"Unit-testing" deep learning with synthetic data for more informative evaluation

Alexander Kuhnle

Supervisor: Prof Ann Copestake
Department of Computer Science and Technology
University of Cambridge

Language Sciences Symposium

19th June 2018

Overview

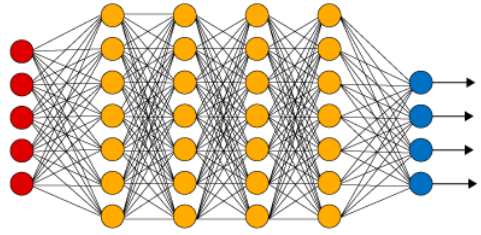
- Visual question answering
- Problems with the VQA Dataset
- Potential of synthetic data
- Evaluation and generalization

Deep learning in Natural Language Processing

labeled dataset

(>100k data points)

~90% training split



deep neural network

(recurrent sequence model)

~10% test split

evaluation

Visual question answering

Where is this cat laying?

Is the cat awake?

What color is the cat?

Is the cat facing the computer?

Is the cat typing?

Is the cat playing with the mouse?

What object is shining on the animal?

What objects is the cat sitting behind?

How many cats?

How many items are on the bookcase?

Are these two children related?

Is the dog begging for food?

$\Rightarrow$ **Visual Turing test?**

Examples from VQA Dataset (`http://visualqa.org/browser/`)

Other popular datasets

SNLI – Stanford Natural Language Inference Corpus

C: A soccer game with multiple males playing.
H: Some men are playing a sport.
→ entailment

C: A smiling costumed woman is holding an umbrella.
H: A happy woman in a fairy costume holds an umbrella.
→ neutral

C: A man inspects the uniform of a figure in some East Asian country.
H: The man is sleeping
→ contradiction

SQuAD – Stanford Question Answering Dataset

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

(1) What causes precipitation to fall?
⇒ gravity

(2) What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
⇒ graupel

(3) Where do water droplets collide with ice crystals to form precipitation? ⇒ within a cloud

Examples from Bowman et al. (`https://arxiv.org/abs/1508.05326`) and Rajpurkar et al. (`https://arxiv.org/abs/1606.05250`)

Question-answer biases

- What sport is...? ⇒ tennis (41%)

- How many...? ⇒ two (39%)

- Do you see a...? ⇒ yes (87%)

Examples from Goyal et al. (`https://arxiv.org/abs/1612.00837`)

Complete question/image understanding

- What...? ⇒ umbrella
- What season...? ⇒ summer
- What season of...? ⇒ summer
- ...
- What season of year was this photo taken in? ⇒ summer

- What does the red sign say? ⇒ stop

Examples from Agrawal et al. (https://arxiv.org/abs/1606.07356) and Devi Parikh's slides (https://newgeneralization.github.io/)

Sensitivity to question words

- How symmetrical are the white bricks on either side of the building? $\Rightarrow$ very
- How **spherical** are the white bricks on either side of the building? $\Rightarrow$ very
- How **soon** are the bricks **fading** on either side of the building? $\Rightarrow$ very
- How **fast** are the bricks **speaking** on either side of the building? $\Rightarrow$ very

Example from Mudrakarta et al. `https://arxiv.org/abs/1805.05492`).

Crowd-sourced real-world datasets

Solve the problem/dataset?

Deep learning will find a way to make effective use of the data.

Evaluate model capabilities?

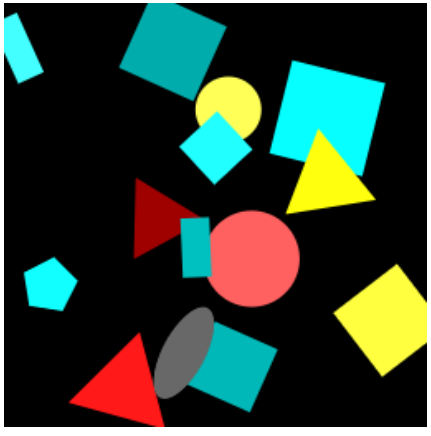
Are these datasets appropriate to investigate this question?

- Natural?
- Difficult?
- Specific?

$\Rightarrow$ **Synthetic data!**

ShapeWorld examples: relations and quantifiers

- A magenta square is to the right of a green shape.
- A yellow shape is not in front of a square.
- A circle is farther from an ellipse than a gray cross.
- A cross is not the same color as a green rectangle.
- The lowermost green shape is a cross.
- A red shape is the same shape as a green shape.



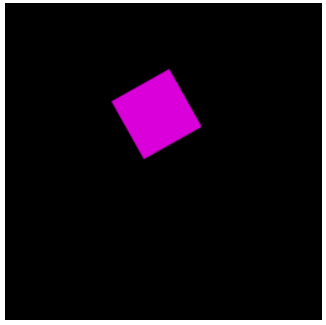
- Less than one triangle is cyan.
- At least half the triangles are red.
- More than a third of the shapes are cyan squares.
- Exactly all the five squares are red.
- More than one of the seven cyan shapes is a square.
- Twice as many red shapes as yellow shapes are circles.

Properties and comparison

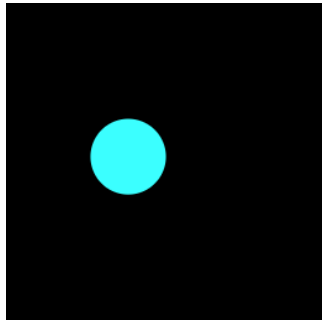
real-world data	**vs**	**synthetic data**
uncontrolled content	$\longleftrightarrow$	clean content
sparse instance coverage	$\longleftrightarrow$	targeted instance coverage
monolithic benchmark	$\longleftrightarrow$	tailored unit tests
test interpolation ability	$\longleftrightarrow$	test extrapolation ability

$\Rightarrow$ **Complementary evaluation paradigms**

What type of generalization do we expect/desire?

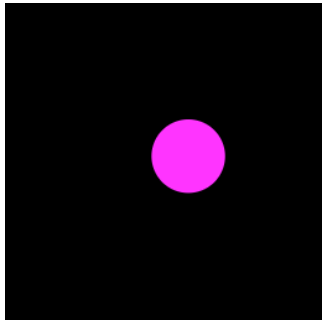


magenta square



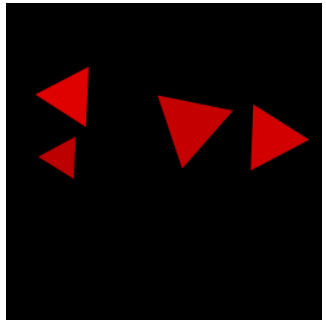
cyan circle

$\Longrightarrow$



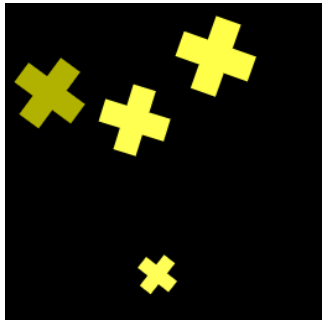
magenta circle

three crosses



four triangles

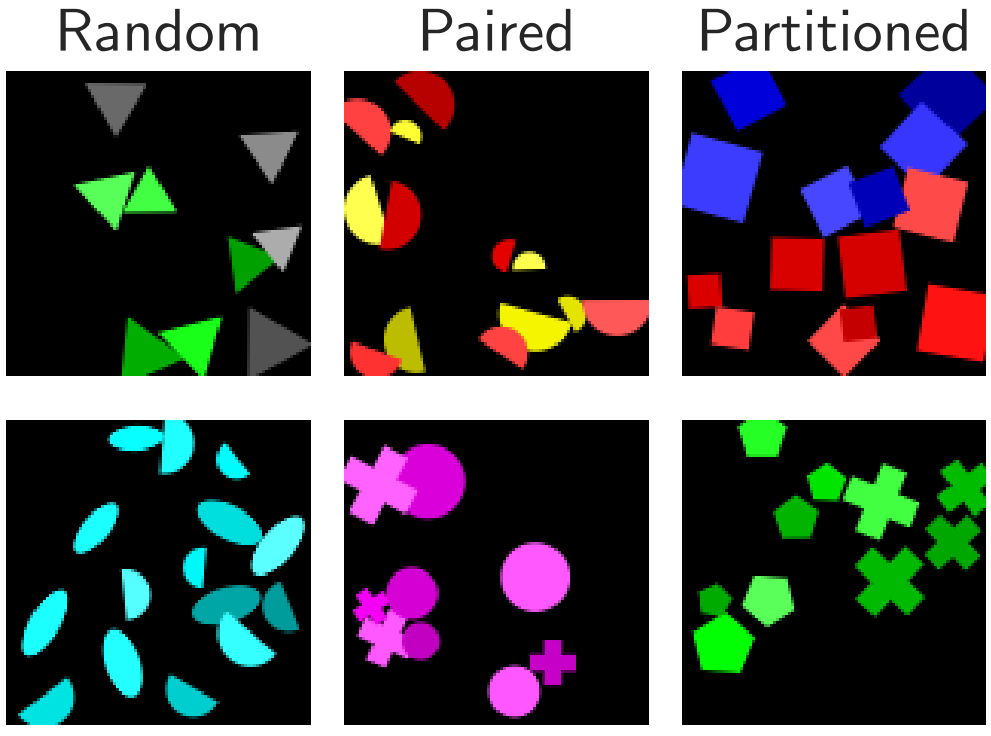
$\Longrightarrow$



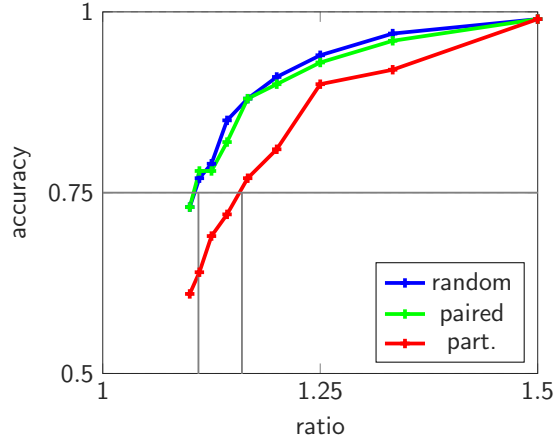
four crosses

Example use case: replication of psychology experiment

(inspired by *The meaning of "most"*, Pietroski et al., 2009)



+ "More/less than half the shapes are X?"




Thank you for your attention!
Questions?