

Captioning in ShapeWorld

The captioning task and datasets

Image captioning

- Automatically generate descriptions for images
- Multi-modal task: NLP+CV

Real-world datasets

- MSCOCO (Lin et al., 2014), Flickr30k (Young et al., 2014), Visual Genome (Krishna et al., 2017)



Caption: A woman is skiing down a snow covered hill.

Figure 1. An example image from MSCOCO.

The captioning task and datasets

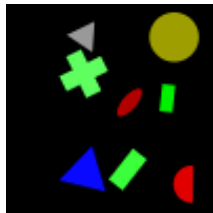
Image captioning

- Automatically generate descriptions for images
- Multi-modal task: NLP+CV

Real-world datasets

Synthetic datasets

- ShapeWorld (Kuhnle and Copestake, 2017)



- There is a gray triangle.
- Exactly one cross is green.
- A red ellipse is to the right of a green rectangle.

Figure 2. ShapeWorld example: descriptive statements in the context of multiple shapes (truthful descriptions in green, and wrong descriptions in red).

Neural image captioning models

An encoder-decoder architecture

- Encoder: Convolutional neural networks (CNNs) as the visual pipeline
- Decoder: Recurrent neural networks (RNNs) as the linguistic pipeline

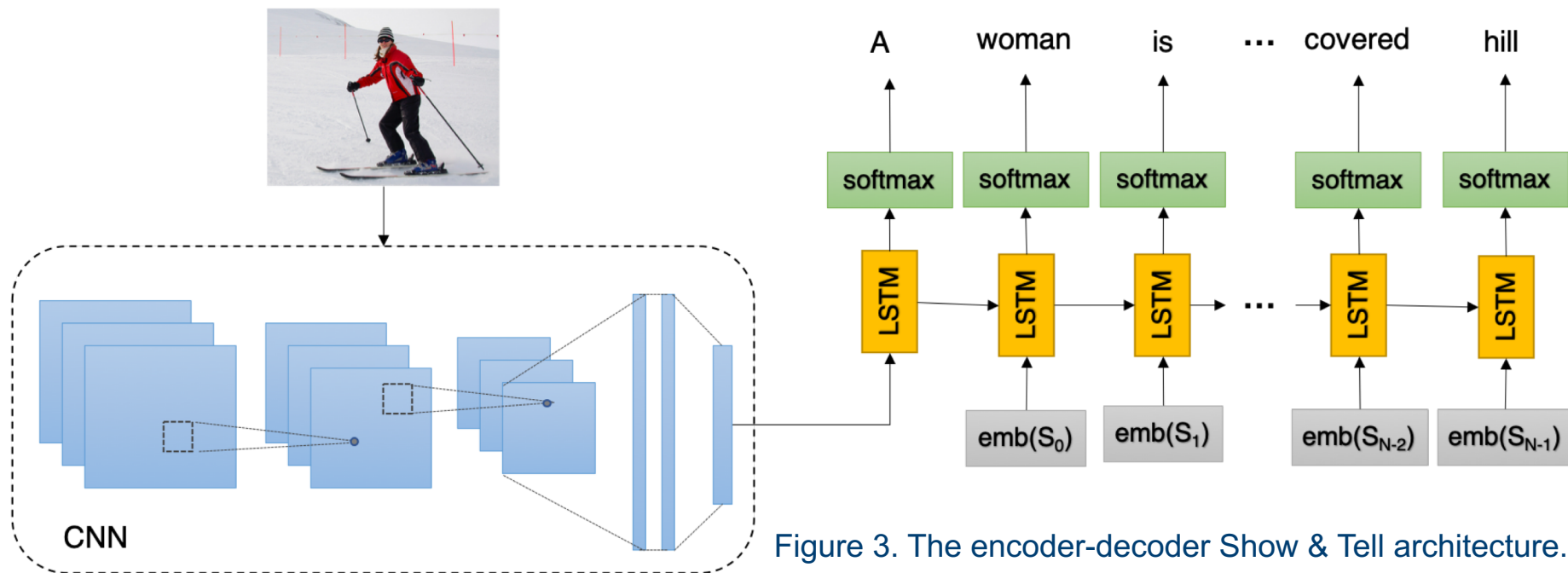


Figure 3. The encoder-decoder Show & Tell architecture.

The GTD evaluation framework

Existing evaluation metrics

- BLEU, METEOR, CIDEr, ROUGE, SPICE
- Use a set of reference captions as an approximate of the image content
- Captions are compared to a set of human judgements about the image



Caption 1: A man is skate boarding down a path and a dog is running by his side.

Caption 2: A man walking his dog on a quiet country road.

Figure 4. An example image and captions from MSCOCO.

The GTD evaluation framework

GTD evaluation framework: grammaticality, truthfulness and diversity

- Grammaticality
 - Parseability with the English Resource Grammar (ERG)
- Truthfulness
 - Whether a caption is compatible with the image content
 - Compare a caption parse with the underlying representation of the world model of an image
- Diversity

$$\text{diversity} = \frac{\#\{\text{model-generated}\}}{\#\{\text{ShapeWorld-generated}\}}$$

Experimental setup

Two image captioning models

- Show & Tell (Vinyals et al., 2015)
- LRCN (Donahue et al., 2015)

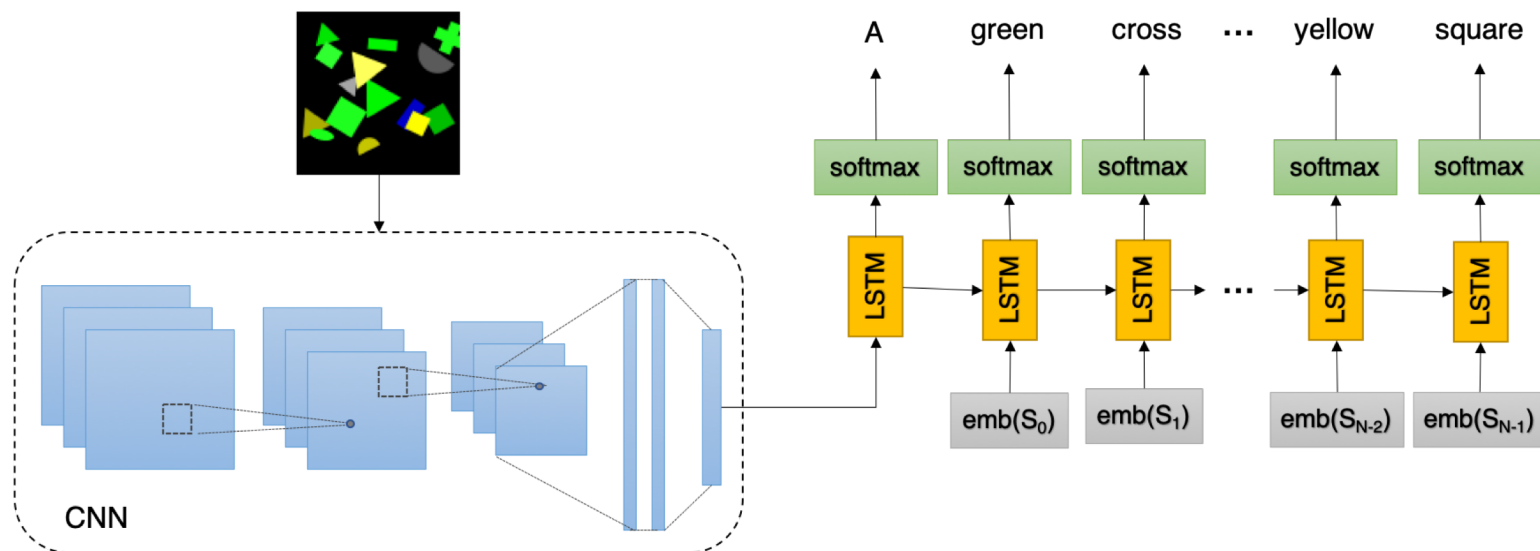


Figure 3. The Show & Tell architecture.

Experimental setup

Two image captioning models

- Show & Tell (Vinyals et al., 2015)
- LRCN (Donahue et al., 2015)

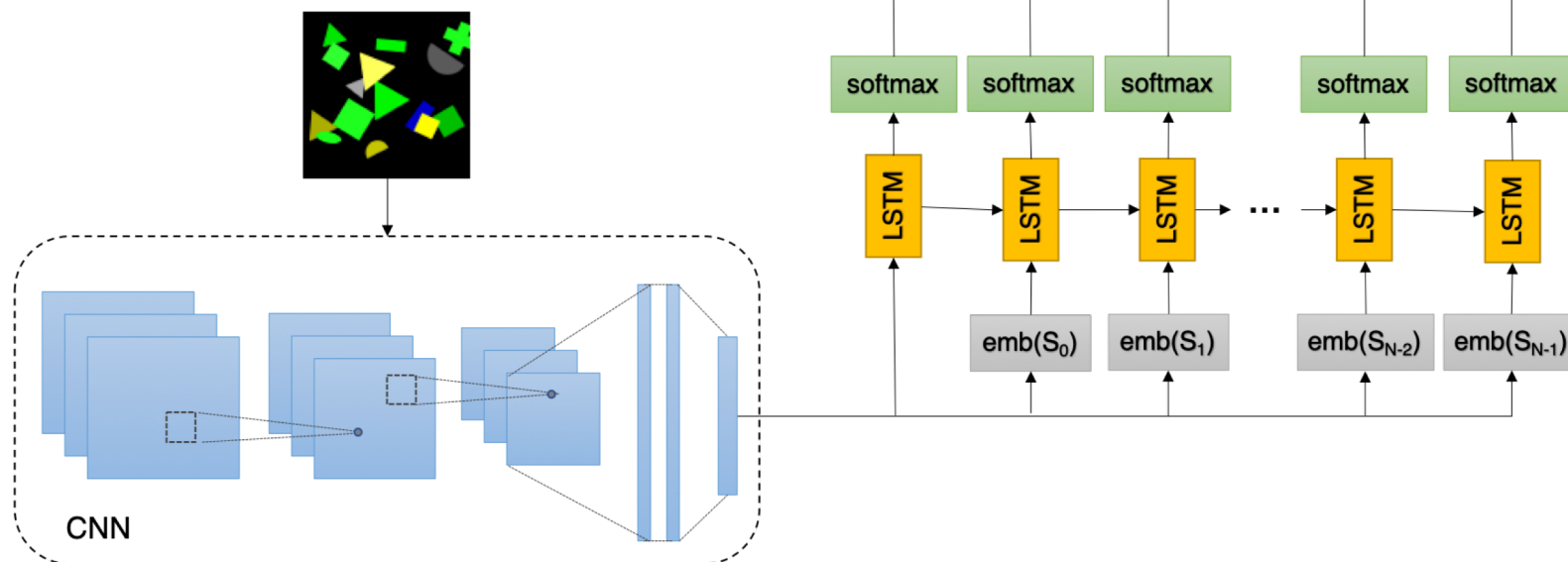




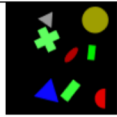



Figure 5. The LRCN architecture.

Experimental setup

The ShapeWorldICE datasets

Type	Variant	Caption	Image
Existential	OneShape	There is a green cross. A rectangle is green. There is a cyan shape.	
	MultiShapes	A shape is a gray triangle. There is a square. There is a yellow shape.	
Spatial	TwoShapes	A square is above a red pentagon. A yellow square is above a yellow pentagon. A square is to the left of a pentagon.	
	MultiShapes	A blue triangle is to the left of a semicircle. A circle is above a green rectangle. A semicircle is to the left of a circle.	
Quantification	Count	Exactly two rectangles are green. Exactly one shape is a yellow circle. Exactly zero shapes are ellipses.	
	Ratio	A quarter of the shapes are rectangles. A third of the rectangles are magenta. Half the shapes are green.	

Experimental results

- LRCN shows clearly superior performance in terms of truthfulness
- Incorporating visual features at each time step is beneficial

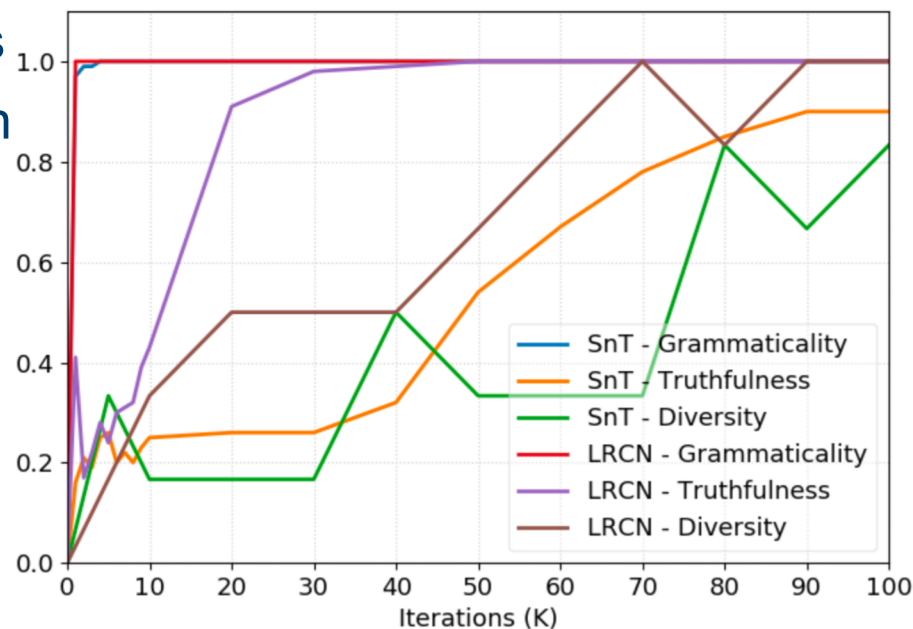


Figure 6. Performance comparison of the Show&Tell model and the LRCN model on *Existential-MultiShapes*.

Experimental results

- Perfect grammaticality achieved for all caption types

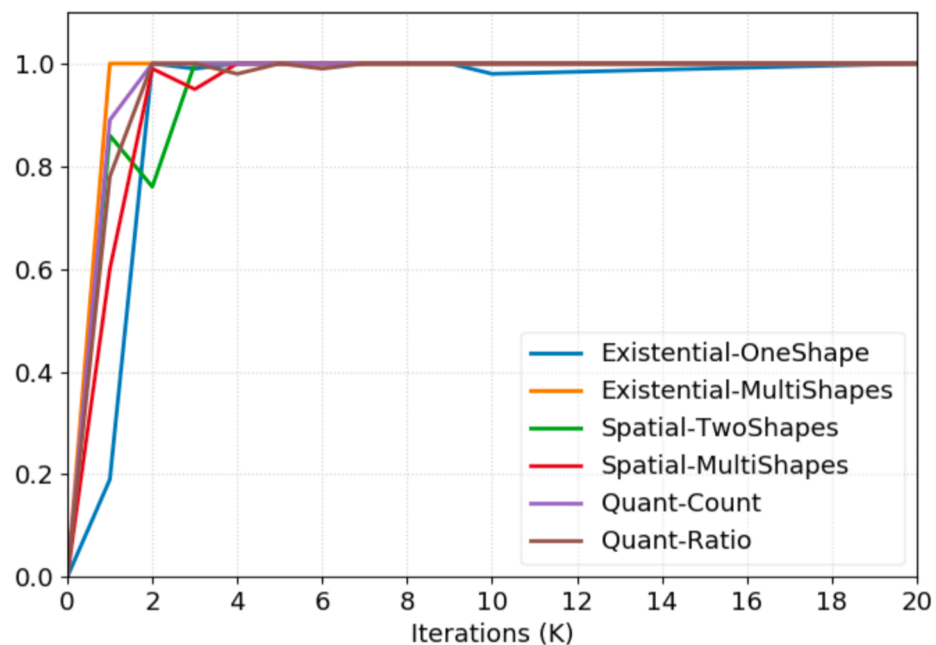
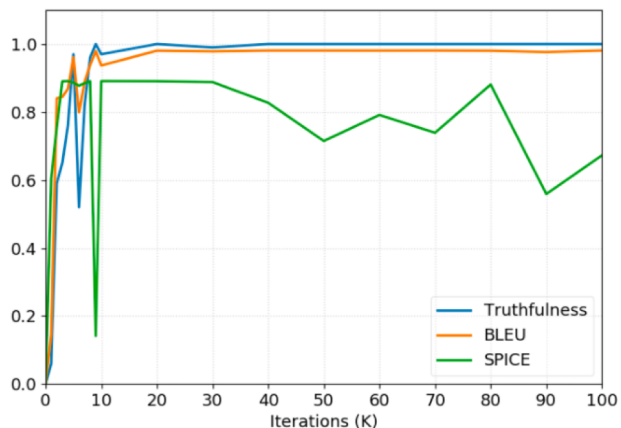


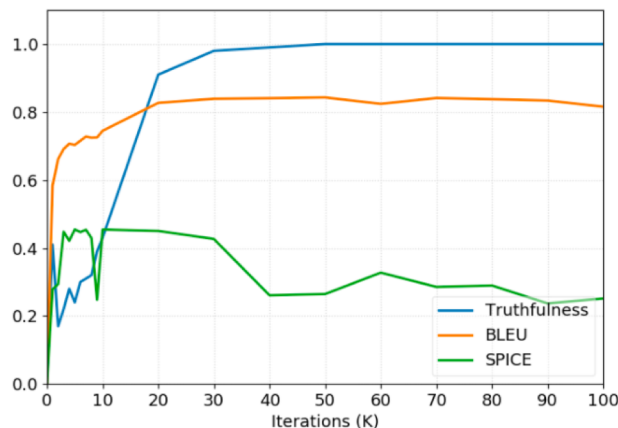
Figure 7. Ratio of grammatical sentences produced by LRCN for different ShapeWorld datasets in the first training iterations (stays at 100% afterwards).

Experimental results

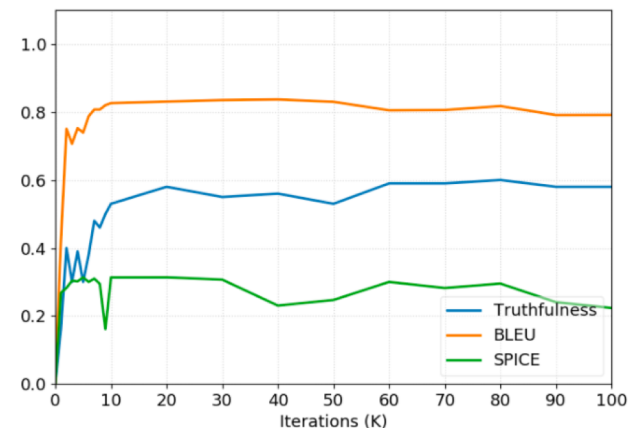
- Low or no correlation between the BLEU / SPICE scores and caption truthfulness



(a) Existential-OneShape



(b) Existential-MultiShapes



(c) Spatial-MultiShapes

Figure 8. Learning curves for LRCN on *Existential-OneShape*, *Existential-MultiShapes* and the *Spatial-MultiShapes*.

Experimental results

- Failure to learn complex spatial relations
- The counting tasks are non-trivial

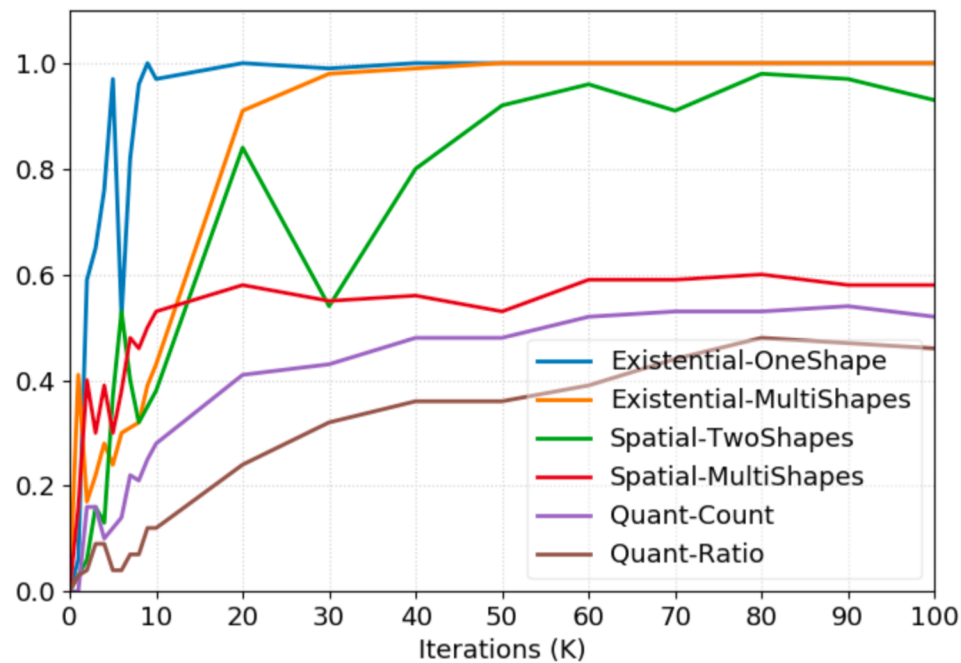


Figure 9. Truthfulness ratios of sentences produced by LRCN for different ShapeWorld datasets.

Experimental results

- Caption diversity benefits from varied language constructions in complex datasets (the *Spatial* and *Quantification* datasets)

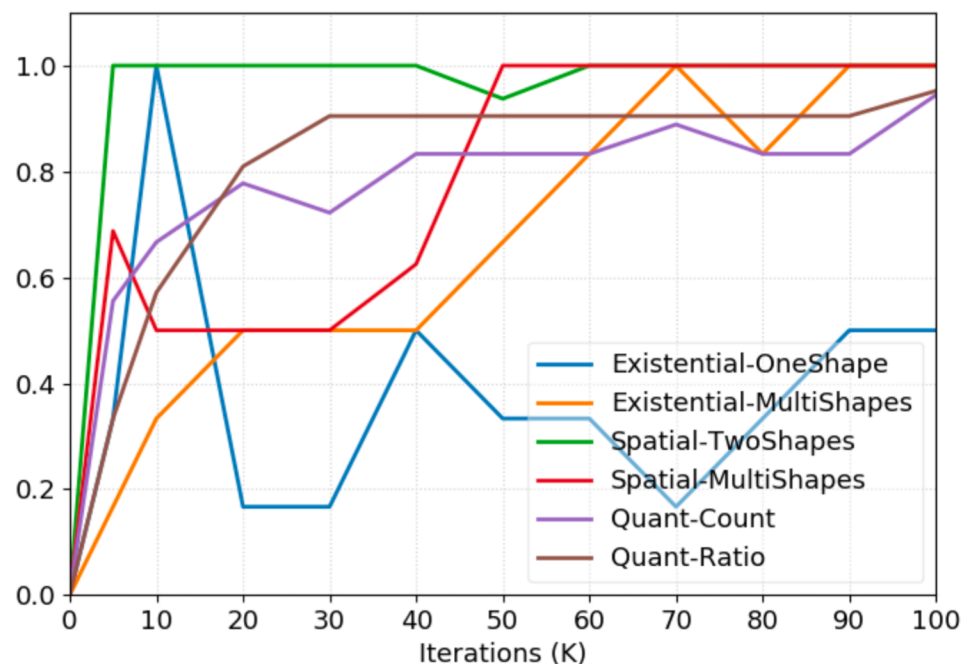


Figure 10. Diversity ratios of sentences produced by LRCN on different ShapeWorld datasets.

Conclusions

- Synthetic datasets enables detailed, diagnostic evaluation of multimodal deep learning systems
- The GTD framework can serve as a supplementary evaluation method to existing standard evaluations
- Future work
 - More complex linguistic variants (relative clauses, coreference, etc)
 - Incorporate the GTD signal to the training process (via a GTD-aware loss)

Q & A