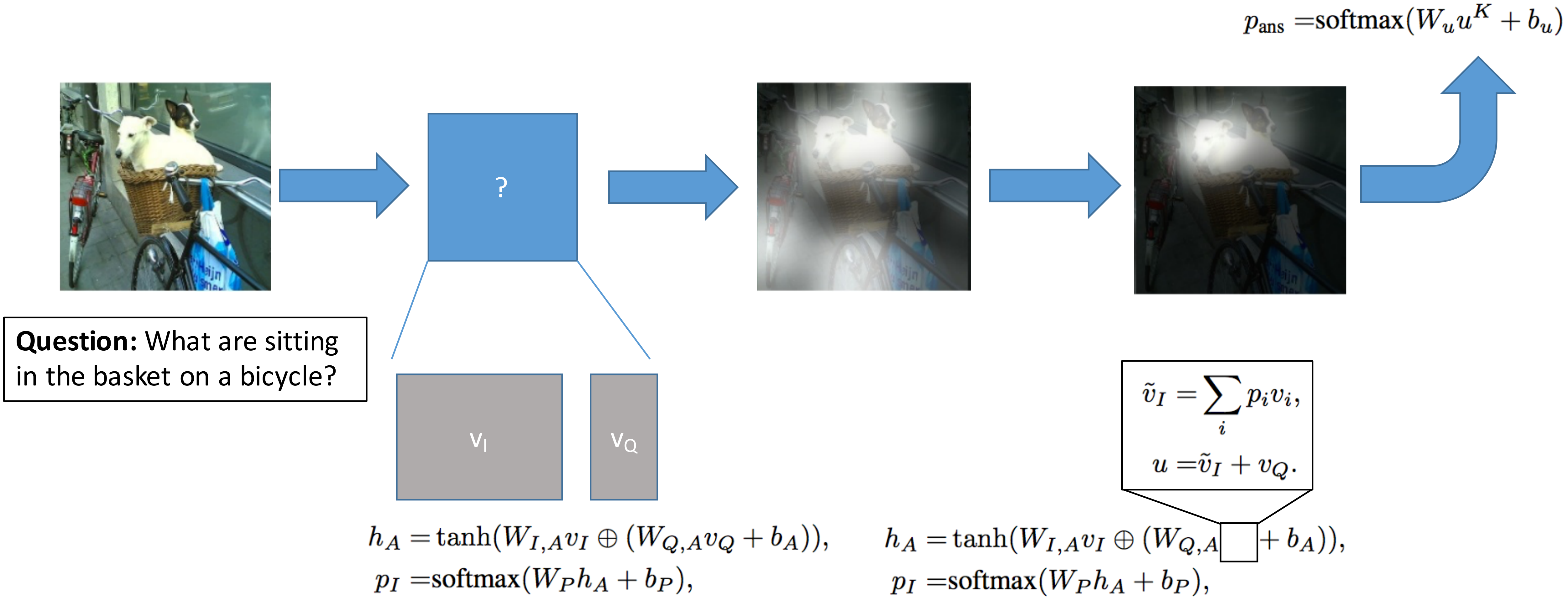
Question: What are sitting in the basket on a bicycle?

$$
\begin{aligned}
h_A =& \tanh(W_{I,A} v_I \oplus (W_{Q,A} v_Q + b_A)),\\
p_I =& \mathrm{softmax}(W_P h_A + b_P),
\end{aligned}
$$

$$
\boxed{
\begin{aligned}
\tilde{v}_I =& \sum_i p_i v_i,\\
u =& \tilde{v}_I + v_Q.
\end{aligned}
}
$$

$$
\begin{aligned}
h_A =& \tanh(W_{I,A} v_I \oplus (W_{Q,A} \boxed{} + b_A)),\\
p_I =& \mathrm{softmax}(W_P h_A + b_P),
\end{aligned}
$$

Visual Attention for Question Answering



Questions for discussion

- When might we want to modulate the question/language representation over time, and when might we prefer to modulate the visual feature representation?
  - Corresponds to recursing on v_Q or v_I

Visual attention **with** deep neural nets

Paper discussions:

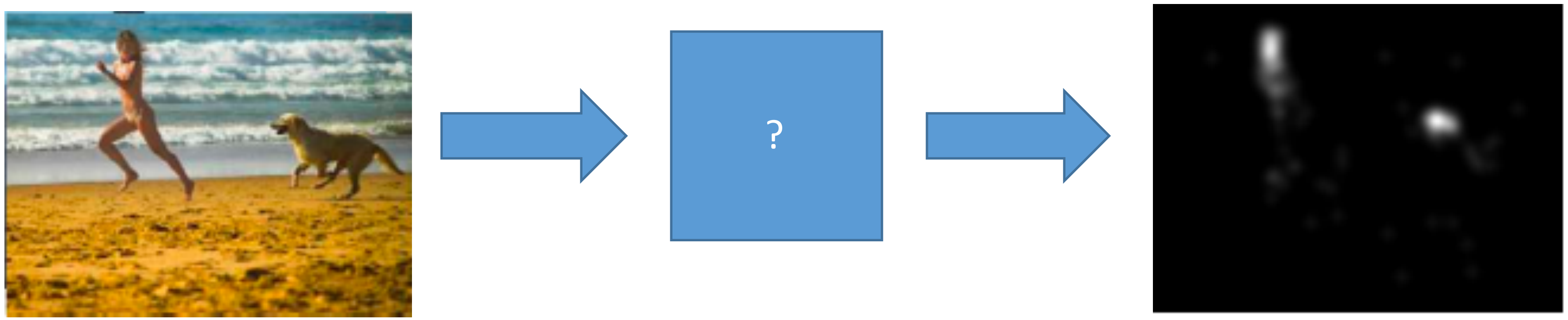
“SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks”, X. Huang, C. Shen, X. Boix, Q. Zhao [CVPR 2015]

“DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations”, S. Kruthiventi, K. Ayush, R. Babu [arXiv Oct 2015]

“Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet”, M. Kümmerer, L. Theis, M. Bethge [ICLR 2015 workshop]

“Predicting Eye Fixations using Convolutional Neural Networks”, N. Liu, J. Han, D. Zhang, S. Wen, T. Liu [CVPR 2015]

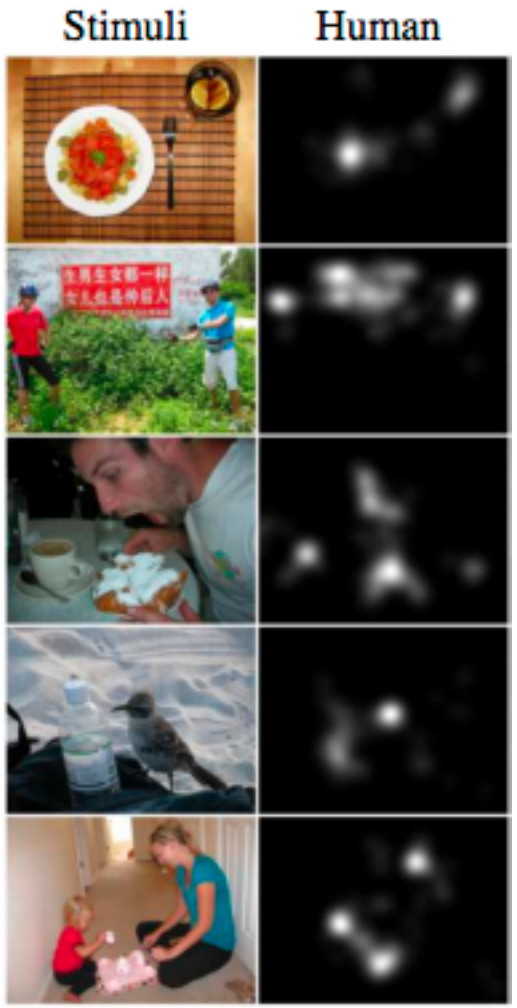
Saliency Prediction with Neural Nets



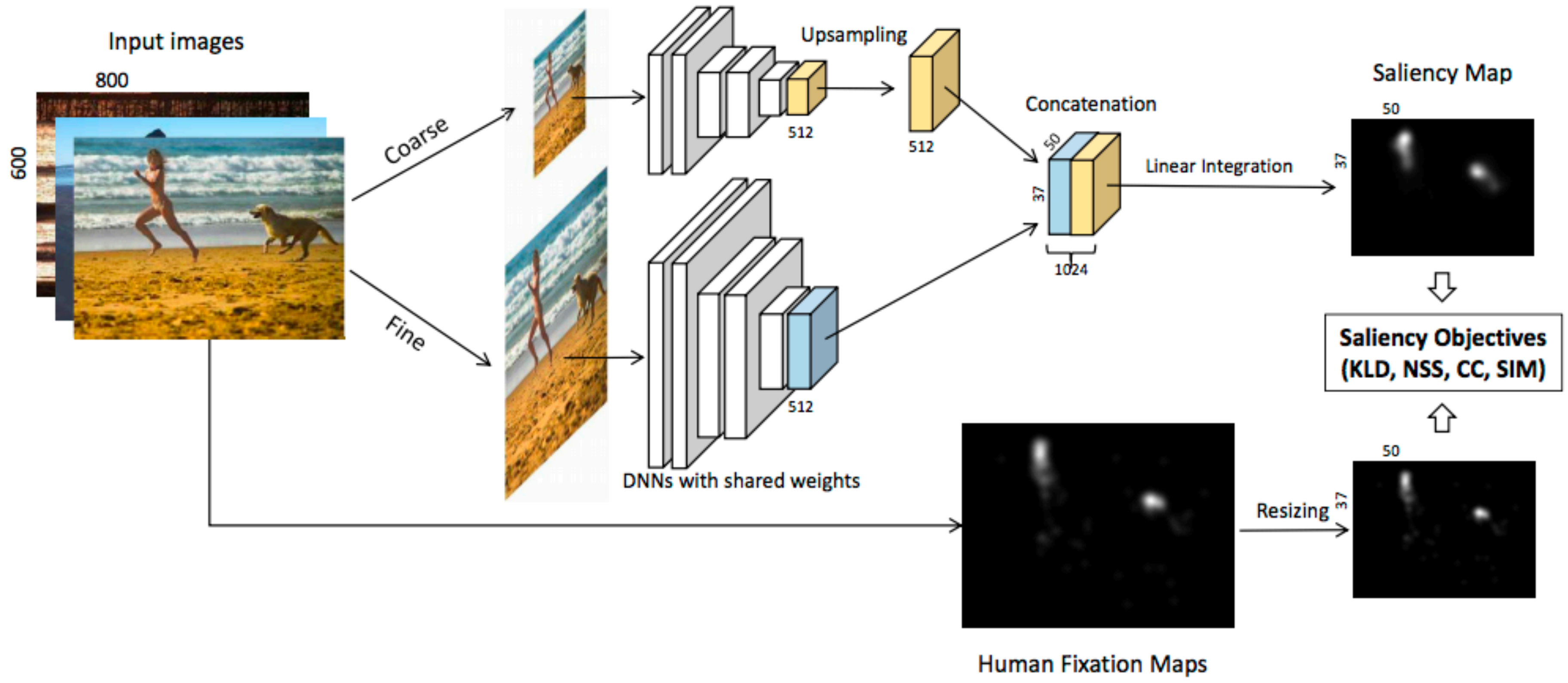
Saliency Prediction with Neural Nets

- Bottom-up pop-out
- Semantic objects of interest
- Salient non-object regions (”abstract concepts”)
- Multi-scale, context-sensitive

- **Challenge:** very small datasets

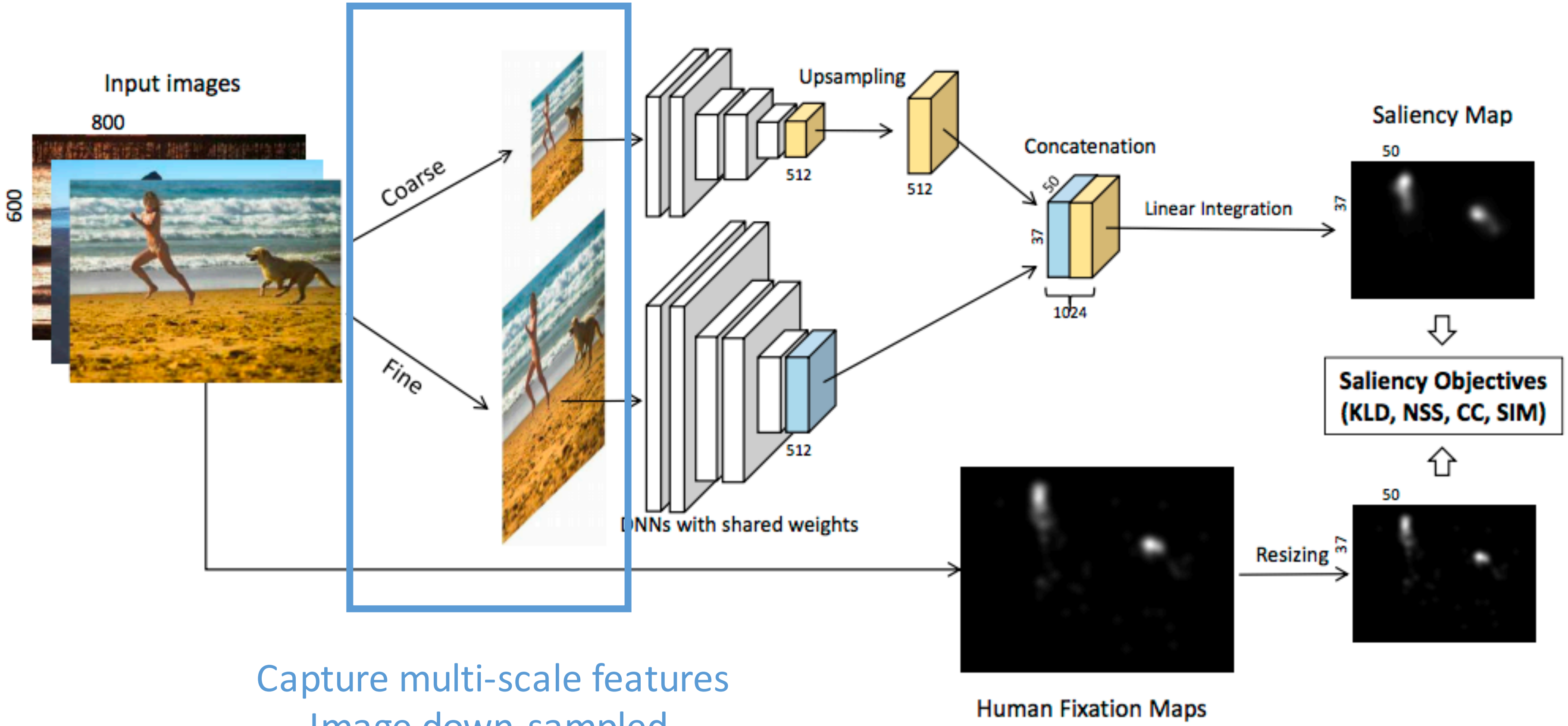


Saliency Prediction with Neural Nets



“SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks”, X. Huang, C. Shen X. Boix, Q. Zhao. [ICCV 2015]

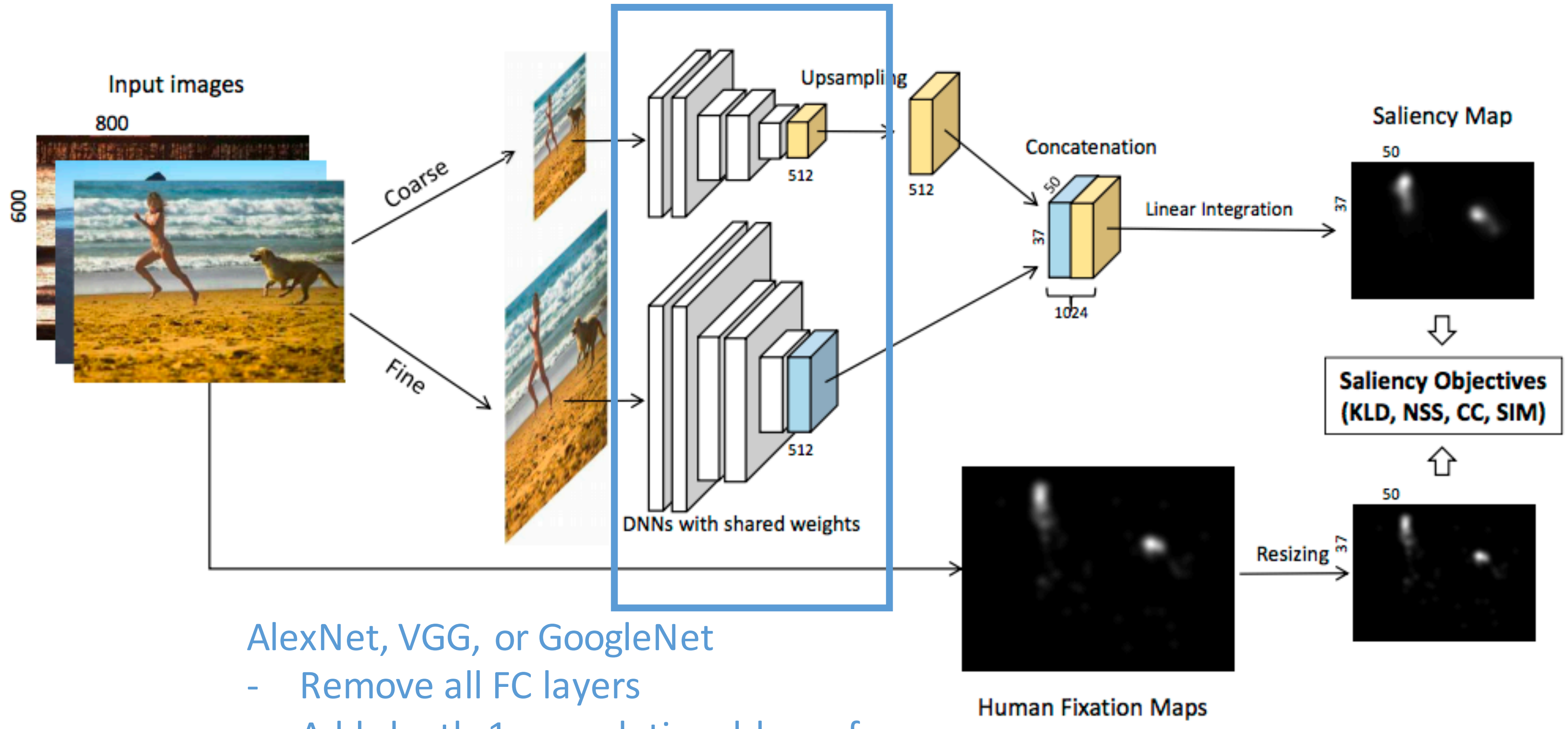
Saliency Prediction with Neural Nets



Capture multi-scale features
- Image down-sampled
- Same DNN applied

"SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks", X. Huang, C. Shen X. Boix, Q. Zhao. [ICCV 2015]

Saliency Prediction with Neural Nets

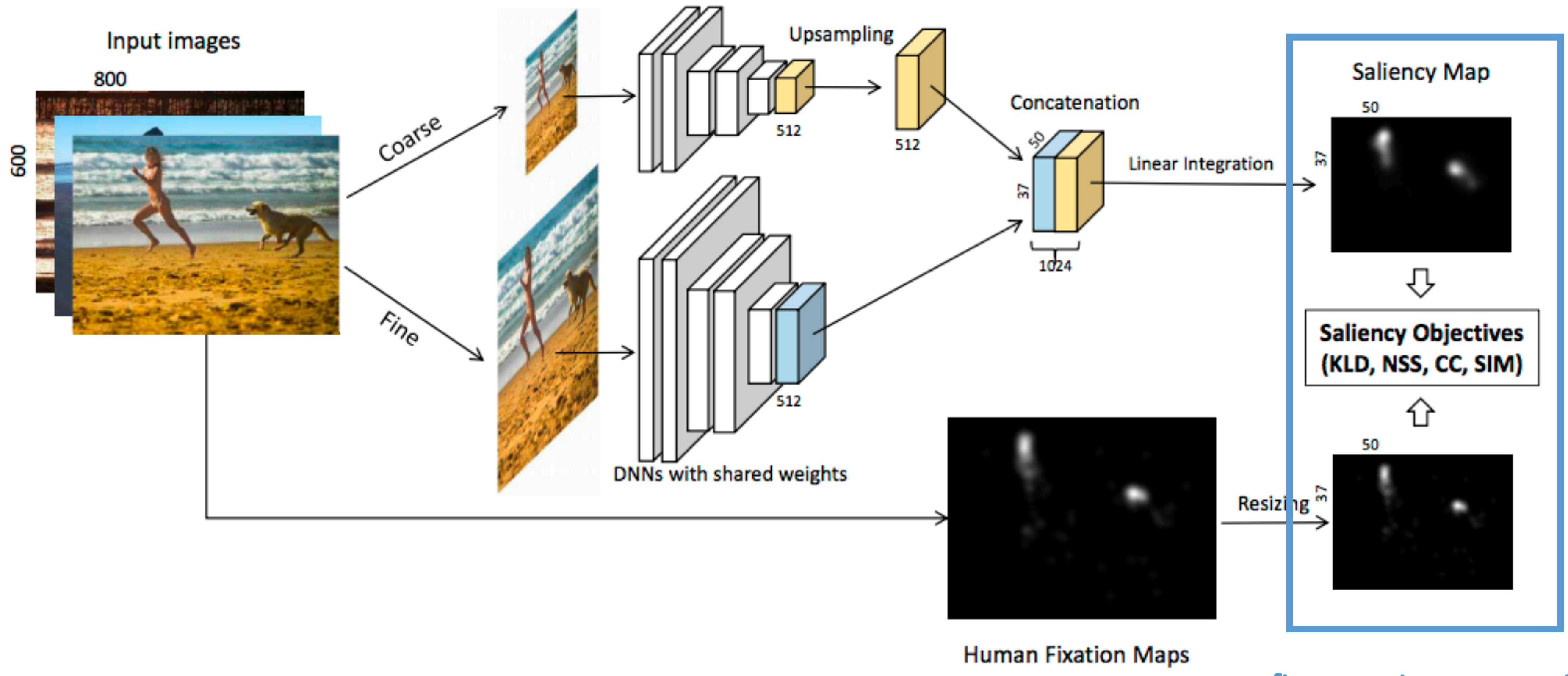


AlexNet, VGG, or GoogleNet

- Remove all FC layers
- Add depth-1 convolutional layer for saliency prediction (after combining responses from both scales)

"SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks", X. Huang, C. Shen X. Boix, Q. Zhao. [ICCV 2015]

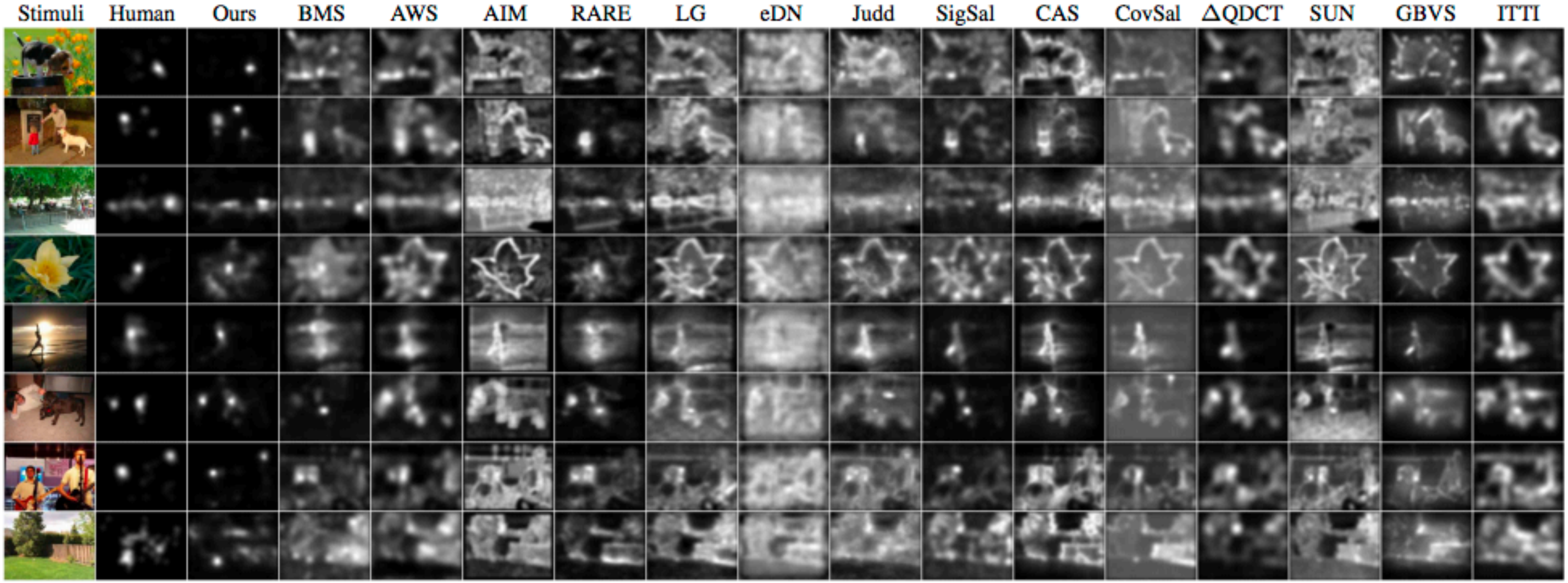
Saliency Prediction with Neural Nets



- fine-tuning pretrained networks
- optimize saliency evaluation metrics directly

"SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks", X. Huang, C. Shen X. Boix, Q. Zhao. [ICCV 2015]

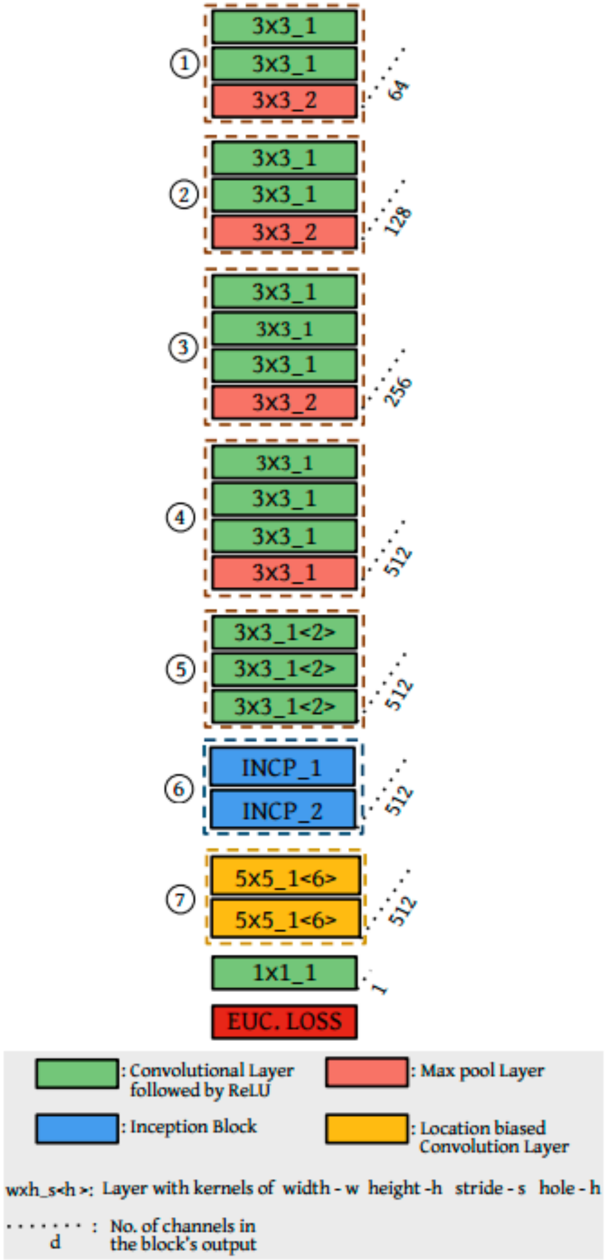
Saliency Prediction with Neural Nets



"SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks", X. Huang, C. Shen X. Boix, Q. Zhao. [ICCV 2015]

Saliency Prediction with Neural Net

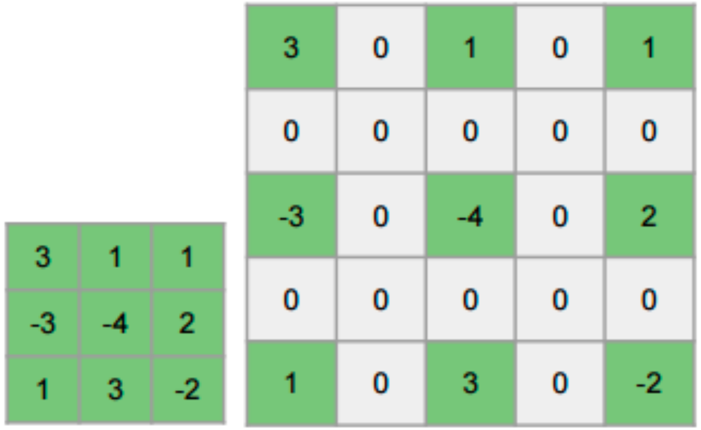
- First 5 layers initialized with VGG-16 weights
  - Note: as channel depth doubles, spatial dimensions are halved with stride-2 pooling layers



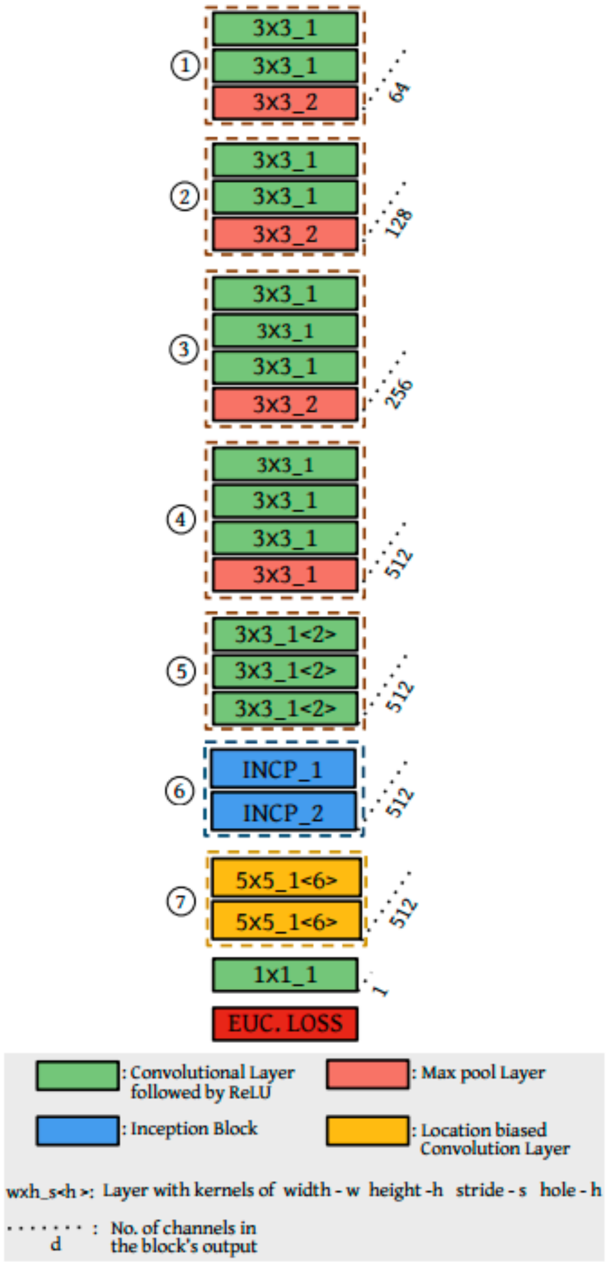
"DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations", S. Kruthiventi, K. Ayush, R. Babu [arXiv Oct 2015]

Saliency Prediction with Neural Net

- First 5 layers initialized with VGG-16 weights
  - Note: as channel depth doubles, spatial dimensions are halved with stride-2 pooling layers
- Holes of size 2 introduced in kernels of 5th layer to increase receptive field without increasing memory footprint



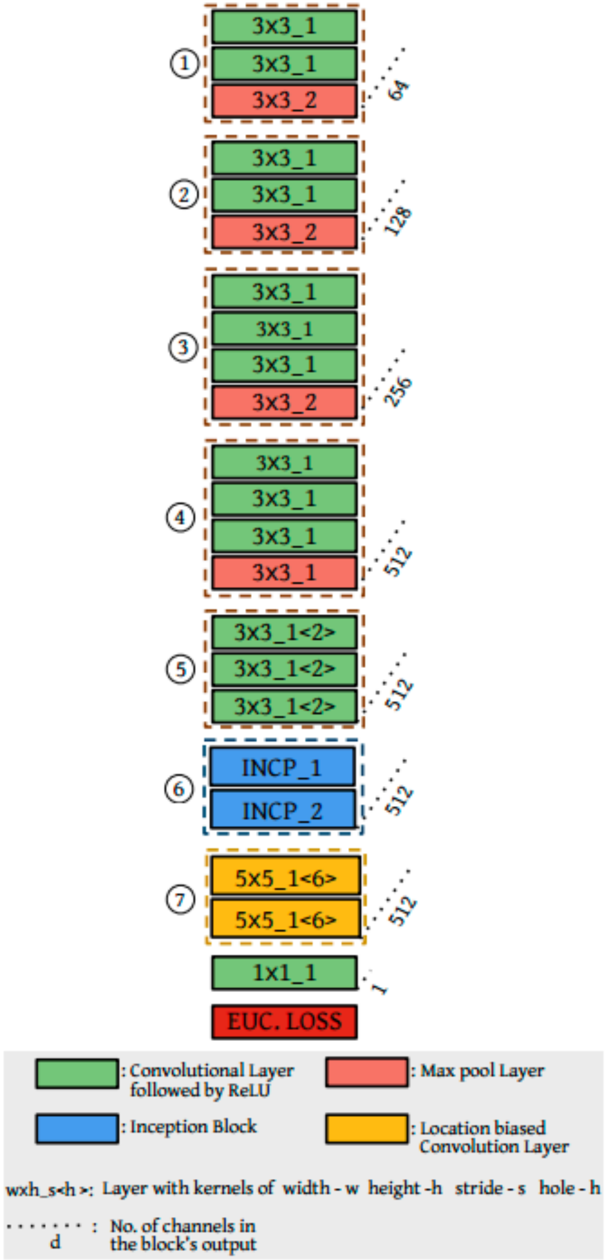
(a) Conv. kernel of size 3x3 (b) Conv. kernel of size 3x3 with hole - 2



"DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations", S. Kruthiventi, K. Ayush, R. Babu [arXiv Oct 2015]

Saliency Prediction with Neural Net

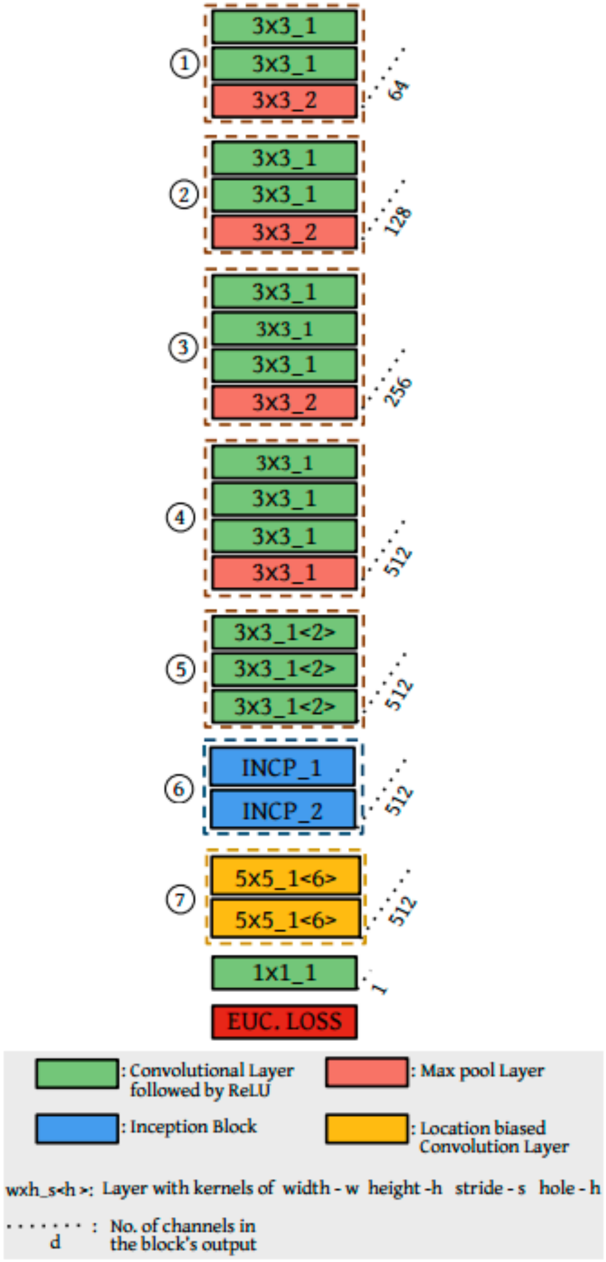
- First 5 layers initialized with VGG-16 weights
  - Note: as channel depth doubles, spatial dimensions are halved with stride-2 pooling layers
- Holes of size 2 introduced in kernels of 5th layer to increase receptive field without increasing memory footprint
- Two inception-style convolutional modules to capture multi-scale semantic structure
- Convolutional layers in 7th layer with holes of size 6 operate on large receptive fields for more global context



"DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations", S. Kruthiventi, K. Ayush, R. Babu [arXiv Oct 2015]

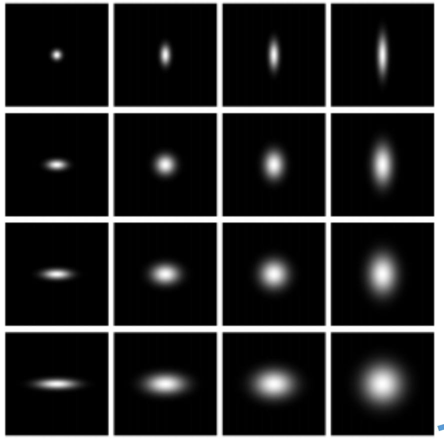
Saliency Prediction with Neural Net

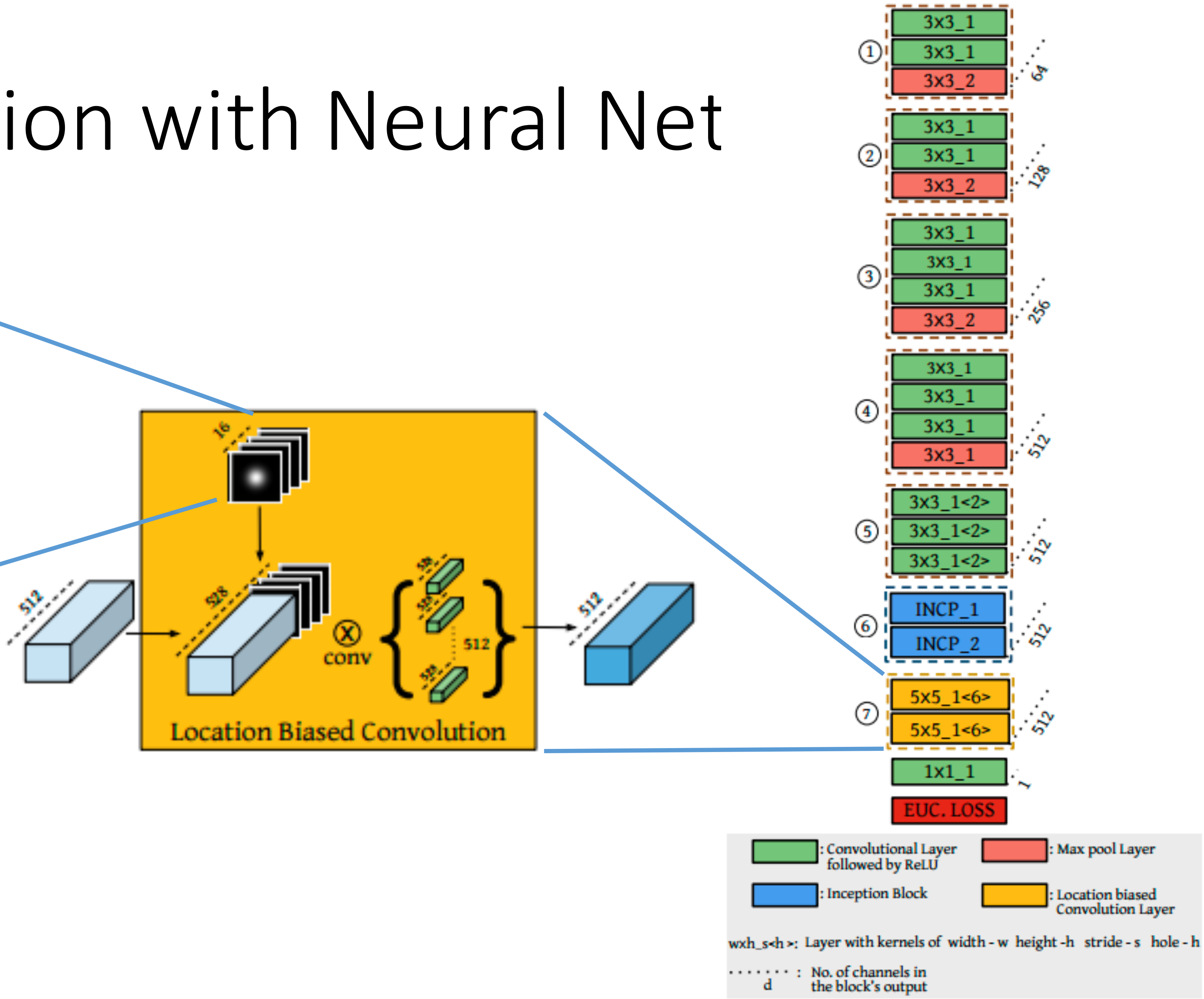
- First 5 layers initialized with VGG-16 weights
  - Note: as channel depth doubles, spatial dimensions are halved with stride-2 pooling layers
- Holes of size 2 introduced in kernels of 5th layer to increase receptive field without increasing memory footprint
- Two inception-style convolutional modules to capture multi-scale semantic structure
- Convolutional layers in 7th layer with holes of size 6 operate on large receptive fields for more global context
- Final layer up-sampled to depth-1 saliency map



"DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations", S. Kruthiventi, K. Ayush, R. Babu [arXiv Oct 2015]

Saliency Prediction with Neural Net

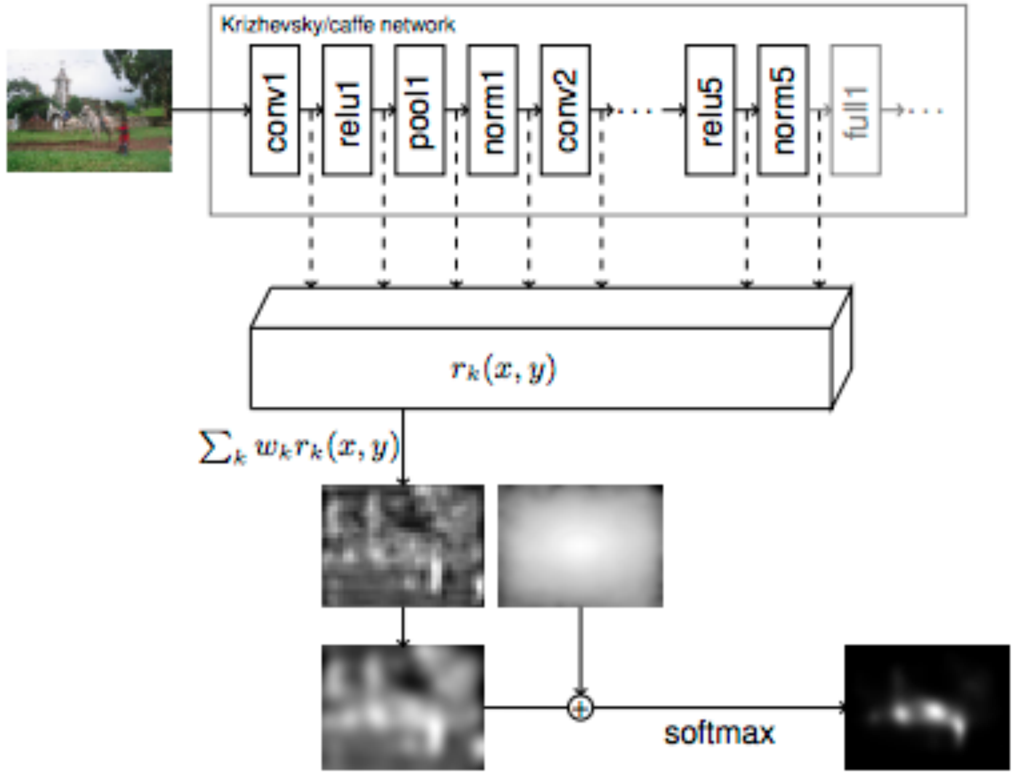




Introducing location-biased behavior without drastically increasing number of network parameters

"DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations", S. Kruthiventi, K. Ayush, R. Babu [arXiv Oct 2015]

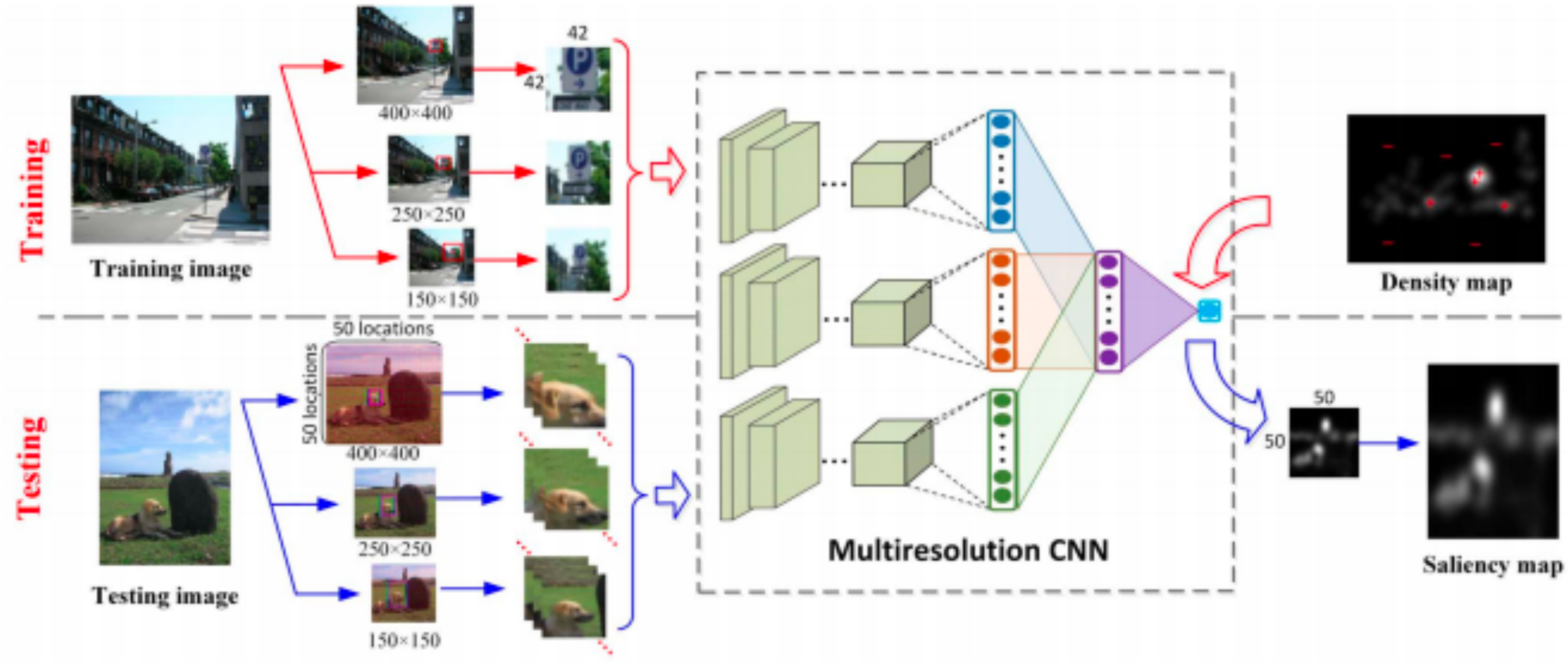
Saliency Prediction with Neural Nets



- Remove final FC layers
- Rescale responses of all other layers to largest size (-> 3712 filter responses per image location)
- Each filter individually normalized across dataset, then Gaussian blurred with some sigma
- Saliency map is weighted combination between these post-processed filters and a center bias
- L1 regularization on weights to encourage sparsity
- Softmax produces final output map

"Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet", M. Kümmerer, L. Theis, M. Bethge [ICLR 2015 workshop]

Fixation Prediction with Neural Nets



"Predicting Eye Fixations using Convolutional Neural Networks", N. Liu, J. Han, D. Zhang, S. Wen, T. Liu [CVPR 2015]

Questions for discussion

- Can directly optimizing for visual attention/saliency lead to benefits for other computer vision applications **or** should visual attention naturally come out of the specific application?
- Can models of visual attention/saliency help bootstrap individual tasks or lead to generalization across tasks?