Deep Learning

Sheng Zha

Amazon AI

What can deep learning do?

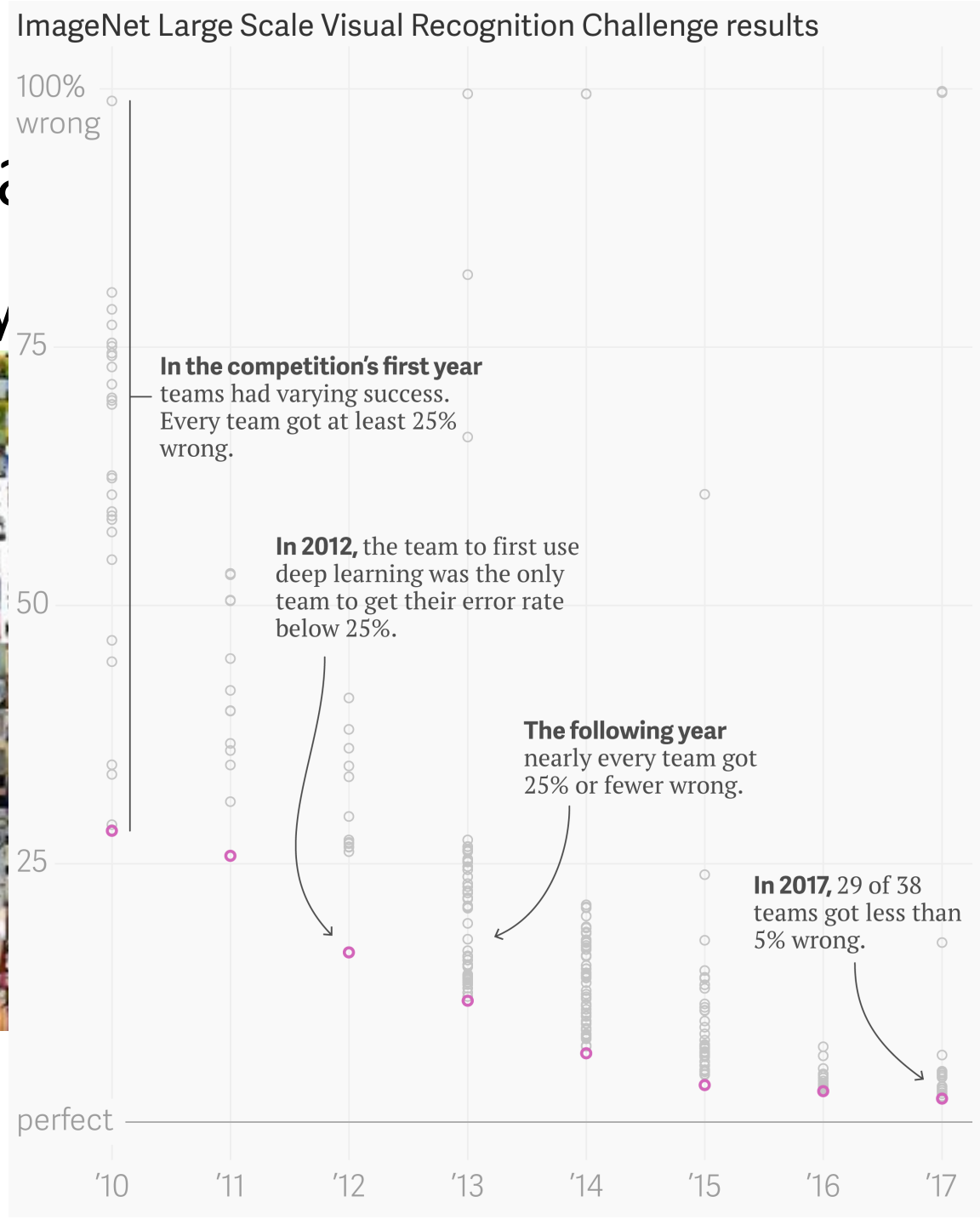
What can deep learning do?

- Classify images

http://www.image-net.org/

Wha

- Classify



Yanofsky, Quartz
https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

rg/

earning

What can deep learning do?

- Detect and segment object

https://github.com/matterport/Mask_RCNN

What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

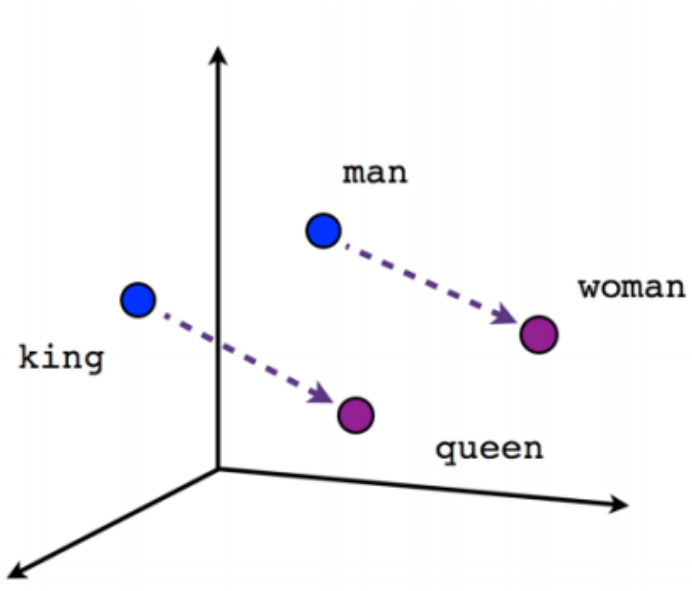
What can deep learning do?

https://github.com/zhanghang1989/MXNet-Gluon-Style-Transfer/

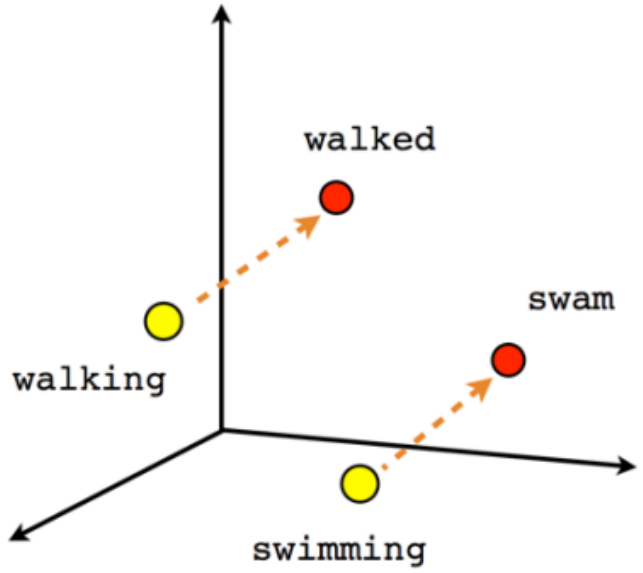
What can deep learning do?

Karras et al, ICLR 2018

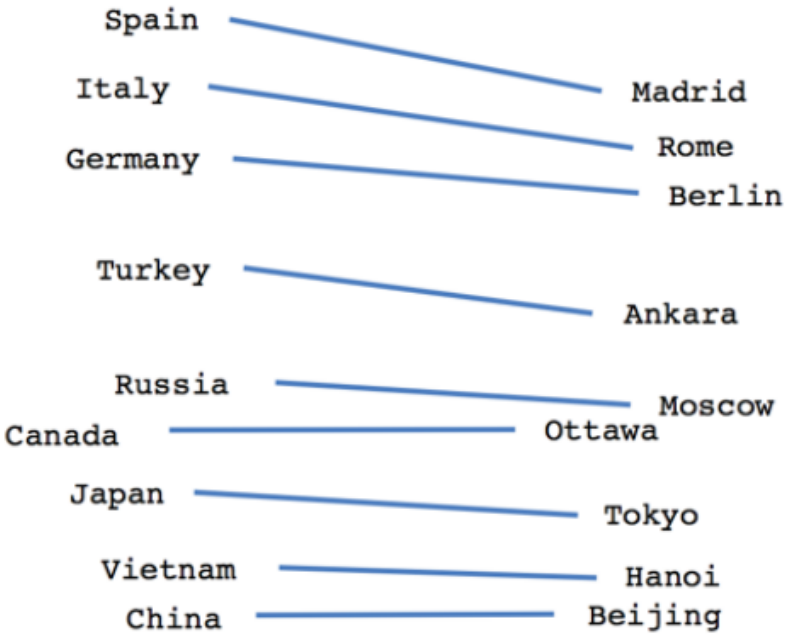
What can deep learning do?



Male-Female



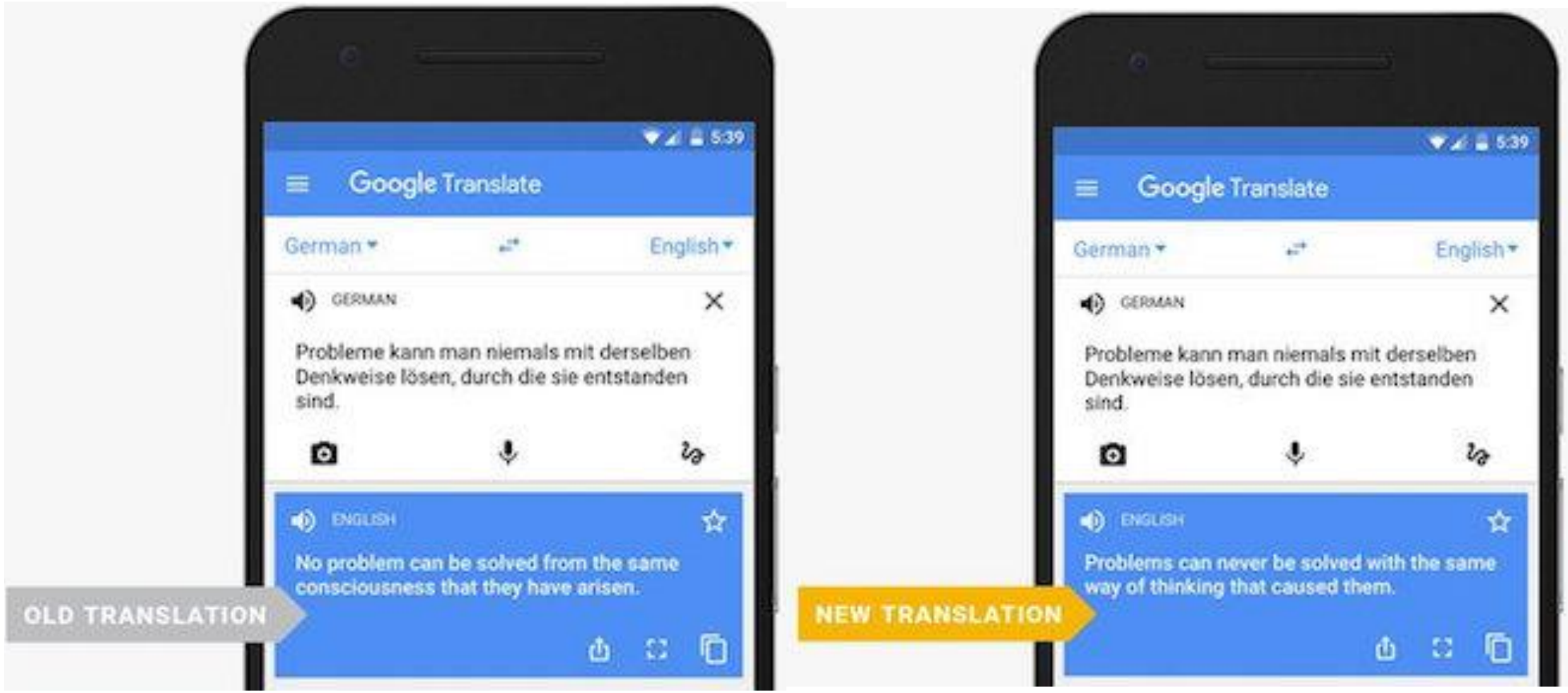
Verb tense



Country-Capital

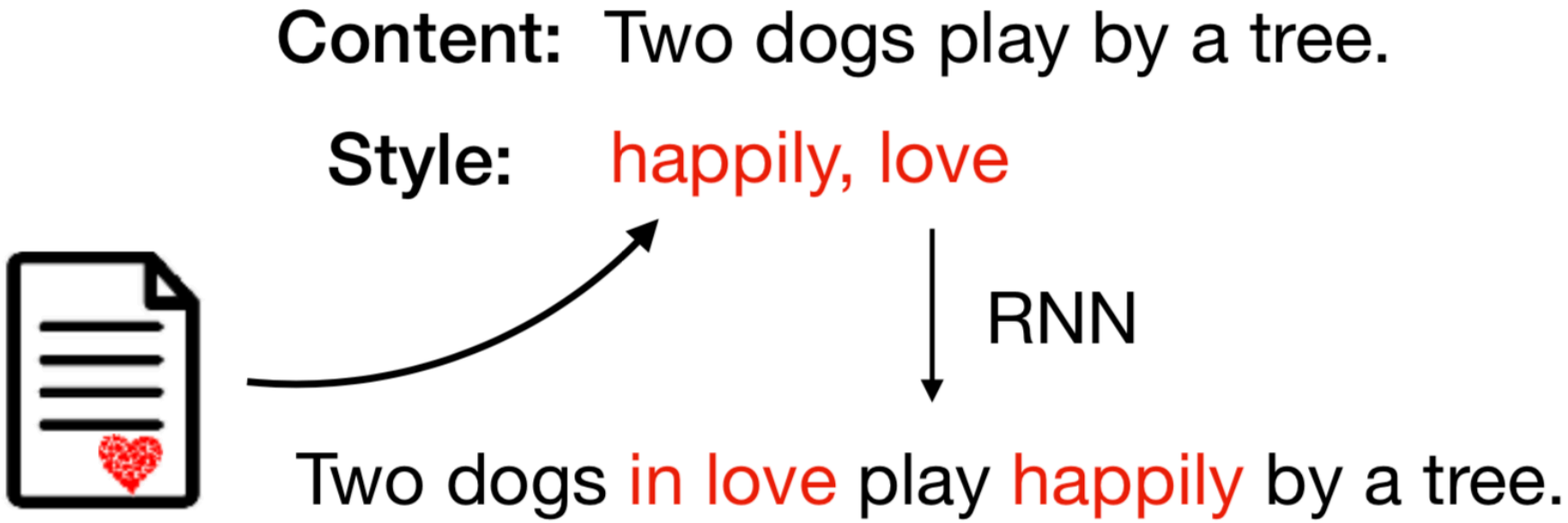
https://www.tensorflow.org/tutorials/word2vec

What can deep learning do?



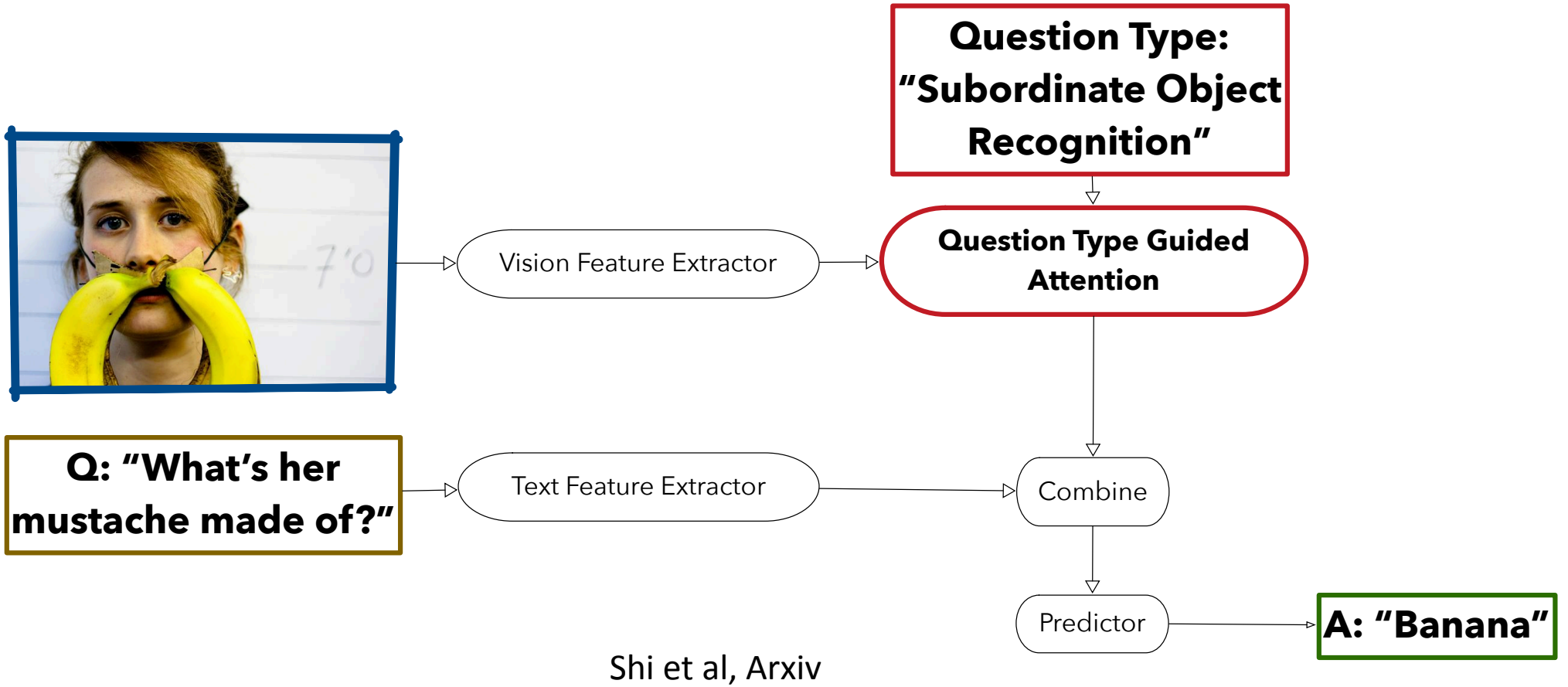
https://www.pcmag.com/news/349610/google-expands-neural-networks-for-language-translation

What can deep learning do?



Li et al, NACCL 2018

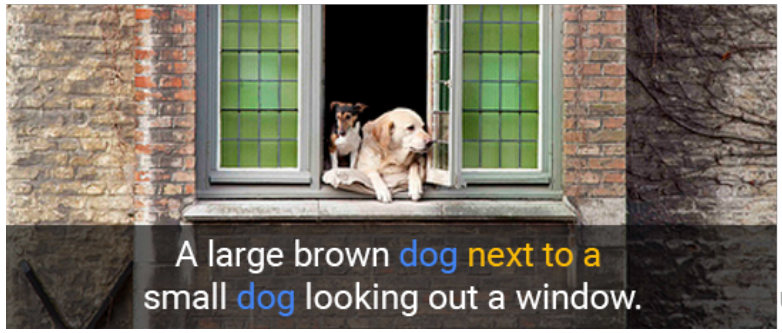
What can deep learning do?



Shi et al, Arxiv

What can deep learning do?

Human captions from the training set

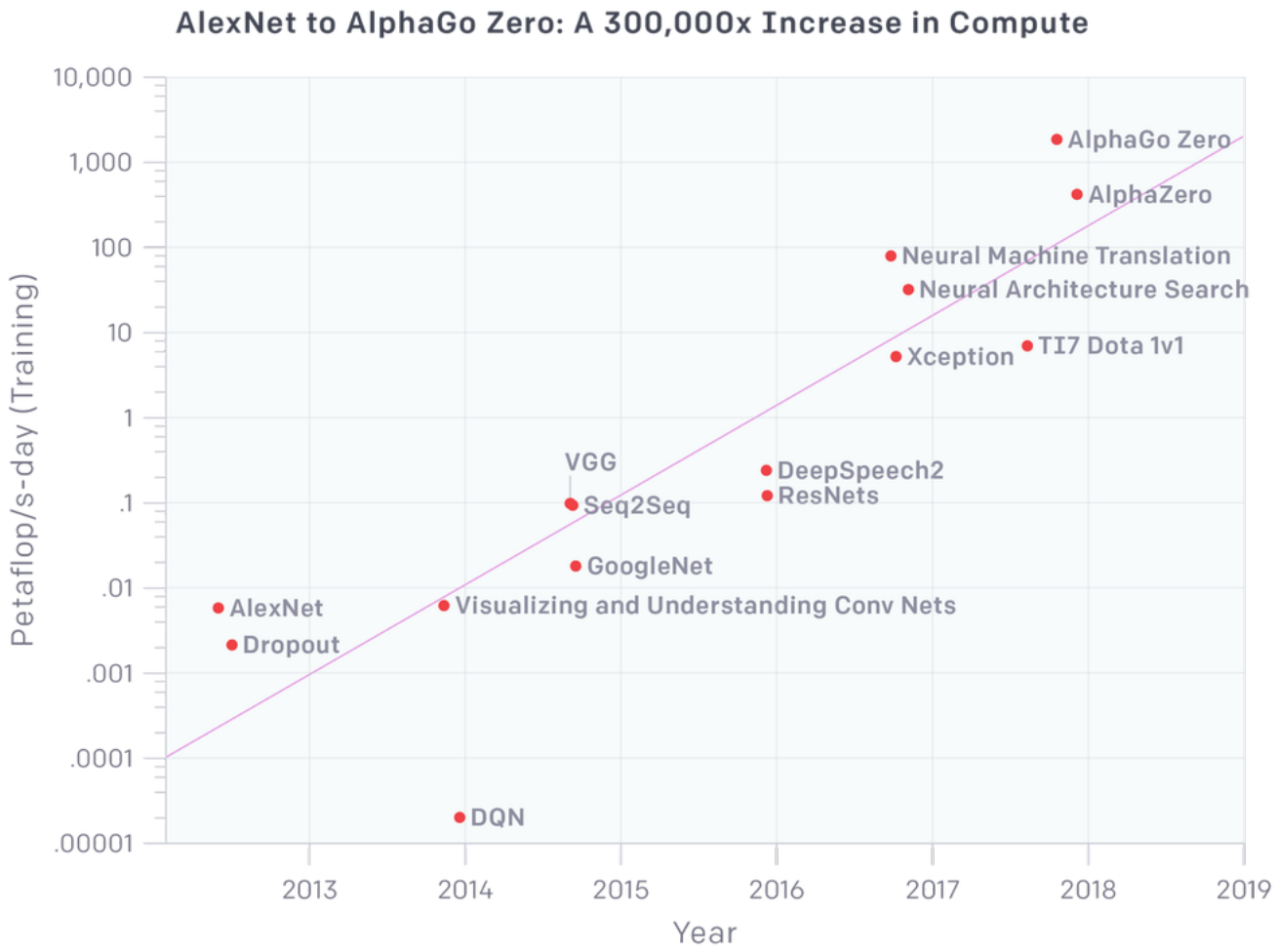


Shallue et al, Google Research Blog
https://ai.googleblog.com/2016/09/show-and-tell-image-captioning-open.html

Automatically captioned

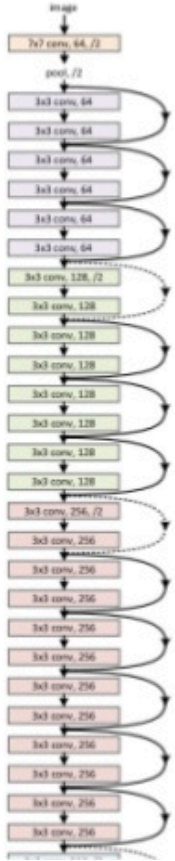
How did deep learning happen?

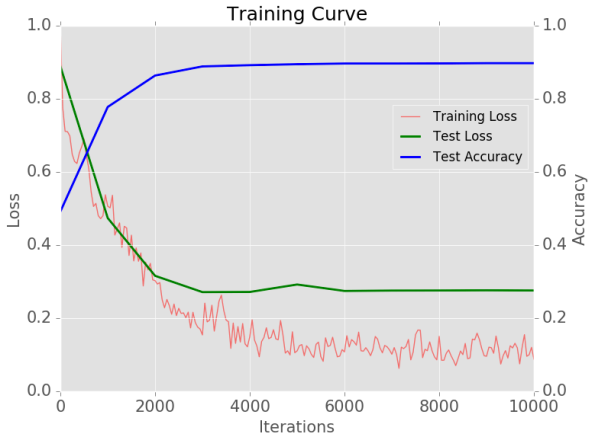
How did deep learning happen?



Amodei et al, OpenAI blog
https://blog.openai.com/ai-and-compute/

How did deep learning happen?





How did deep learning happen?

	Titan	Summit
Compute Nodes	18,688	4,608
GPU	1x NVIDIA K20x per node; 18,688 GPUs total	6x NVIDIA Volta per node; 27,648 GPUs total
CPU	1x AMD Opteron 16-core per node; 18,688 CPUs total	2x IBM Power9 per node; 9,216 CPUs total
Node Performance	1.44 teraflops	49 teraflops
Peak Performance	27 petaflops	200 petaflops
System Memory	710 terabytes	10 petabytes
CPU-GPU-Memory Interconnect	PCI Gen 2 (8GB/s)	NVLink (300GB/s)
Power Consumption	9MW	13MW

Foertter, NVIDIA Developer Blog
https://devblogs.nvidia.com/summit-gpu-supercomputer-enables-smarter-science/

Programming for deep learning

Programming for deep learning used to be…

```
void computeLogregSoftmaxGrad(NVMatrix& labels, NVMatrix& probs, NVMatrix& target, bool add, float coeff) {
    int numCases = probs.getLeadingDim();
    int numOut = probs.getFollowingDim();
    assert(labels.getNumElements() == numCases);
    assert(probs.isContiguous());
    assert(target.isContiguous());
    assert(labels.isContiguous());
    assert(probs.isTrans());

    dim3 threads(LOGREG_GRAD_THREADS_X, LOGREG_GRAD_THREADS_Y);
    dim3 blocks(DIVUP(numCases, LOGREG_GRAD_THREADS_X), DIVUP(numOut, LOGREG_GRAD_THREADS_Y));
    if (!add) {
        target.resize(probs);
        kLogregSoftmaxGrad<false><<<blocks, threads>>>(probs.getDevData(), labels.getDevData(), target.getDevData(),
                                                       numCases, numOut, coeff);
    } else {
        kLogregSoftmaxGrad<true><<<blocks, threads>>>(probs.getDevData(), labels.getDevData(), target.getDevData(),
                                                      numCases, numOut, coeff);
    }

    getLastCudaError("computeLogregSoftmaxGrad: Kernel execution failed");
}
```

Krizhevsky et al, code from cuda-convnet (AlexNet)

Programming for deep learning nowadays

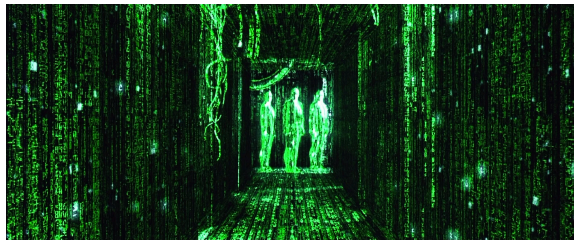
```
loss_func = gluon.loss.SoftmaxCrossEntropyLoss()
loss = loss_func(output)
loss.backward()
```

Programming for deep learning nowadays

```
loss_fu                    ro
loss = l
loss.ba
```

Programming for deep learning nowadays


```
output = model(data)
```



[crazy GPU stuff]

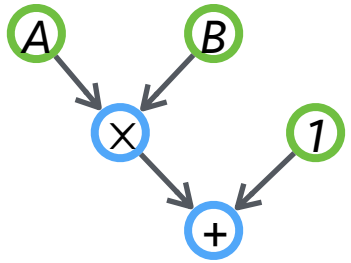
DL Programming — Imperative


```
import numpy as np
a = np.ones(10)
b = np.ones(10) * 2
c = b * a
print c
d = c + 1
```

Easy to debug, easy to code

- **Advantages**
  - Straightforward and flexible
  - Use language natively (loops, control flow, debugging)
- **Disadvantages**
  - Hard to optimize

DL Programming — Declarative (Symbolic)




```
A = Variable('A')
B = Variable('B')
C = B * A
D = C + 1
f = compile(D)
d = f(A=np.ones(10),
      B=np.ones(10)*2)
```

C can share memory with *D*

- **Advantages**:
  - Opportunities for optimization
  - Easy to serialize models
  - More portable across languages
- **Disadvantages**:
  - Hard to debug
  - Unsuitable for dynamic graphs
  - Can't use native code

MXNet Gluon

Q: Why GLUON?

A:  + 

Today's goals

- Knowledge
  - What deep learning is and why it's powerful
  - What recurrent neural networks is and its application in NLP

- Experience
  - How to train your deep learning model
  - How to solve real world NLP problem with deep learning

Morning Schedule

- Morning:
  - Deep Learning I
    - What deep learning can do. √
    - Start playing with Gluon. ←
  - Deep Learning II
    - What is deep learning. How does the magic happen.
  - Deep Learning III
    - Use recurrent neural networks for NLP.

Setup for morning sessions

- Get it on your laptop
  - Get Conda (anaconda/miniconda)
    - https://conda.io/miniconda.html
  - Get today's notebooks
    - git clone https://github.com/szha/gluon-crash-course -b jsalt
  - Set up Conda environment (inside gluon-crash-course folder)
    - conda env create -f env/environment.yml
  - Open Jupyter Notebook, with notedown plugin
    - source activate jsalt_gluon_nlp
    - https://github.com/szha/gluon-crash-course#prerequisites

Resources

- Deep Learning — the Straight Dope
  - https://gluon.mxnet.io/
- A 60-minute Gluon Crash Course
  - https://gluon-crash-course.mxnet.io/
- GluonCV
  - http://gluon-cv.mxnet.io/
- GluonNLP
  - https://gluon-nlp.mxnet.io/
- MXNet User Forum
  - http://discuss.mxnet.io/
- MXNet Documentation
  - https://mxnet.apache.org/