Evaluating multi-modal deep learning systems with micro-worlds

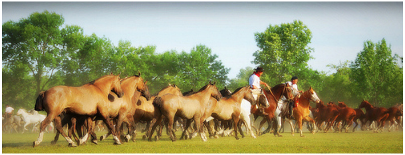
Alexander Kuhnle

Supervisor: Ann Copestake

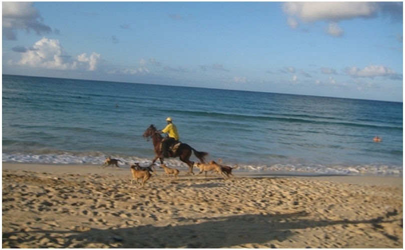
Computer Laboratory
University of Cambridge

Graduate Open Day, 2016

Image captioning



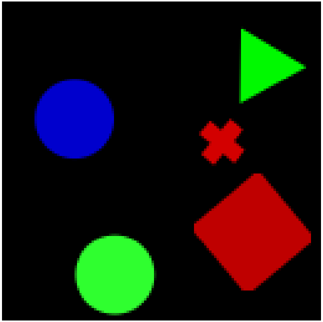
- *“A large herd of horses riding on either side of two men.”*
- *“A man with a horde of horses, he appears to be herding them.”*
- *“There is a herd of horses running and there is two people in the center of the herd on horses directing them.”*
- *“Men are riding horses among other horses.”*
- *“There 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.”*

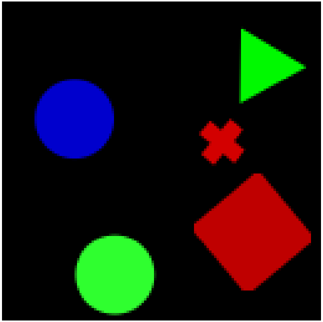
Photos and captions from the Microsoft COCO dataset (`http://mscoco.org/`).

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



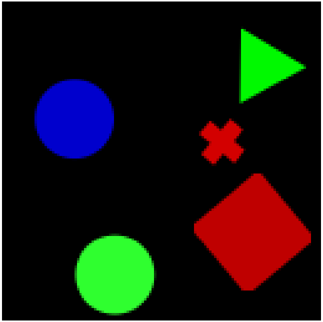
- *"There is a red square."*
- *"Some shapes are green."*
- *"All circles are to the left of a red square."*
- *"The left-most shape is a blue circle."*
- *"Most circles are blue."* (???)

⇒ Clear, noise-free representation

⇒ Much less different object types, etc

⇒ Still, structurally complex situations

Abstract images



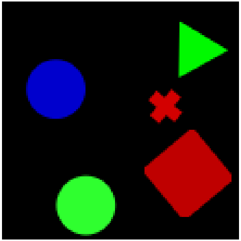
- *"There is a red square."*
- *"Some shapes are green."*
- *"All circles are to the left of a red square."*
- *"The left-most shape is a blue circle."*
- *"Most circles are blue."* (???)

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



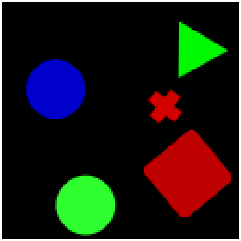
Internal representation:

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

⇒ Randomly sampled

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

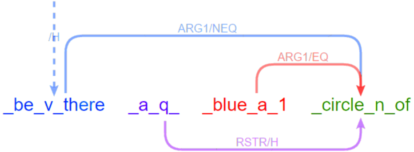
Data generation



Internal representation:

[{shape=”triangle”, color=”green”, pos, etc},
{shape=”circle”, color=”blue”, pos, etc},
{shape=”cross”, color=”red”, pos, etc},
{shape=”square”, color=”red”, pos, etc},
{shape=”circle”, color=”green”, pos, etc}]

⇒ Caption generation via semantic graph representation:



“There is a blue circle.”

Formal semantics

"There is a red square." $\exists s \in W\colon \mathrm{square}(s.\mathrm{shape}) \wedge \mathrm{red}(s.\mathrm{colour})$

"Some shapes are green." $\exists s_1 \left[\neq s_2\right] \in W\colon \mathrm{green}(s_1.\mathrm{colour})\ \left[\wedge\ \mathrm{green}(s_2.\mathrm{colour})\right]$

"All circles are to the left of a red square."

$$\forall s_1 \in W\colon \mathrm{circle}(s_1.\mathrm{shp}) \Rightarrow \Big(\exists s_2 \in W\colon \mathrm{square}(s_2.\mathrm{shp}) \wedge \mathrm{red}(s_2.\mathrm{clr}) \wedge s_1.\mathrm{x} < s_2.\mathrm{x}\Big)$$

"The left-most shape is a blue circle."

$$\forall s_1 \in W\colon \Big(\forall s_2 \in W\colon s_1.\mathrm{x} \leq s_2.\mathrm{x}\Big) \Rightarrow \mathrm{circle}(s_1.\mathrm{shape}) \wedge \mathrm{blue}(s_1.\mathrm{colour})$$

"Most circles are blue."

$$\begin{aligned} &S_1 = \{s \in W\colon \mathrm{circle}(s.\mathrm{shape}) \wedge \mathrm{blue}(s.\mathrm{color})\} \\ &S_2 = \{s \in W\colon \mathrm{circle}(s.\mathrm{shape})\} \\ &|S_1| \;/\; |S_2| >[=]\ 0.5 \end{aligned}$$

Thank you for your attention!

Questions?