UNIVERSITY OF WATERLOO

Factoid Question Answering

CS 898 – Project

June 12, 2017

Salman Mohammed
David R. Cheriton School of Computer Science
University of Waterloo

Motivation

Source: https://www.apple.com/newsroom/2017/01/hey-siri-whos-going-to-win-the-super-bowl/

Google

who is the wife of bill gates?

All News Images Videos Maps More Settings Tools

About 21,000,000 results (1.10 seconds)

Bill Gates / Spouse

Melinda Gates

m. 1994

Melinda Ann Gates, DBE is an American philanthropist, former Microsoft employee, and wife of Microsoft founder Bill Gates. She is co-founder of the Bill & Melinda Gates Foundation. Wikipedia

More about Melinda Gates

Source: Google

Examples

Q: Who is the Falcons quarterback in 2012?

A: Matt Ryan

Q: Where did George Harrison live before he died?

A: Liverpool

Q: Who were the parents of Queen Elizabeth I?

A: Anne Boleyn, Henry VIII of England

Task

simple factoid question answering

answers reference a single fact in the knowledge-base

Freebase – large knowledge base

17.8M million facts, 4M unique entities, 7523 relation types

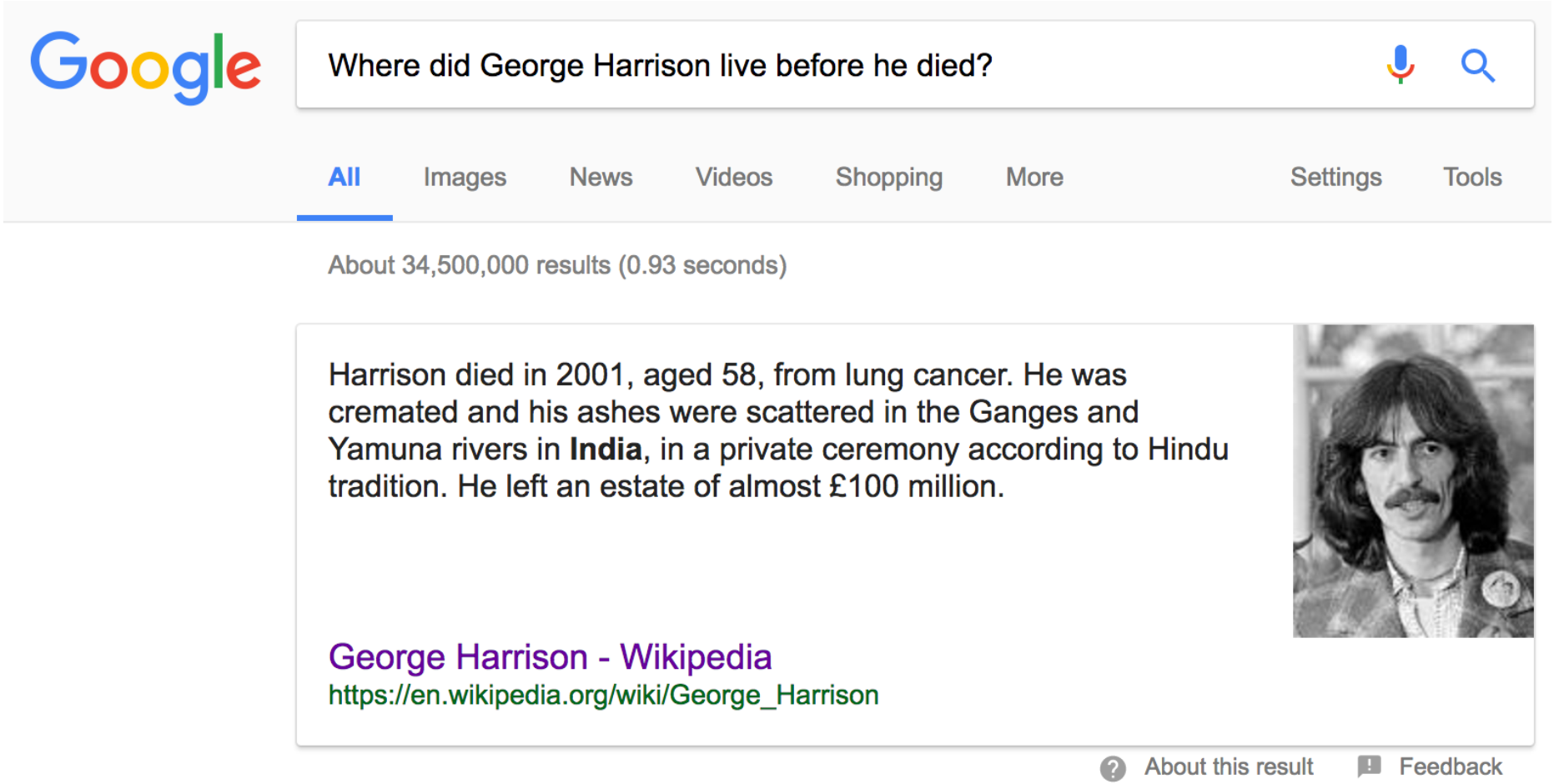
fact: Bahamas country/currency Bahamian_dollar

different from complex questions

Q: Who does David James play for in 2011?

Q: What year did Messi and Henry play together in Barcelona?

Not that simple…



Approach

Q: Who were the parents of Queen Elizabeth I?

A: Anne Boleyn, Henry VIII of England

Entity: Queen Elizabeth I

Freebase Entity MID: *m.02rg_*

Relation: */people/person/parents*

Lookup Freebase: query (entityid, relation)

Difficulties

no consistent way to do entity name → ID conversion

'JFK' could refer to a person, president, film, airport.

evaluate correct answer

'Cuban Convertible Peso' vs. 'Cuban Peso'

state-of-the-art accuracy: ~76%

many facts

long pipeline

Assuming you know…

Word Vectors

dense vector representation for words

word2vec, GloVe

Fully Connected Neural Networks

every node in a layer connected to all nodes in the previous layer

fixed size input(image) and output(classes)

Recurrent Neural Networks

modelling sequences

reasoning about previous events to make decision

Recurrent NNs

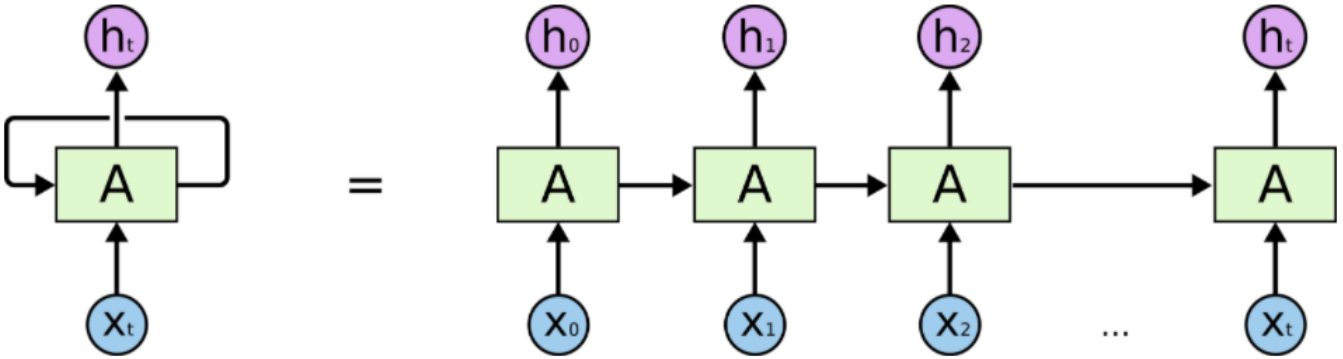
Input: x_t

word embedding

Memory/State: h_t

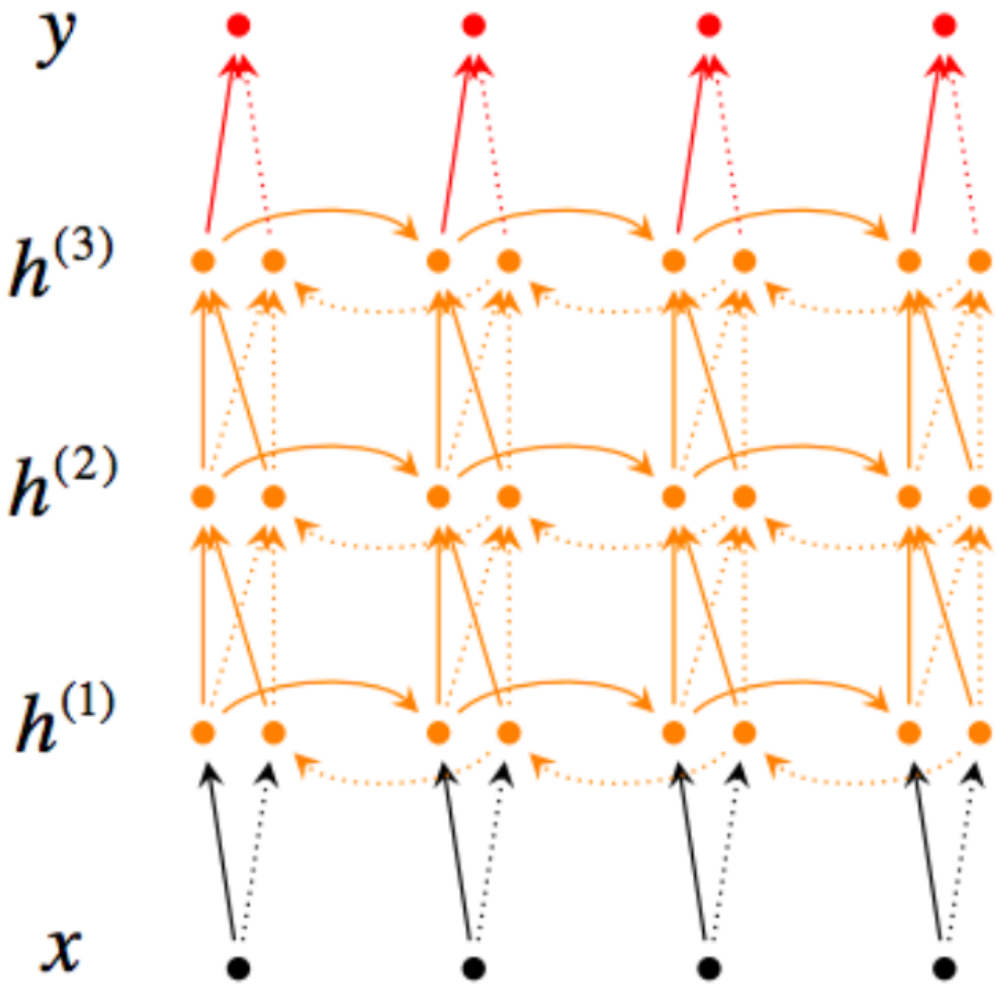
embedding based on current input and previous state

final state: think "sentence embedding"



An unrolled recurrent neural network.

Deep Bi-directional RNNs



Source: http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

Problem with RNNs

Learning long-term dependencies

"I grew up in France … I speak fluent _____."

Vanishing/Exploding gradient problem

notice that the same weight matrix is multiplied at each time step during forward and backward propagation

Long Short Term Memory Networks (LSTMs)

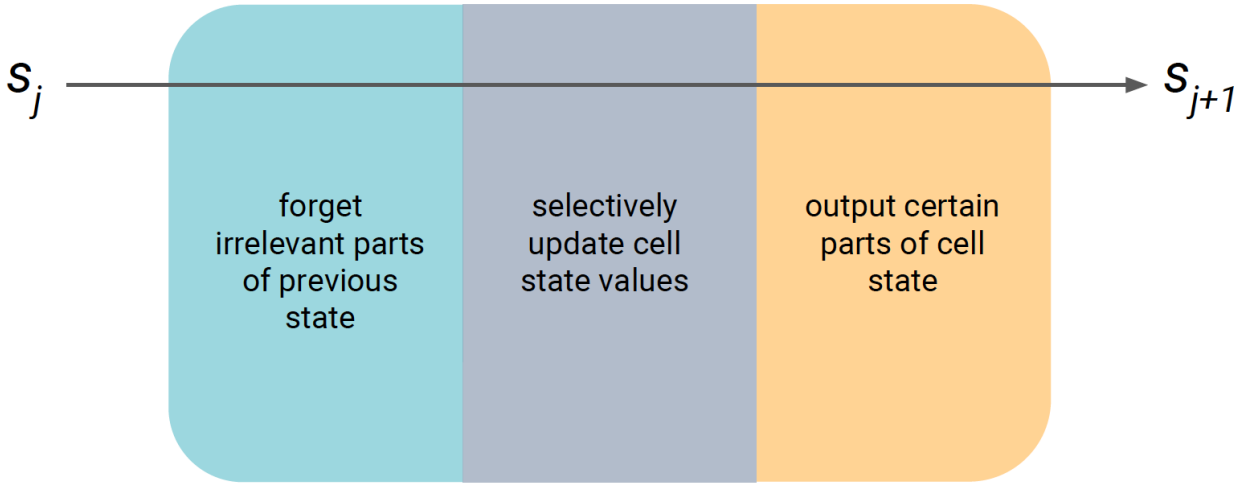
Avoid long term dependency problem

remember information for a long time

Idea: gated cells

complex node with gates controlling what information is passed through

maintains an additional "cell state" - c_t



Source: http://introtodeeplearning.com/Sequence%20Modeling.pdf


Method
Source: Google

Approach

Q: Who were the parents of Queen Elizabeth I?

A: Anne Boleyn, Henry VIII of England

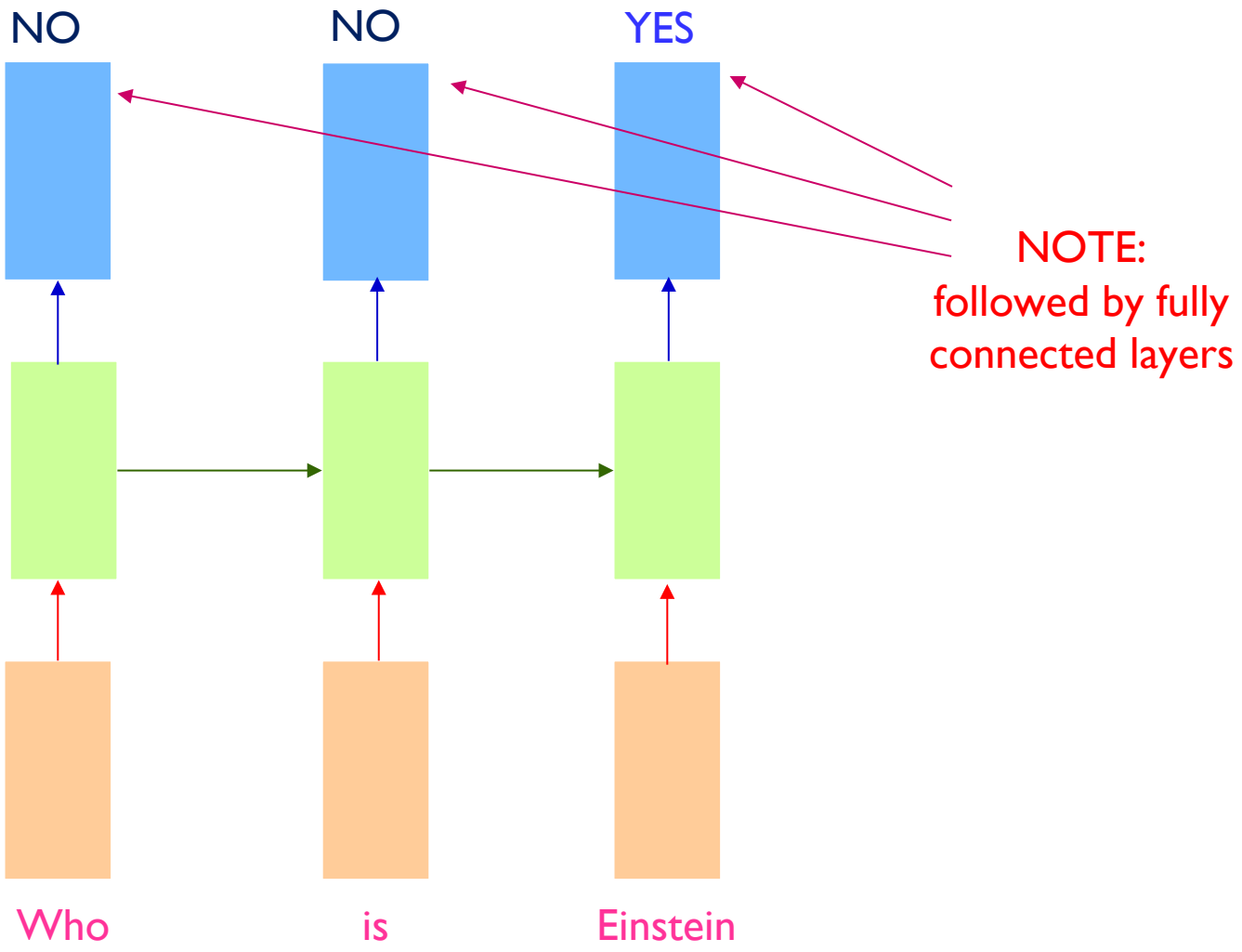
Entity: Queen Elizabeth I

Freebase Entity MID: *m.02rg_*

Relation: */people/person/parents*

Lookup Freebase: query (entityid, relation)

Entity Detection



Entity Linking

'Einstein' → '*m.013tyr*'

more than one entity refers to 'Einstein'

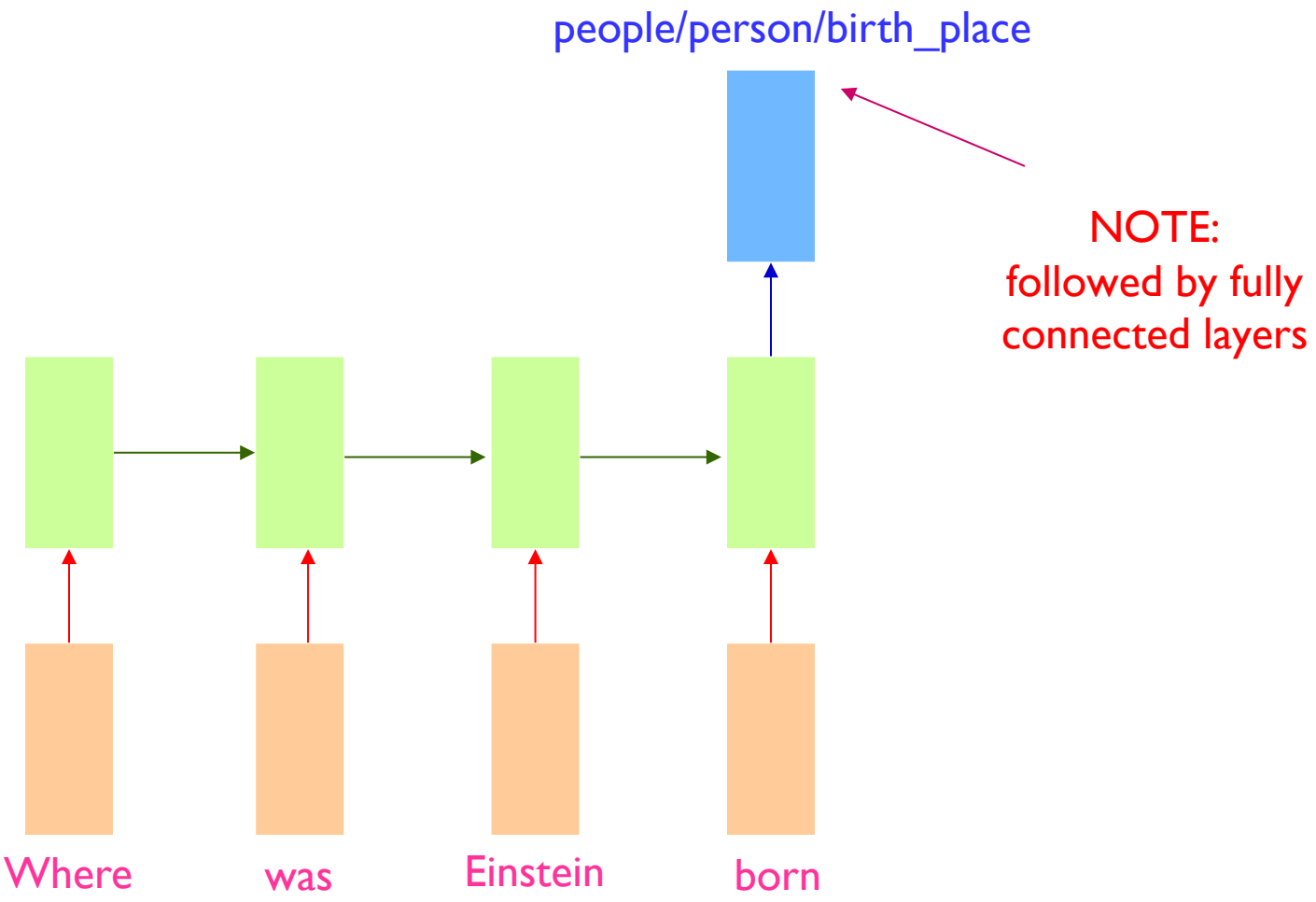
build a Lucene index of all entities

store entity MID as docid

store the name variants in different fields

ranked retrieval – BM25

Relation Prediction



Relation Prediction

- Dataset: Simple Questions
- Training set: ~76,000 examples
- Validation set: ~11,000 examples
- Number of classes: 1,837 relation types
- Model: Bi-directional LSTM (4 layers)
- Accuracy on validation set: ~81%

Other Ideas

joint-model the (entity, relation) pair

rank entities, relations and then, joint-model them

convolutional networks with attention modules

character level CNN for entity detection

word level CNN for relation prediction

Source: Google

Tricks of the Trade

Activation function: try ReLU

prevents from shrinking gradients

Optimization algorithm: try Adam

computes adaptive learning rate; usually faster convergence

read: http://sebastianruder.com/optimizing-gradient-descent/index.html

Weight initialization: use Xavier initialization

make sure weights start out 'just right'

Prevent overfitting: dropout, L2 regularization

dropout prevents feature co-adaptation

remember to scale model weights at test time for droput

Tricks of the Trade (cont'd)

Random Hyperparameter Search

grid search is a bad idea; read: https://arxiv.org/abs/1206.5533

some hyper-parameters more important than others

Batch Normalization

make activations unit gaussian distribution at the beginning of the training

insert BatchNorm layer immediately after fully-connected/convolutional layers

Initialize recurrent weight matrix, W^{hx} & W^{hh}, to identity matrix

helps vanishing gradient problem. read: https://arxiv.org/pdf/1504.00941.pdf

Gradient clipping

helps exploding gradient problem

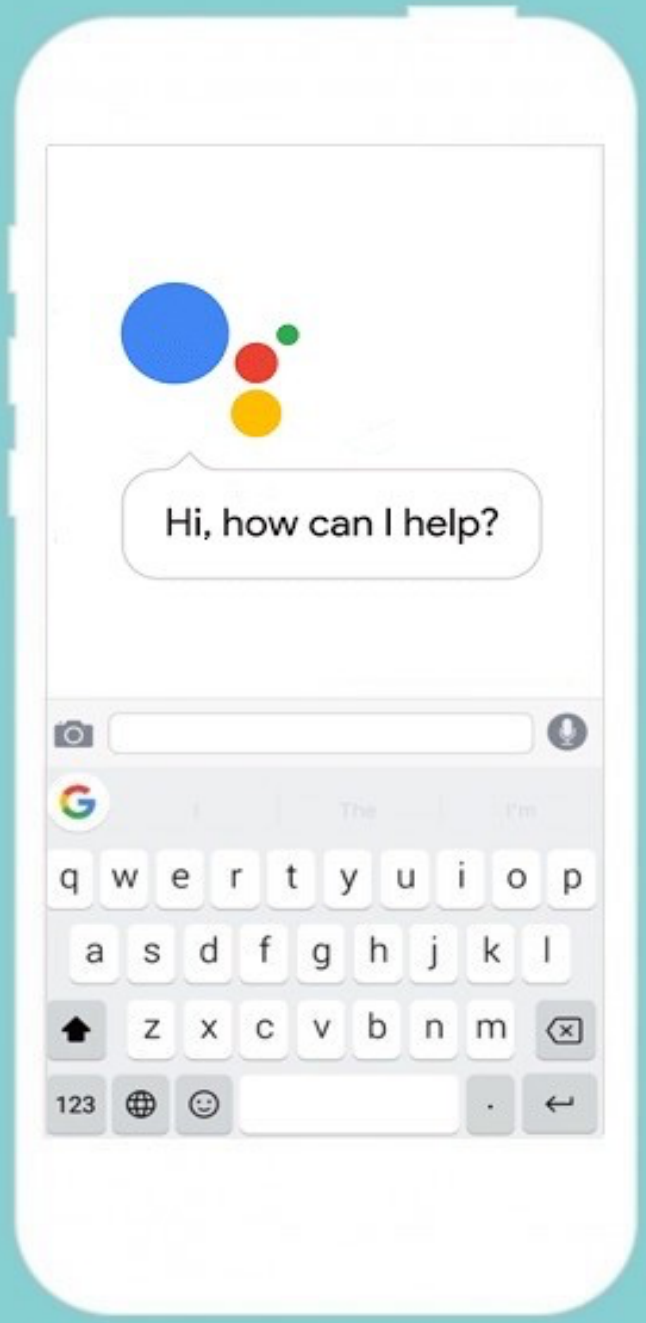
Acknowledgement

Wengpen Yin et al.
https://arxiv.org/abs/1606.03391

Ferhan Ture, Oliver Jojic
https://arxiv.org/abs/1606.05029

Christopher Olah
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Jimmy Lin
slide template taken from https://lintool.github.io/bigdata-2017w



Questions?

Source: Google