Evaluating multi-modal deep learning systems with micro-worlds

Alexander Kuhnle
Supervisor: Ann Copestake

Computer Laboratory
University of Cambridge

Graduate Open Day, 2016
Image captioning

- “A large herd of horses riding on either side of two men.”
- “A man with a horde of horses, he appears to be herding them.”
- “There is a herd of horses running and there is two people in the center of the herd on horses directing them.”
- “Men are riding horses among other horses.”
- “There are many horses standing in the field.”
- “A man in yellow, is riding his horse on the beach.”
- “A person riding horseback on the beach with a pack of dogs running along.”
- “A person wearing a yellow shirt is riding a horse with some dogs on a beach.”
- “A man riding on the back of a brown horse on a beach.”
- “A man rides a horse along the beach with a pack of dogs.”

Photos and captions from the Microsoft COCO dataset (http://mscoco.org/).
Abstract images

- “There is a red square.”
- “Some shapes are green.”
- “All circles are to the left of a red square.”
- “The left-most shape is a blue circle.”
- “Most circles are blue.”
Abstract images

- “There is a red square.”
- “Some shapes are green.”
- “All circles are to the left of a red square.”
- “The left-most shape is a blue circle.”
- “Most circles are blue.” (???)

⇒ Clear, noise-free representation
⇒ Much less different object types, etc
⇒ Still, structurally complex situations
Abstract images

- “There is a red square.”
- “Some shapes are green.”
- “All circles are to the left of a red square.”
- “The left-most shape is a blue circle.”
- “Most circles are blue.” (???)

Experimental setup / network architecture:

1. CNN yields image embedding
2. LSTM yields caption embedding
3. Fuse both to decide appropriateness of caption given image

(More sophisticated architectures later.)
Data generation

**Internal representation:**

\[
\left[ \{ \text{shape} = "\text{triangle}" , \text{color} = "\text{green}" , \text{pos, etc}\}, \\
\{ \text{shape} = "\text{circle}" , \text{color} = "\text{blue}" , \text{pos, etc}\}, \\
\{ \text{shape} = "\text{cross}" , \text{color} = "\text{red}" , \text{pos, etc}\}, \\
\{ \text{shape} = "\text{square}" , \text{color} = "\text{red}" , \text{pos, etc}\}, \\
\{ \text{shape} = "\text{circle}" , \text{color} = "\text{green}" , \text{pos, etc}\} \right]
\]

⇒ Randomly sampled

⇒ Both image (simple) and caption (more complex) can be generated from it
Data generation

Internal representation:
\[
[ \{ \text{shape}="\text{triangle}", \text{color}="\text{green}\}, \text{pos}, \text{etc}\},
\{ \text{shape}="\text{circle}", \text{color}="\text{blue}\}, \text{pos}, \text{etc}\},
\{ \text{shape}="\text{cross}", \text{color}="\text{red}\}, \text{pos}, \text{etc}\},
\{ \text{shape}="\text{square}", \text{color}="\text{red}\}, \text{pos}, \text{etc}\},
\{ \text{shape}="\text{circle}", \text{color}="\text{green}\}, \text{pos}, \text{etc}\} \]
\]

⇒ Caption generation via semantic graph representation:

“\text{There is a blue circle.}”
Formal semantics

“There is a red square.” \( \exists s \in W : \text{square}(s.\text{shape}) \land \text{red}(s.\text{colour}) \)

“Some shapes are green.” \( \exists s_1[\neq s_2] \in W : \text{green}(s_1.\text{colour}) \land \text{green}(s_2.\text{colour}) \)

“All circles are to the left of a red square.”
\[ \forall s_1 \in W : \text{circle}(s_1.\text{shape}) \Rightarrow (\exists s_2 \in W : \text{square}(s_2.\text{shape}) \land \text{red}(s_2.\text{clr}) \land s_1.x < s_2.x) \]

“The left-most shape is a blue circle.”
\[ \forall s_1 \in W : (\forall s_2 \in W : s_1.x \leq s_2.x) \Rightarrow \text{circle}(s_1.\text{shape}) \land \text{blue}(s_1.\text{colour}) \]

“Most circles are blue.”
\[ S_1 = \{ s \in W : \text{circle}(s.\text{shape}) \land \text{blue}(s.\text{color}) \} \]
\[ S_2 = \{ s \in W : \text{circle}(s.\text{shape}) \} \]
\[ |S_1| / |S_2| \geq 0.5 \]
Thank you for your attention!

Questions?