Day 4 Lecture 6

Attention Models
Attention Models: Motivation

The whole input volume is used to predict the output...

Image: H x W x 3
Attention Models: Motivation

Image: H x W x 3

The whole input volume is used to predict the output...

...despite the fact that not all pixels are equally important
Attention Models: Motivation

Attention models can relieve computational burden

Helpful when processing big images!
Attention Models: Motivation

Attention models can relieve computational burden

Helpful when processing big images!
Encoder & Decoder

From previous lecture...

The whole input sentence is used to produce the translation

Attention Models

Bahdanau et al. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015
Attention Models

Idea: Focus in different parts of the input as you make/refine predictions in time

E.g.: Image Captioning

A bird flying over a body of water
LSTM Decoder

The LSTM decoder “sees” the input only at the beginning!
Attention for Image Captioning

Image: $H \times W \times 3$

Features: $L \times D$

Slide Credit: CS231n
Attention for Image Captioning

Image: $H \times W \times 3$

Features: $L \times D$

Attention weights $(L \times D)$

$h_0$

$\mathbf{a}_1$

Slide Credit: CS231n
Attention for Image Captioning

Image: $H \times W \times 3$

Features: $L \times D$

Weighted combination of features

Weighted features: $D$

Attention weights (LxD)

Predicted word

First word

Weighted combination

CNN
Attention for Image Captioning

Image: H x W x 3

Features: L x D

Weighted combination of features

Weighted features: D

Attention weights (LxD)

First word

predicted word

Slide Credit: CS231n
Attention for Image Captioning

14x14 Feature Map

1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015
Attention for Image Captioning

Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015
Attention for Image Captioning

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.

---

Xu et al. [Show, Attend and Tell: Neural Image Caption Generation with Visual Attention](https://icml.cc/2015/papers/1227.pdf), ICML 2015
Soft Attention

**CNN**

**Image:** $H \times W \times 3$

**Grid of features** (Each D-dimensional)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>d</td>
<td></td>
</tr>
</tbody>
</table>

**Distribution over grid locations**

$p_a + p_b + p_c + p_d = 1$

**Context vector $z$** (D-dimensional)

$$z = p_a a + p_b b + p_c c + p_d d$$

Derivative $dz/dp$ is nice!

Train with gradient descent

---

**Soft attention:**

Summarize ALL locations


---

Slide Credit: **CS231n**
Soft Attention

Image: H x W x 3

Grid of features (Each D-dimensional)

\[
p_a + p_b + p_c + p_d = 1
\]

Soft attention:
Summarize ALL locations
\[
z = p_a a + p_b b + p_c c + p_d d
\]

Differentiable function
Train with gradient descent

- Still uses the whole input!
- Constrained to fix grid

Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

Slide Credit: CS231n
Hard attention

Input image: 
H x W x 3

Box Coordinates: 
(xc, yc, w, h)

Cropped and rescaled image: 
X x Y x 3

Gradient is 0 almost everywhere
Gradient is undefined at x = 0

Hard attention:
Sample a subset of the input

Not a differentiable function!
Can’t train with backprop :(

need reinforcement learning
Hard attention

**Classify** images by attending to arbitrary regions of the *input*

**Generate** images by attending to arbitrary regions of the *output*

Gregor et al. DRAW: A Recurrent Neural Network For Image Generation. ICML 2015
Hard attention

Gregor et al. DRAW: A Recurrent Neural Network For Image Generation. ICML 2015
Hard attention

Read text, generate handwriting using an RNN that attends at different arbitrary regions over time

Hard attention

Input image: $H \times W \times 3$

Box Coordinates: $(x_c, y_c, w, h)$

Cropped and rescaled image: $X \times Y \times 3$

Not a differentiable function!

Can’t train with backprop :(
Spatial Transformer Networks

Input image: \(H \times W \times 3\)

Box Coordinates: \((xc, yc, w, h)\)

Cropped and rescaled image: \(X \times Y \times 3\)

Not a differentiable function!

Can't train with backprop :(

Make it differentiable

Train with backprop :)

Jaderberg et al. Spatial Transformer Networks, NIPS 2015
Spatial Transformer Networks

**Input image:**
\( H \times W \times 3 \)

**Box Coordinates:**
\( (x_c, y_c, w, h) \)

**Cropped and rescaled image:**
\( X \times Y \times 3 \)

**Can we make this function differentiable?**

**Idea:** Function mapping 
*pixel coordinates* \((x_t, y_t)\) of output to 
*pixel coordinates* \((x_s, y_s)\) of input

\[
\begin{pmatrix}
  x_s^i \\
  y_s^i
\end{pmatrix} =
\begin{bmatrix}
  \theta_{11} & \theta_{12} & \theta_{13} \\
  \theta_{21} & \theta_{22} & \theta_{23}
\end{bmatrix}
\begin{pmatrix}
  x_t^i \\
  y_t^i \\
  1
\end{pmatrix}
\]

Repeat for all pixels in *output* to get a
**sampling grid**

Then use **bilinear interpolation** to compute output

Network attends to input by predicting \( \theta \)

Jaderberg et al. *Spatial Transformer Networks*. NIPS 2015

Slide Credit: [CS231n](#)
Spatial Transformer Networks

Differentiable module

Easy to incorporate in any network, anywhere!

Insert spatial transformers into a classification network and it learns to attend and transform the input

Jaderberg et al. Spatial Transformer Networks. NIPS 2015
Spatial Transformer Networks

Fine-grained classification

Jaderberg et al. Spatial Transformer Networks. NIPS 2015
Visual Attention

Visual Attention

Action Recognition in Videos


Kuen et al. Recurrent Attentional Networks for Saliency Detection. CVPR 2016

Salient Object Detection
Other examples

Resources

• CS231n Lecture @ Stanford [slides][video]
• More on Reinforcement Learning
• Soft vs Hard attention
• Handwriting generation demo
• Spatial Transformer Networks - Slides & Video by Victor Campos
• Attention implementations:
  ○ Seq2seq in Keras
  ○ DRAW & Spatial Transformers in Keras
  ○ DRAW in Lasagne
  ○ DRAW in Tensorflow