CS6501: Deep Learning for Visual Recognition
Seq2Seq Model & Text-to-Image Synthesis

Presenter: Fuwen Tan
Today’s Class

• Mini-batch training of the RNN model
  • Special “End-of-Sequence” token: <end>
  • Padding

• Sequence-to-sequence model
  • Neural Machine Translation[1]

• Text-to-Image Synthesis[2]

[1] Effective Approaches to Attention-based Neural Machine Translation. Thang Luong, Hieu Pham, and Christopher D. Manning. EMNLP 2015
A RNN model will never end

“Hello”, “world”, “!”, “!”, “!”, “!”, “!”, …
Unless: set the maximum length before hand

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>“hello”</th>
<th>“world”</th>
<th>“java”</th>
<th>“is”</th>
<th>“better”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 2</td>
<td>“hello”</td>
<td>“hoos”</td>
<td>“I”</td>
<td>“like”</td>
<td>“python”</td>
</tr>
<tr>
<td>Sample 3</td>
<td>“one”</td>
<td>“plus”</td>
<td>“eight”</td>
<td>“equals”</td>
<td>“to”</td>
</tr>
</tbody>
</table>

I want sentences of 5 words
Or: learn to predict the END.

Ground-truth: “Hello”, “world”, “<end>”

Training: learn to generate the ground-truth sequence with “<end>”. Testing: generate the sequence until an “<end>” is predicted.
Computing loss: what if \#ground-truth \neq \#prediction

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>“hello”</th>
<th>“world”</th>
<th>“&lt;end&gt;”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction 1</td>
<td>“hello”</td>
<td>“&lt;end&gt;”</td>
<td>“foo”</td>
</tr>
<tr>
<td>Prediction 2</td>
<td>“hello”</td>
<td>“how”</td>
<td>“are”</td>
</tr>
</tbody>
</table>

\[ \text{loss}_1 \quad \text{loss}_2 \quad \text{loss}_3 \]
Mini-batch training: padding

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>“hello”</th>
<th>“how”</th>
<th>“are”</th>
<th>“you”</th>
<th>“today”</th>
<th>“&lt;end&gt;”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 2</td>
<td>“a”</td>
<td>“dog”</td>
<td>“is”</td>
<td>“driving”</td>
<td>“&lt;end&gt;”</td>
<td>“&lt;pad&gt;”</td>
</tr>
<tr>
<td>Sample 3</td>
<td>“hello”</td>
<td>“world”</td>
<td>“&lt;end&gt;”</td>
<td>“&lt;pad&gt;”</td>
<td>“&lt;pad&gt;”</td>
<td>“&lt;pad&gt;”</td>
</tr>
</tbody>
</table>
## Mini-batch training: padding

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>“hello”</th>
<th>“how”</th>
<th>“are”</th>
<th>“you”</th>
<th>“today”</th>
<th>“&lt;end&gt;”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 2</td>
<td>“a”</td>
<td>“dog”</td>
<td>“is”</td>
<td>“driving”</td>
<td>“&lt;end&gt;”</td>
<td>“&lt;pad&gt;”</td>
</tr>
<tr>
<td>Sample 3</td>
<td>“hello”</td>
<td>“world”</td>
<td>“&lt;end&gt;”</td>
<td>“&lt;pad&gt;”</td>
<td>“&lt;pad&gt;”</td>
<td>“&lt;pad&gt;”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>1.0</th>
<th>1.0</th>
<th>1.0</th>
<th>1.0</th>
<th>1.0</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 2</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Sample 3</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Generating text that makes sense: Language Model

Unconditional: $h_0 = 0$
Generating text with a goal: Machine Translation

I am a student

Conditional: $h_0 = \overline{h}_4$

Seq2Seq model

$\texttt{Seq2Seq:} \quad \text{"suis"} \quad \leftarrow \quad f(h_2) = \text{softmax}(W_s h_2)$
Seq2Seq model with perfect word alignments

Seq2Seq:  

```
<table>
<thead>
<tr>
<th>l</th>
<th>am</th>
<th>a</th>
<th>student</th>
</tr>
</thead>
</table>
```

```
<table>
<thead>
<tr>
<th>je</th>
<th>suis</th>
<th>un</th>
<th>étudiant</th>
</tr>
</thead>
</table>
```

```
<table>
<thead>
<tr>
<th>f(h₂)</th>
</tr>
</thead>
</table>
```

Ideally:

```
<table>
<thead>
<tr>
<th>l</th>
<th>am</th>
<th>a</th>
<th>student</th>
</tr>
</thead>
</table>
```

```
<table>
<thead>
<tr>
<th>je</th>
<th>suis</th>
<th>un</th>
<th>étudiant</th>
</tr>
</thead>
</table>
```

```
<table>
<thead>
<tr>
<th>f'(h₂, &quot;am&quot;)</th>
</tr>
</thead>
</table>
```

Seq2Seq model with perfect word alignments

Seq2Seq: "suis" ← $f(h_2)$

Ideally: "suis" ← $f'(h_2, "am")$

Or: "suis" ← $f'(h_2, \overline{h}_2)$
Seq2Seq model with attention

Ideally: “suis” ← \( f'(h_2, \bar{h}_2) \)

In practice: “suis” ← \( f'(h_2, c_2) \)

\[
c_2 = \sum_{k=1}^{4} w_{2,k} \bar{h}_k
\]

Pray that \( S: w_{2,2} = 1, w_{2,k\neq2} = 0 \) is true

Or train the model such that \( S \) is almost true
Seq2Seq model with attention

Key assumption: \( \overline{h}_2 \approx h_2 \approx h_0 - "je" \approx \overline{h}_4 - "je" \)

\[
w_{2,k} = \frac{\exp \left( \text{score}(h_2, \overline{h}_k) \right)}{\sum_j \exp \left( \text{score}(h_2, \overline{h}_j) \right)}
\]

\[
\text{score}(h_2, \overline{h}_k) = h_2^T W_a \overline{h}_k
\]

[1] Effective Approaches to Attention-based Neural Machine Translation. Thang Luong, Hieu Pham, and Christopher D. Manning. EMNLP 2015
Seq2Seq model with attention

\[ f'(h_2, c_2) = \text{softmax}(W_s \ \text{tanh}(W_c[h_2; c_2])) \]

“suis”
Perform much better for long sequences

Figure 6: **Length Analysis** – translation qualities of different systems as sentences become longer.

[1] Effective Approaches to Attention-based Neural Machine Translation. Thang Luong, Hieu Pham, and Christopher D. Manning. EMNLP 2015
Also very helpful in image captioning

ECCV 2018 accepted 776 papers

38 of them with “attention” in their titles
**Seq2Seq vs Text-to-Image Synthesis**

Sentence: composition of words

Image: composition of patches
Can we do this?
Challenges

Machine Translation:
“I am a student” $\rightarrow$ “je suis un étudiant”

Text-to-Image Synthesis:
"A person is holding a surfboard"
Challenges: in each step

Machine Translation:

student → étudiant

Text-to-Image Synthesis:

"A person is holding a surfboard"

- object category: person, surfboard
- location: somewhere in the 2D world
- attributes: size, pose, expression, …
Challenges: in each step

Text-to-Image Synthesis:

"A person is holding a surfboard"

- **object category**: person
- **location**: somewhere in the 2D world
- **attributes**: size, pose, expression, …

Learning the distributions of categories, locations, attributes from the training samples
Mike holds a hotdog

<table>
<thead>
<tr>
<th>object</th>
<th>location</th>
<th>attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clip-art of</td>
<td>somewhere on the</td>
<td>pose: hold size, ...</td>
</tr>
<tr>
<td>&quot;Mike&quot;</td>
<td>ground</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clip-art of</td>
<td>In Mike's hand</td>
<td>size: &lt; Mike</td>
</tr>
<tr>
<td>&quot;hotdog&quot;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mike
holds
a
hotdog
Task 1: Abstract Scene Generation

“Mike is surprised at the duck. The duck is standing on the grill. Jenny is running towards Mike and the duck.”
Task 1: Abstract Scene Generation

Object category

58 clip-art objects

Location

28 x 28 grid

Attributes

3 sizes, 2 orientations, 7 poses and 5 expressions for “Mike” and “Jenny”.
Task 2: Scene Layout Generation

“A guy on a motorcycle with some people watching.”
Task 2: Scene Layout Generation

Object category
- 80 object categories from COCO:
  "person", "car", "chair", …

Location
- 28 x 28 grid

Attributes
- 17 sizes, 17 aspect-ratios
Task 3: Composite Image Generation

“Several elephants walking together in a line near water.”
Task 3: Composite Image Generation

Object category

95 object & stuff categories from COCO: “person”, “grass”, “sky”, ...

Location

32 x 32 grid

Attributes

17 sizes, 17 aspect-ratios

a feature vector for patch retrieval
Step-by-step generation of Abstract Scene

Mike is holding a hotdog.  
Jenny is sitting in the sandbox.  
Jenny is holding the shovel.

- **object attn:** sitting sandbox holding
- **attribute attn:** jenny <eos> jenny
- **object attn:** sandbox sitting mike
- **attribute attn:** sandbox <eos> jenny
- **object attn:** mike jenny sitting
- **attribute attn:** holding hotdog mike
- **object attn:** jenny jenny mike
- **attribute attn:** sitting jenny holding
- **object attn:** hotdog shovel holding
- **attribute attn:** shovel holding jenny holding
- **object attn:** shovel holding sandbox
- **attribute attn:** shovel holding <eos>
Step-by-step generation of composite image

Inputs: a room with TV and some different types of couches.

couches, room, TV
room, couches, <eos>
 eos>, room, types

types, of, and

Generated sequences and the top-3 attended words in each step.

Reference image
Step-by-step generation of composite image

Inputs: a person walks on the beach, carrying a surf board.

Generated sequences and the top-3 attended words in each step.

Reference image
<table>
<thead>
<tr>
<th>Input</th>
<th>Zitnick et al. 2013</th>
<th>Text2Scene (w/o Attention)</th>
<th>Text2Scene</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jenny is wearing sunglasses. Mike is holding the red shovel.</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>Mike is wearing a viking head.</td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
</tr>
<tr>
<td>Mike went down the slide fast. Jenny is worried that Mike is hurt.</td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td>Jenny is wearing a chef hat.</td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
<td><img src="image16" alt="Image" /></td>
</tr>
<tr>
<td>Mike is angry at Jenny.</td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
</tr>
<tr>
<td>Jenny is sad that Mike took the frisbee. The pizza is on the table.</td>
<td><img src="image21" alt="Image" /></td>
<td><img src="image22" alt="Image" /></td>
<td><img src="image23" alt="Image" /></td>
<td><img src="image24" alt="Image" /></td>
</tr>
<tr>
<td>Jenny is holding a bucket and shovel.</td>
<td><img src="image25" alt="Image" /></td>
<td><img src="image26" alt="Image" /></td>
<td><img src="image27" alt="Image" /></td>
<td><img src="image28" alt="Image" /></td>
</tr>
<tr>
<td>Mike fell off the swingset. There is rain and lightning in the sky.</td>
<td><img src="image29" alt="Image" /></td>
<td><img src="image30" alt="Image" /></td>
<td><img src="image31" alt="Image" /></td>
<td><img src="image32" alt="Image" /></td>
</tr>
</tbody>
</table>
## More examples

<table>
<thead>
<tr>
<th>Input Caption</th>
<th>Predicted Layout</th>
<th>Reference Layout</th>
<th>Reference Image</th>
<th>Input Caption</th>
<th>Predicted Layout</th>
<th>Reference Layout</th>
<th>Reference Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>An attractive young <strong>woman</strong> leads a grey horse through a paddock.</td>
<td><img src="image1.png" alt="Predicted layout" /></td>
<td><img src="image2.png" alt="Reference layout" /></td>
<td><img src="image3.png" alt="Reference image" /></td>
<td>A couple of <strong>women</strong> ride horses through some water.</td>
<td><img src="image4.png" alt="Predicted layout" /></td>
<td><img src="image5.png" alt="Reference layout" /></td>
<td><img src="image6.png" alt="Reference image" /></td>
</tr>
<tr>
<td>Two giraffes in a zoo enjoy a walk and a snack.</td>
<td><img src="image7.png" alt="Predicted layout" /></td>
<td><img src="image8.png" alt="Reference layout" /></td>
<td><img src="image9.png" alt="Reference image" /></td>
<td>A cat standing <strong>next to</strong> an open refrigerator door.</td>
<td><img src="image10.png" alt="Predicted layout" /></td>
<td><img src="image11.png" alt="Reference layout" /></td>
<td><img src="image12.png" alt="Reference image" /></td>
</tr>
<tr>
<td>A person <strong>holding a surf board</strong> in a body of water.</td>
<td><img src="image13.png" alt="Predicted layout" /></td>
<td><img src="image14.png" alt="Reference layout" /></td>
<td><img src="image15.png" alt="Reference image" /></td>
<td>This is a man <strong>riding a board</strong> in the water.</td>
<td><img src="image16.png" alt="Predicted layout" /></td>
<td><img src="image17.png" alt="Reference layout" /></td>
<td><img src="image18.png" alt="Reference image" /></td>
</tr>
<tr>
<td>A laptop <strong>computer a keyboard and two monitors.</strong></td>
<td><img src="image19.png" alt="Predicted layout" /></td>
<td><img src="image20.png" alt="Reference layout" /></td>
<td><img src="image21.png" alt="Reference image" /></td>
<td>A woman is riding her <strong>bike</strong> down the street in front of traffic.</td>
<td><img src="image22.png" alt="Predicted layout" /></td>
<td><img src="image23.png" alt="Reference layout" /></td>
<td><img src="image24.png" alt="Reference image" /></td>
</tr>
<tr>
<td>A man and a woman stand <strong>under an umbrella</strong> at a street crossing on a rainy day.</td>
<td><img src="image25.png" alt="Predicted layout" /></td>
<td><img src="image26.png" alt="Reference layout" /></td>
<td><img src="image27.png" alt="Reference image" /></td>
<td>Two women walk outside, <strong>both holding up umbrellas.</strong></td>
<td><img src="image28.png" alt="Predicted layout" /></td>
<td><img src="image29.png" alt="Reference layout" /></td>
<td><img src="image30.png" alt="Reference image" /></td>
</tr>
</tbody>
</table>
## More examples

<table>
<thead>
<tr>
<th>Input Caption</th>
<th>Real Image</th>
<th>SG2IM</th>
<th>HDGAN</th>
<th>AttnGAN</th>
<th>Text2Scene [no inpainting]</th>
<th>Text2Scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>A room with a <strong>TV</strong> and some different types of <strong>couches</strong>.</td>
<td><img src="image1" alt="Real Image" /></td>
<td><img src="image2" alt="SG2IM" /></td>
<td><img src="image3" alt="HDGAN" /></td>
<td><img src="image4" alt="AttnGAN" /></td>
<td><img src="image5" alt="Text2Scene" /></td>
<td><img src="image6" alt="Text2Scene" /></td>
</tr>
<tr>
<td>A tall <strong>monitor</strong> is near a <strong>keyboard</strong> and a <strong>mouse</strong>.</td>
<td><img src="image7" alt="Real Image" /></td>
<td><img src="image8" alt="SG2IM" /></td>
<td><img src="image9" alt="HDGAN" /></td>
<td><img src="image10" alt="AttnGAN" /></td>
<td><img src="image11" alt="Text2Scene" /></td>
<td><img src="image12" alt="Text2Scene" /></td>
</tr>
<tr>
<td>a <strong>car bridge</strong> going <strong>over</strong> a commuter <strong>train</strong>.</td>
<td><img src="image13" alt="Real Image" /></td>
<td><img src="image14" alt="SG2IM" /></td>
<td><img src="image15" alt="HDGAN" /></td>
<td><img src="image16" alt="AttnGAN" /></td>
<td><img src="image17" alt="Text2Scene" /></td>
<td><img src="image18" alt="Text2Scene" /></td>
</tr>
<tr>
<td><strong>Three zebras</strong> grazing in a <strong>grassy</strong> area near shrubs.</td>
<td><img src="image19" alt="Real Image" /></td>
<td><img src="image20" alt="SG2IM" /></td>
<td><img src="image21" alt="HDGAN" /></td>
<td><img src="image22" alt="AttnGAN" /></td>
<td><img src="image23" alt="Text2Scene" /></td>
<td><img src="image24" alt="Text2Scene" /></td>
</tr>
<tr>
<td>A <strong>woman sitting</strong> on a <strong>bench</strong> with an <strong>umbrella on</strong> her head.</td>
<td><img src="image25" alt="Real Image" /></td>
<td><img src="image26" alt="SG2IM" /></td>
<td><img src="image27" alt="HDGAN" /></td>
<td><img src="image28" alt="AttnGAN" /></td>
<td><img src="image29" alt="Text2Scene" /></td>
<td><img src="image30" alt="Text2Scene" /></td>
</tr>
<tr>
<td>A <strong>woman is riding her bike</strong> down the street in front of <strong>traffic</strong>.</td>
<td><img src="image31" alt="Real Image" /></td>
<td><img src="image32" alt="SG2IM" /></td>
<td><img src="image33" alt="HDGAN" /></td>
<td><img src="image34" alt="AttnGAN" /></td>
<td><img src="image35" alt="Text2Scene" /></td>
<td><img src="image36" alt="Text2Scene" /></td>
</tr>
</tbody>
</table>
More examples
Questions?