TensorFlow Lite

@NataliePis

09/11/2016 - TensorFlow

09/11/2016 - TensorFlow

14/11/2017 - TensorFlow Lite

09/11/2016 - TensorFlow

14/11/2017 - TensorFlow Lite

19/11/2018 - ?

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

How to ML

1. **Define the problem**
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

How to ML

1. Define the problem
2. **Gather data**
   - relevant to the task
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

How to ML

1. Define the problem
2. Gather data
3. **Prepare data**
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

- clean and pre-process
- randomize
- split: train/test

How to ML

1. Define the problem
2. Gather data
3. **Prepare data**
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

clean and pre-process
randomize
split: train/test
75/25

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. **Choose a model**
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

- learning task
- input type
- possible number of categories

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. **Train the model**
   - assign random values
   - predict the train data
   - adjust weights
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. **Evaluate the model**
   - check test data metrics
7. Tune the hyperparameters
8. Predict

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. **Tune the hyperparameters**
   - or, fine tune
8. Predict

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. **Predict**

TensorFlow

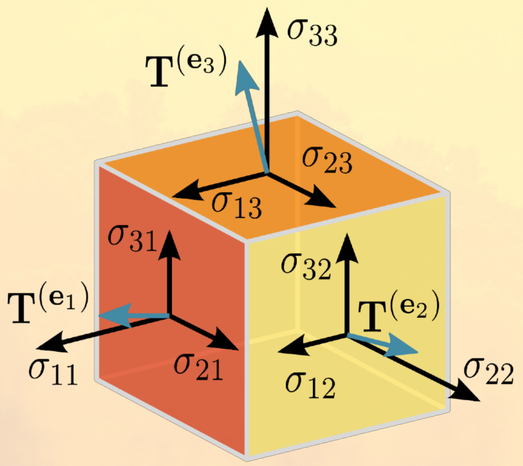
TensorFlow

TensorFlow is an open-source software
for Machine Intelligence,
used mainly for Machine Learning applications
such as neural networks.

TensorFlow

TensorFlow is an open-source software
for Machine Intelligence,
used mainly for Machine Learning applications
such as neural networks.

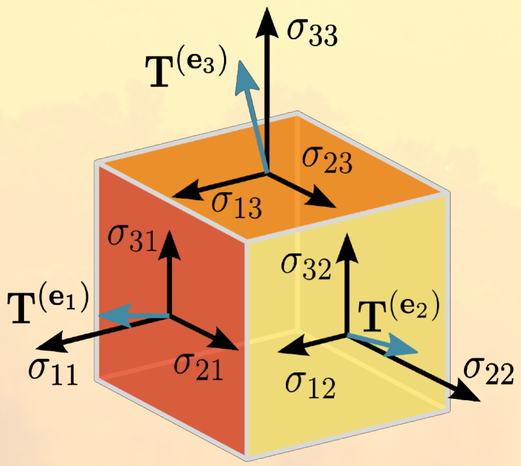
TensorFlow is an open-source software for Machine Intelligence,
used mainly for machine learning applications such as neural networks.

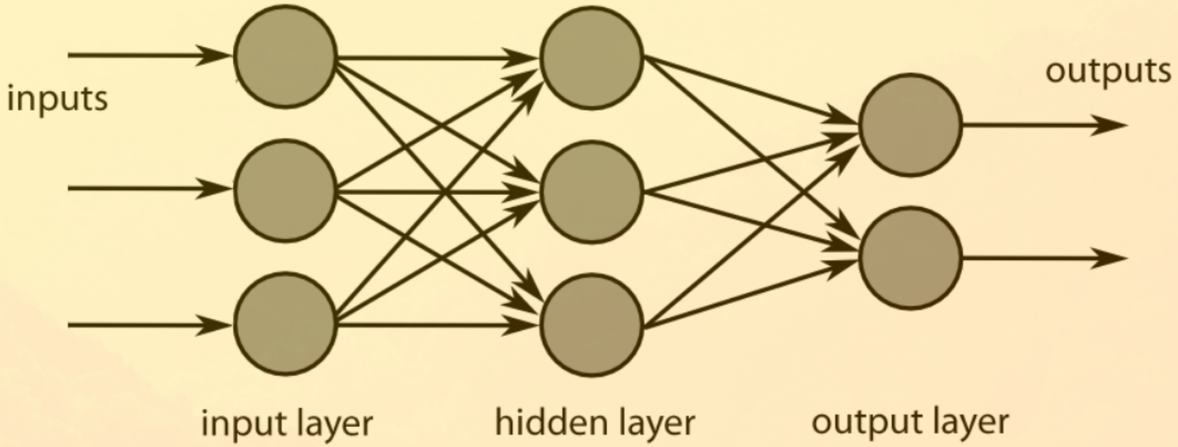


A tensor is a generalization
of vectors and matrices to
potentially higher dimensions

1. Data type
2. Shape
   - number of dimensions
   - number of values / dimension

TensorFlow is an open-source software for Machine Intelligence,
used mainly for machine learning applications such as neural networks.





A tensor is a generalization
of vectors and matrices to
potentially higher dimensions

1. Data type
2. Shape
   - number of dimensions
   - number of values / dimension

The flow part comes to describe:
- the graph (model) is a set of nodes (operations)
- the data (tensors) "flows" through those nodes,
  undergoing mathematical manipulation

You can look at, and evaluate, any node of the graph

HDR
PHOTO
SQUARE
PANO

HDR
PHOTO
SQUARE
PANO

TensorFlow Lite

TensorFlow Lite is the official solution for running machine learning models on mobile and embedded devices

TensorFlow Lite

TensorFlow Lite is the official solution for running machine learning models on mobile and embedded devices

TensorFlow Lite

TensorFlow Lite is the official solution for running machine learning models on mobile and embedded devices

TensorFlow Lite

TensorFlow Lite is the official solution for running machine learning models on mobile and embedded devices

TensorFlow Lite

- Not (yet) for model training
- Made for software developers
- Architected for many platforms of small devices
  - Small binary
  - Fast Inference

Offline

- No server calls
- Free: no need for hosting
- Data stays on device
- Direct access to the device's sensors

TensorFlow Lite

- Not (yet) for model training
- Made for software developers
- Architected for many platforms of small devices
  - Small binary
  - Fast Inference

Offline

- No server calls
- Free: no need for hosting
- Data stays on device
- Direct access to the device's sensors

TensorFlow Lite

?

TensorFlow mobile

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

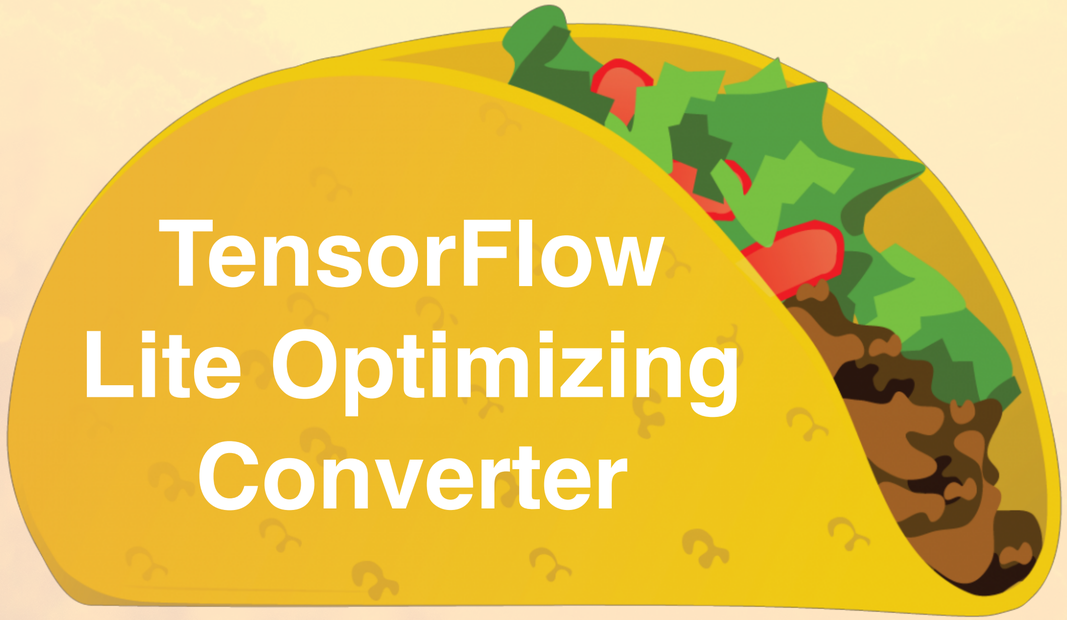
How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. **Optimize the model for mobile**
9. Predict

TensorFlow Lite Converter

https://www.tensorflow.org/lite/convert/

TensorFlow graph

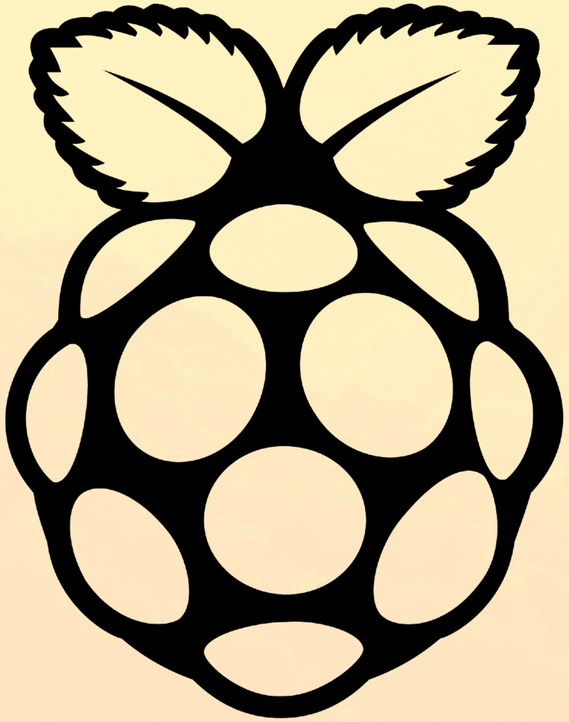


TensorFlow Lite graph

Environments

Environments

Environments



Models

Float Models:

- AutoML mobile image classification
- Image classification

Quantized Models

- Image classification

Models

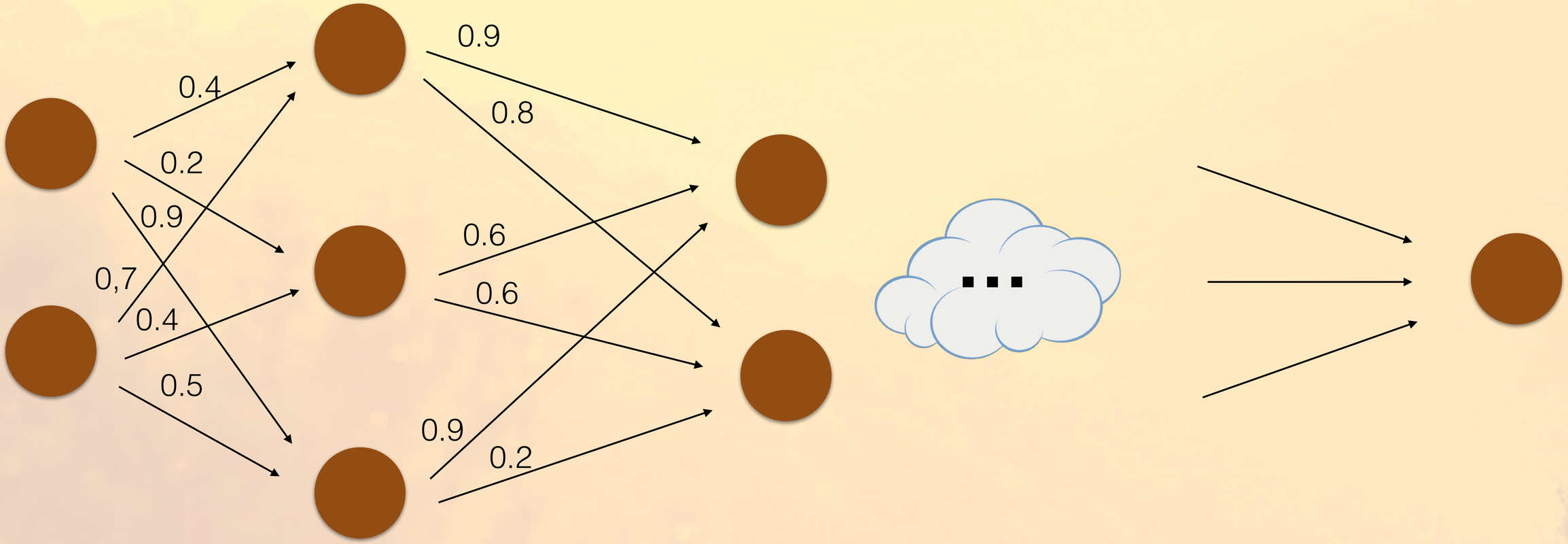
Float Models:

- AutoML mobile image classification
- Image classification

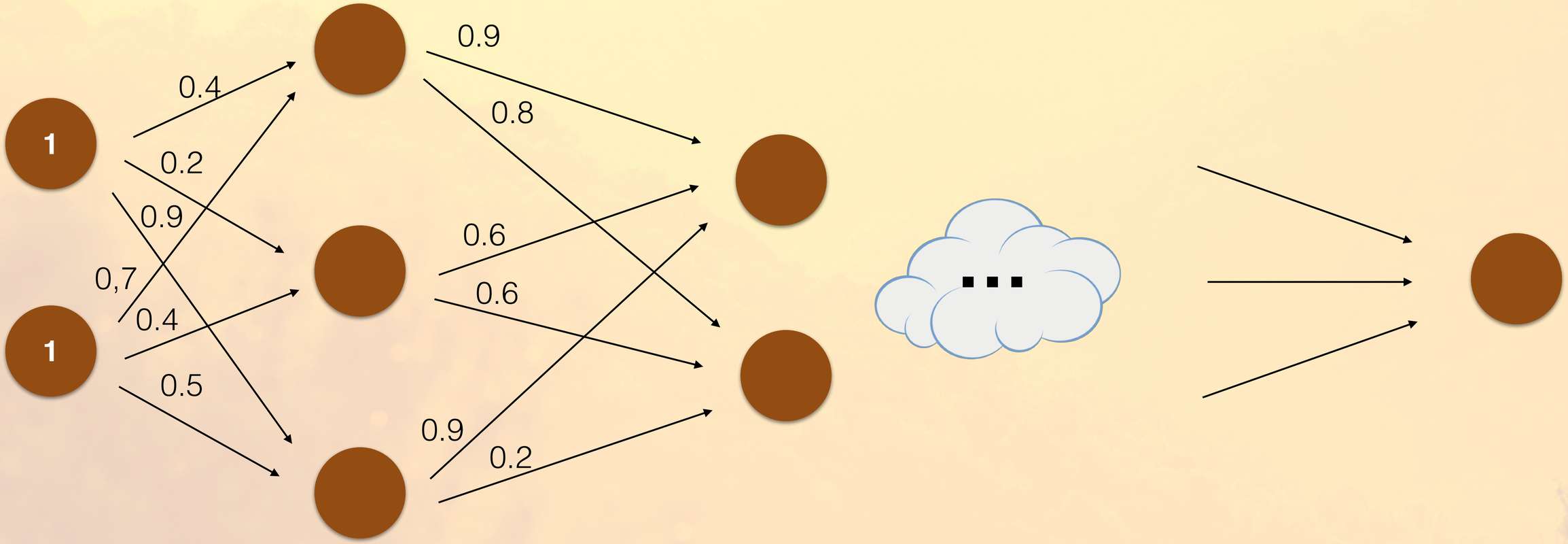
Quantized Models

- Image classification

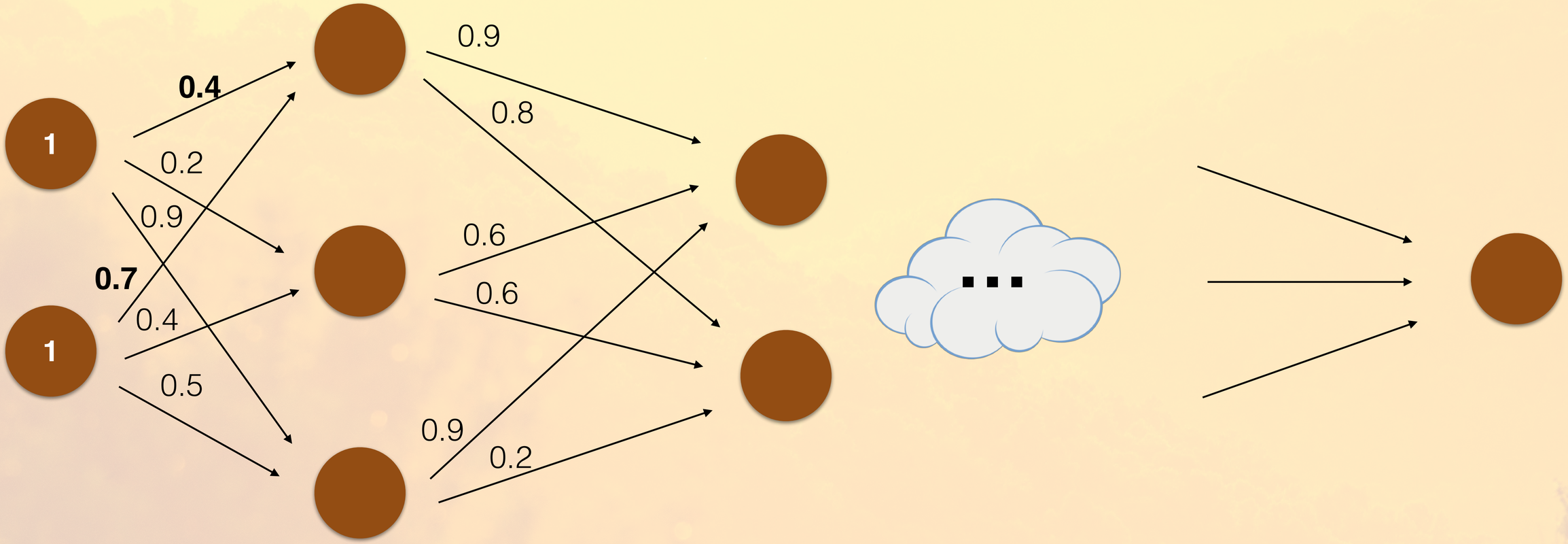
Models



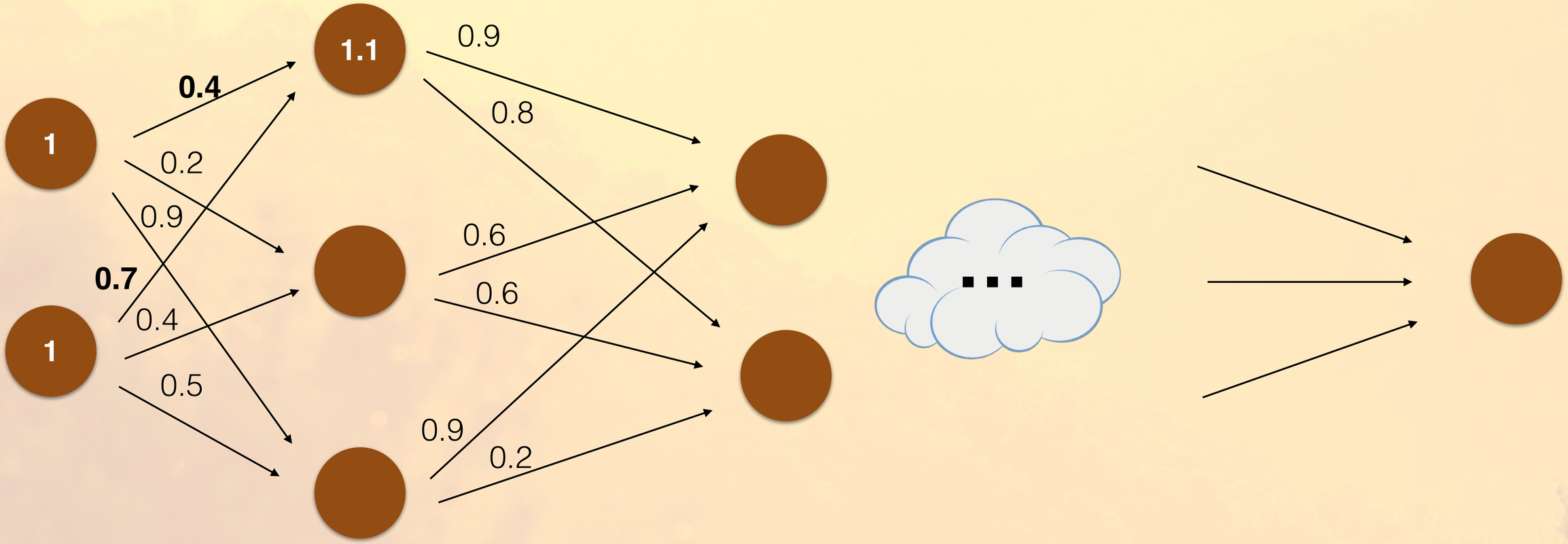
Models



Models

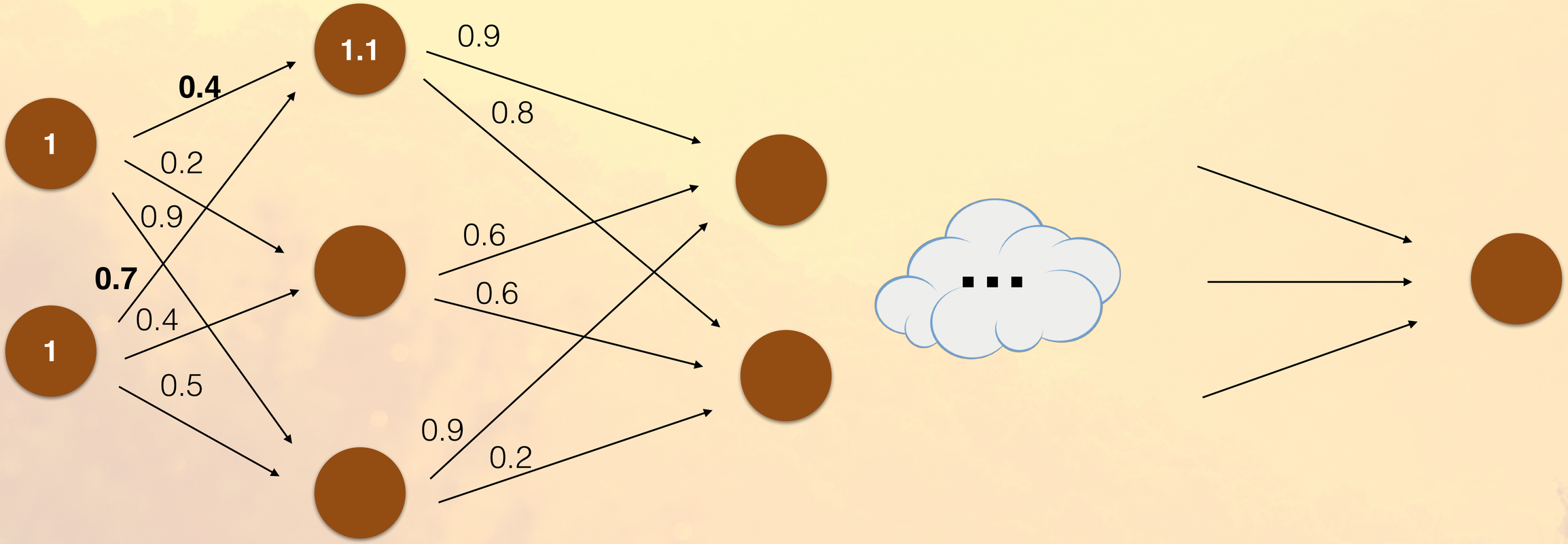


Models



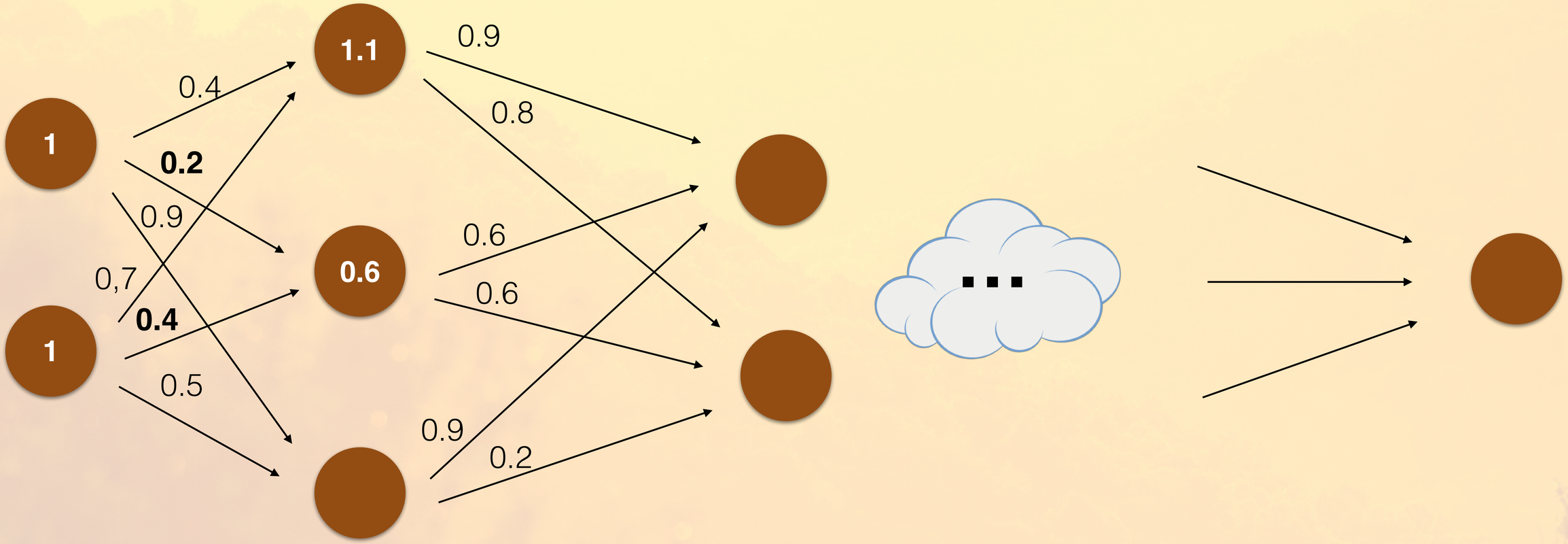
Models

0.4+0.7=1.1



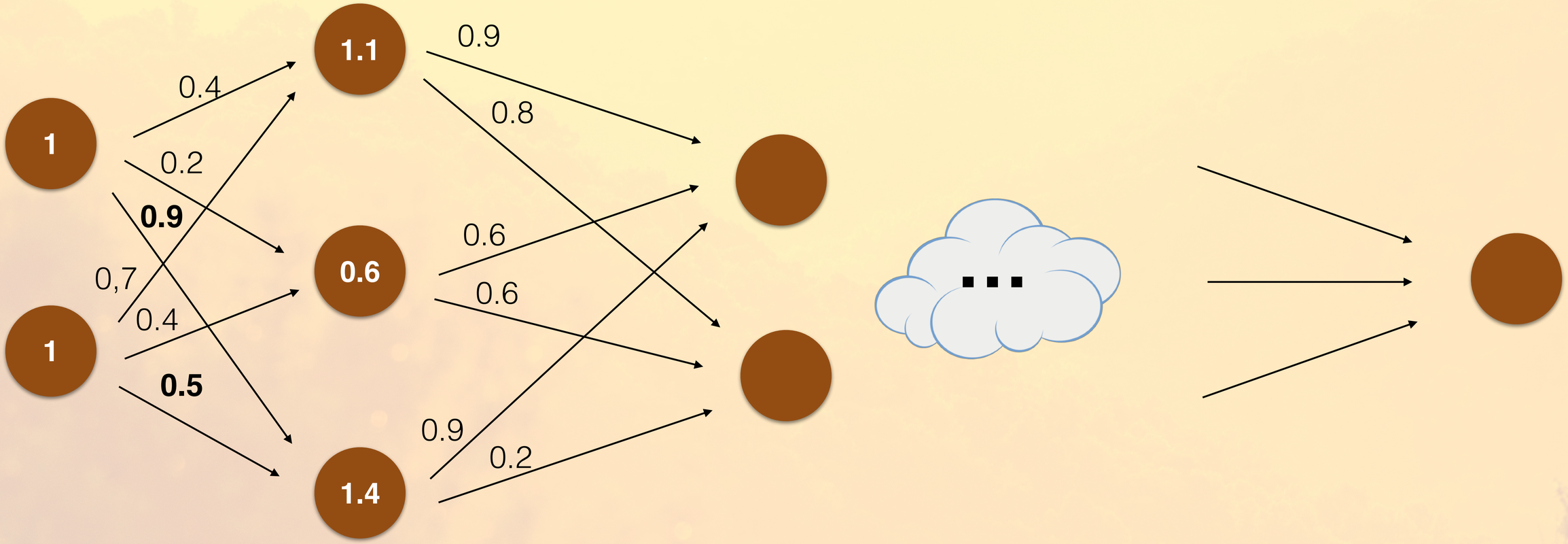
Models

0.2+0.4=0.6

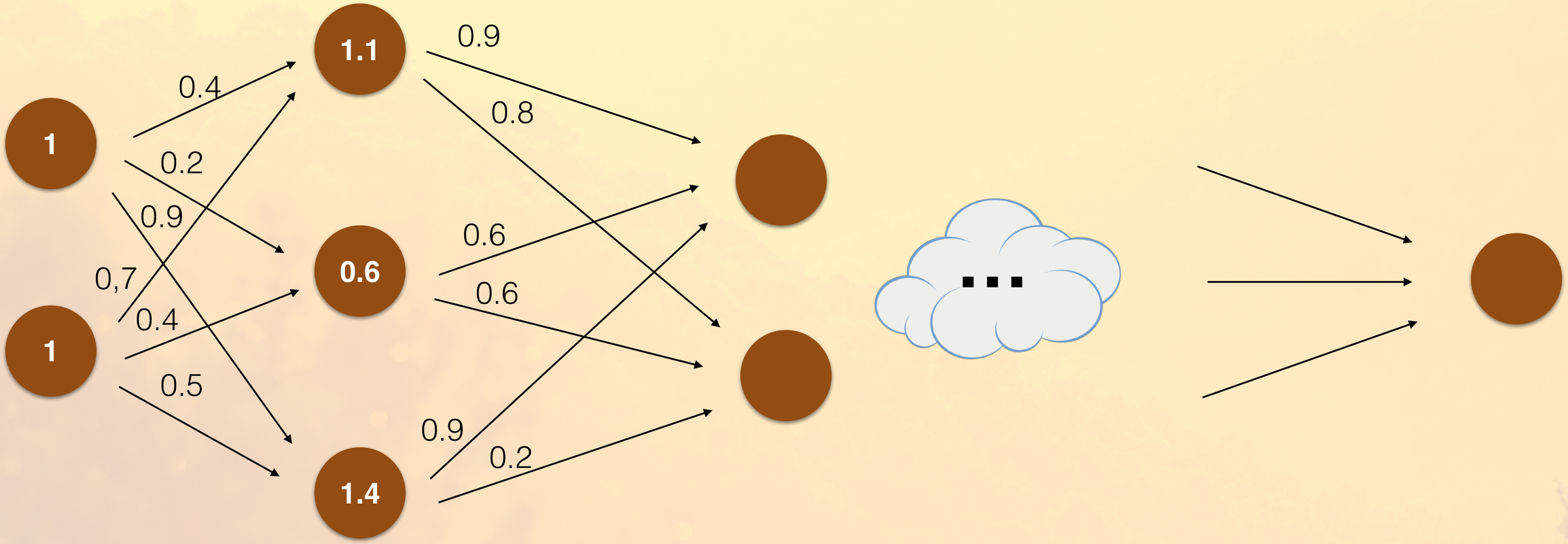


Models

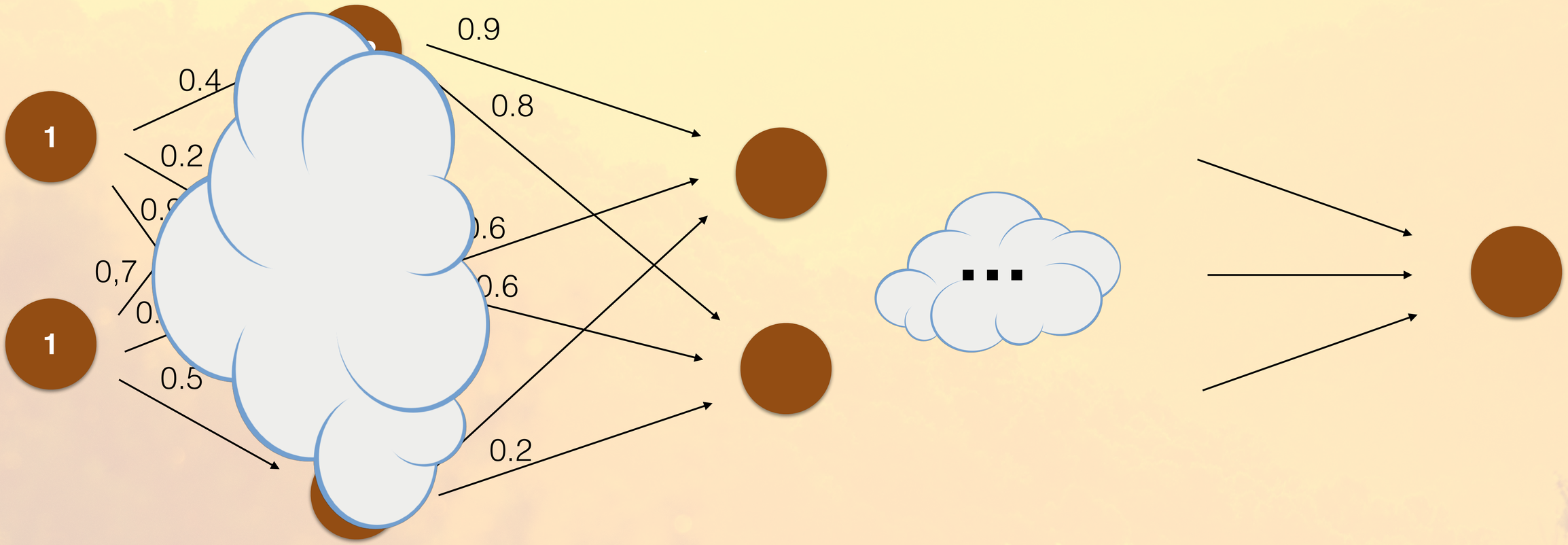
0.9+0.5=1.4



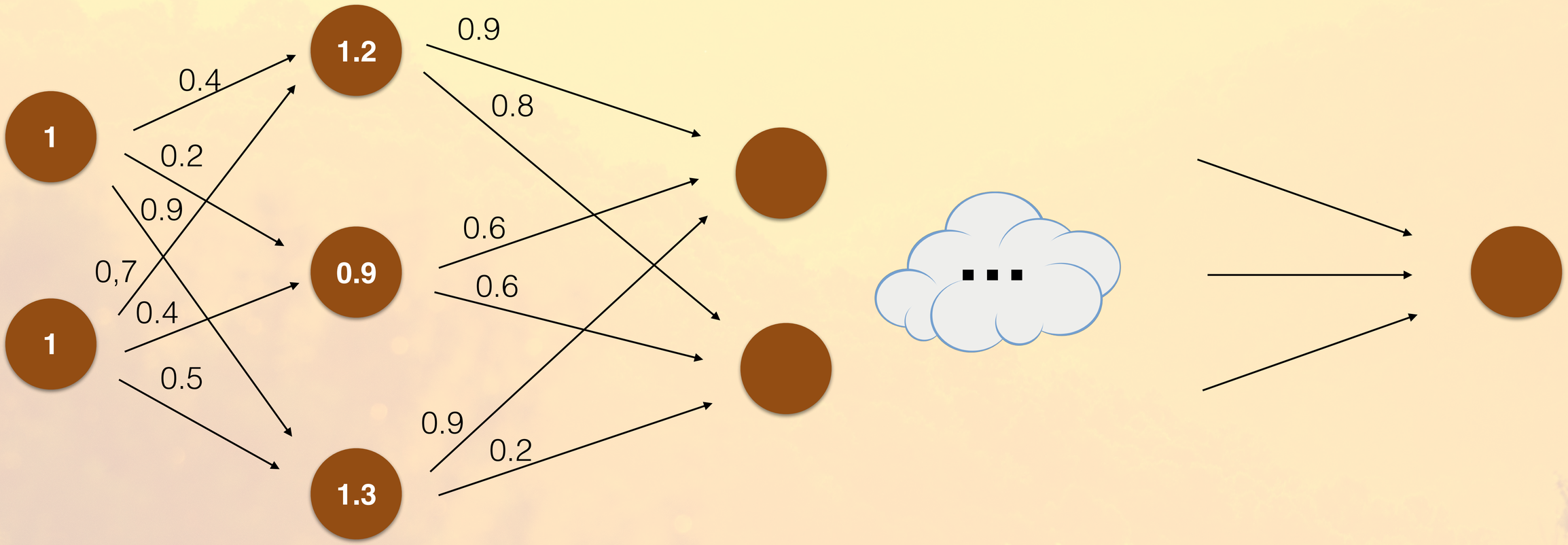
Models



Models

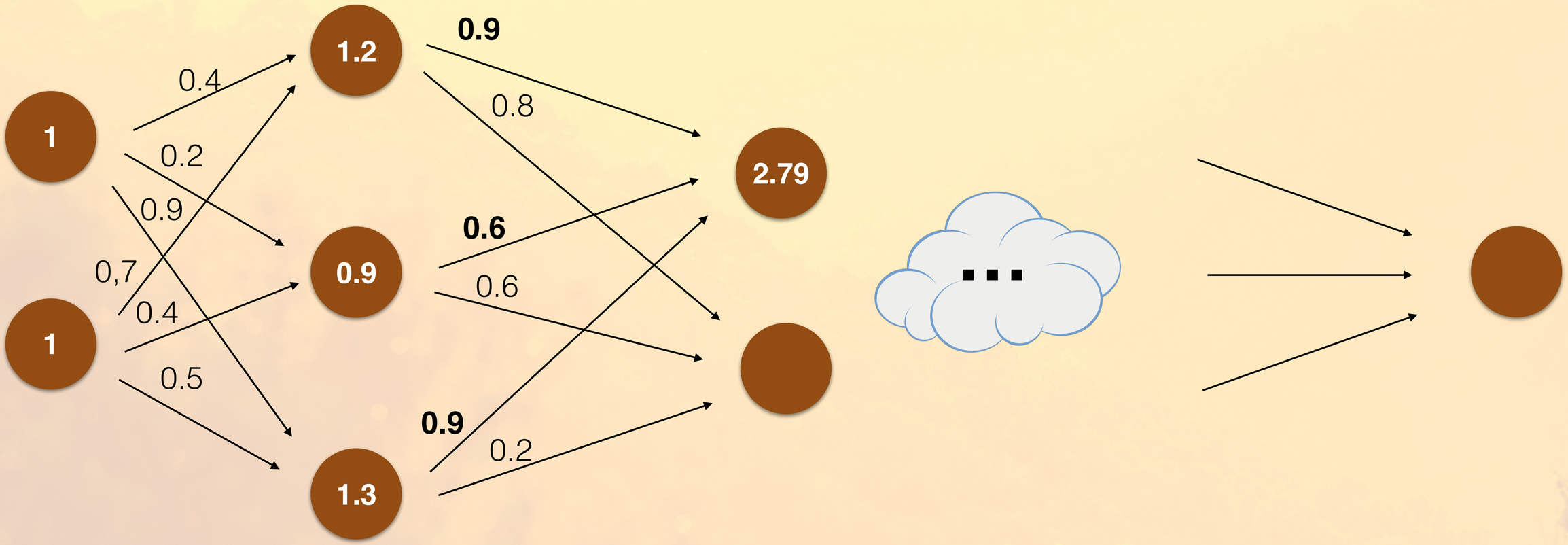


Models



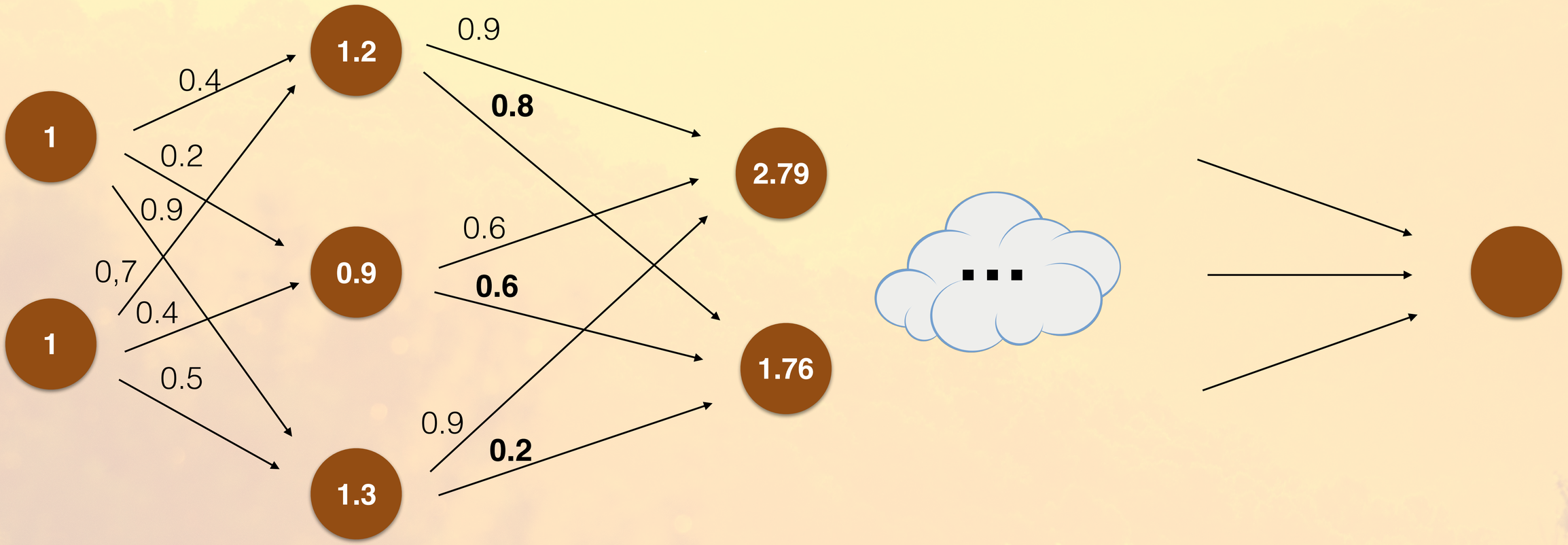
Models

1.2*0.9+0.9*0.6+1.3*0.9=2.79

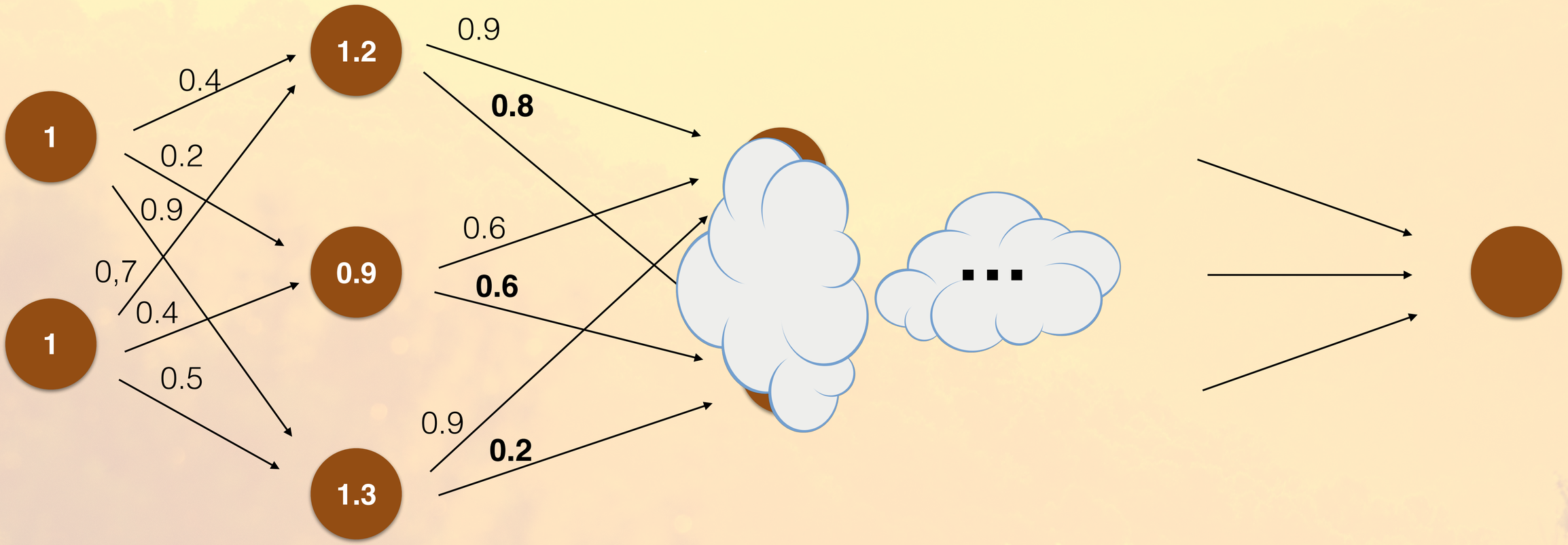


Models

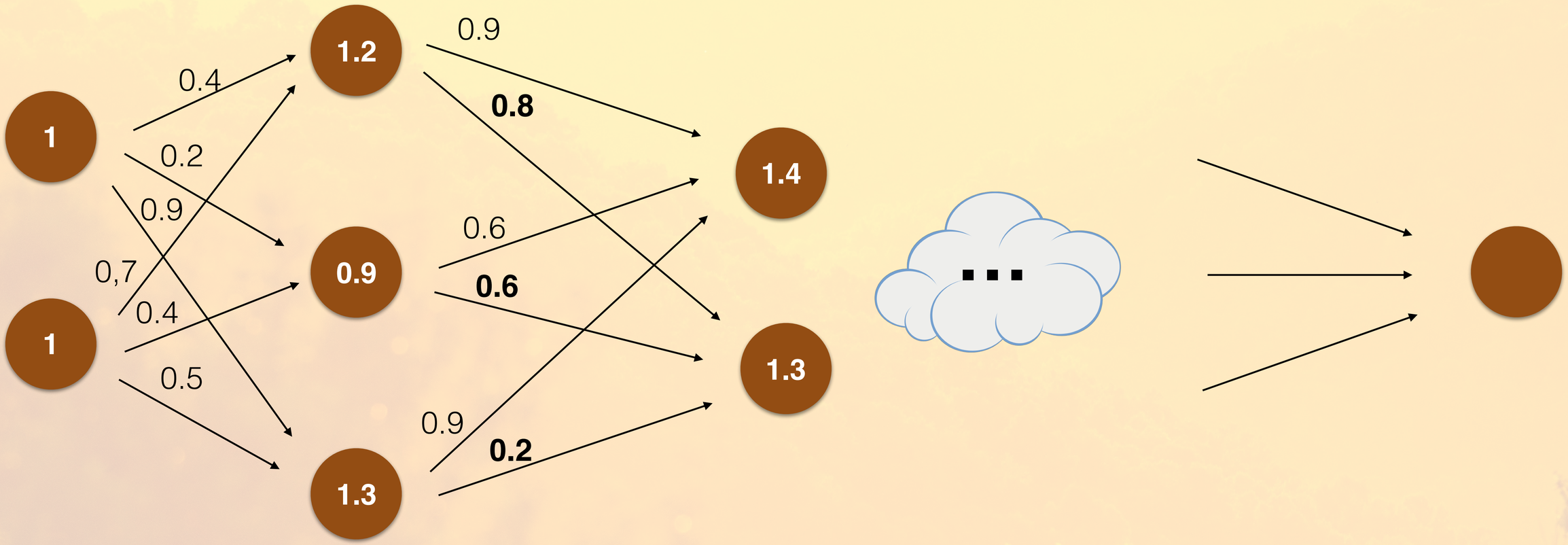
1.2*0.8+0.9*0.6+1.3*0.2=1.76



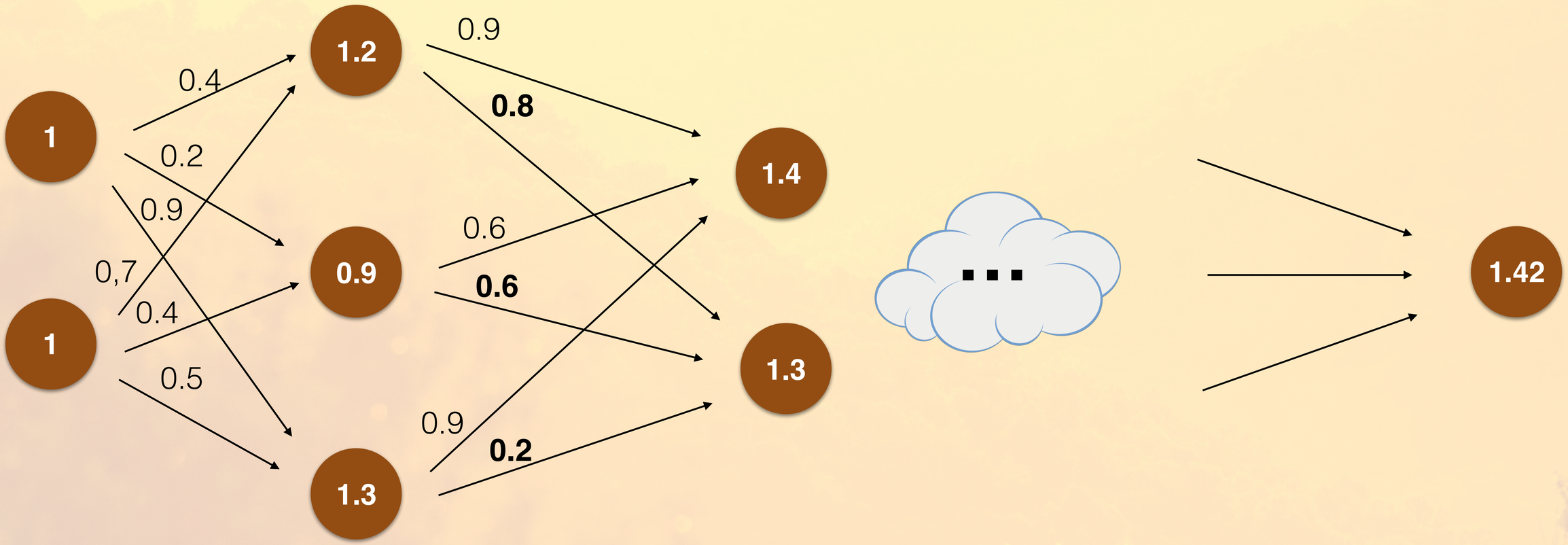
Models

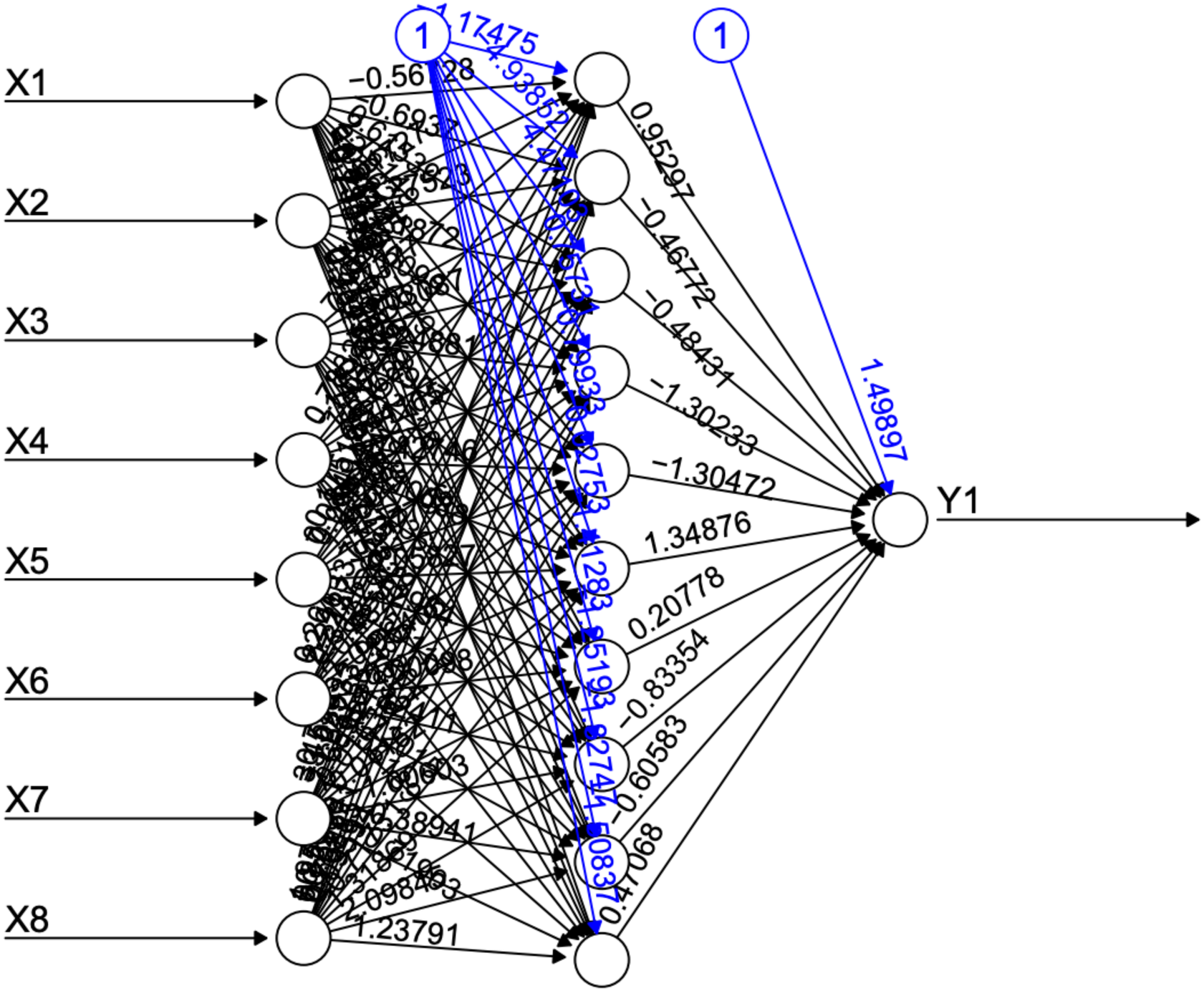


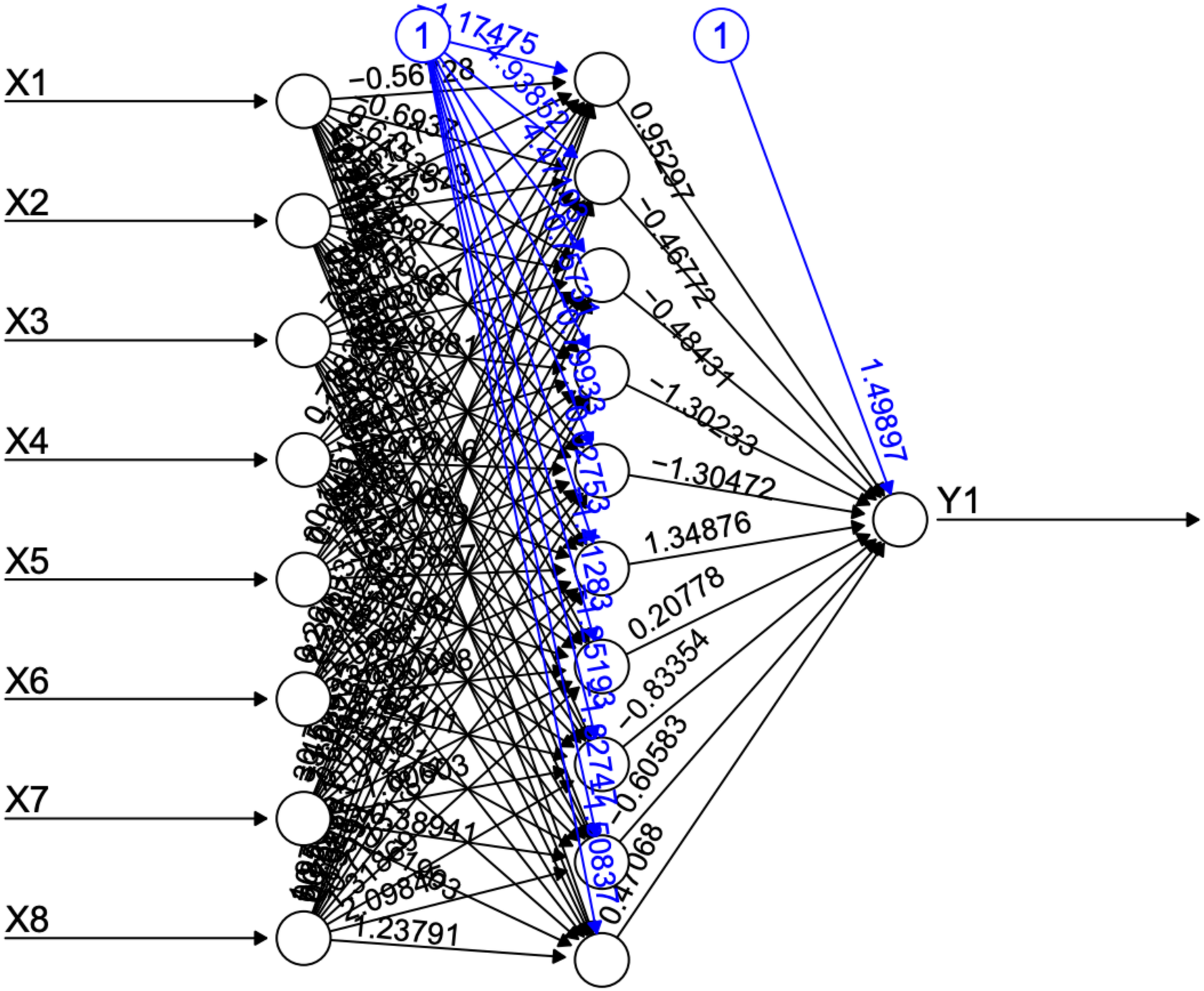
Models

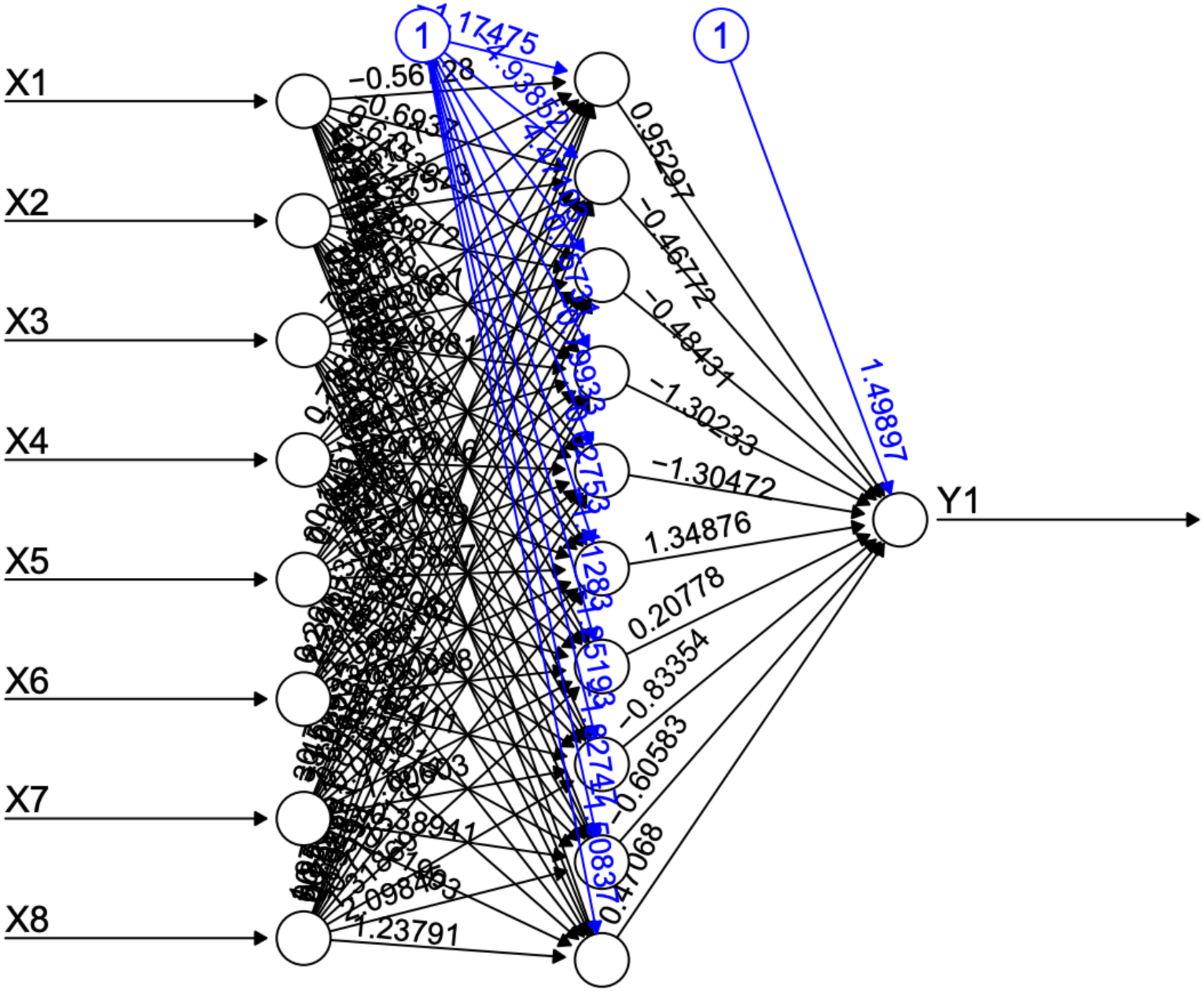


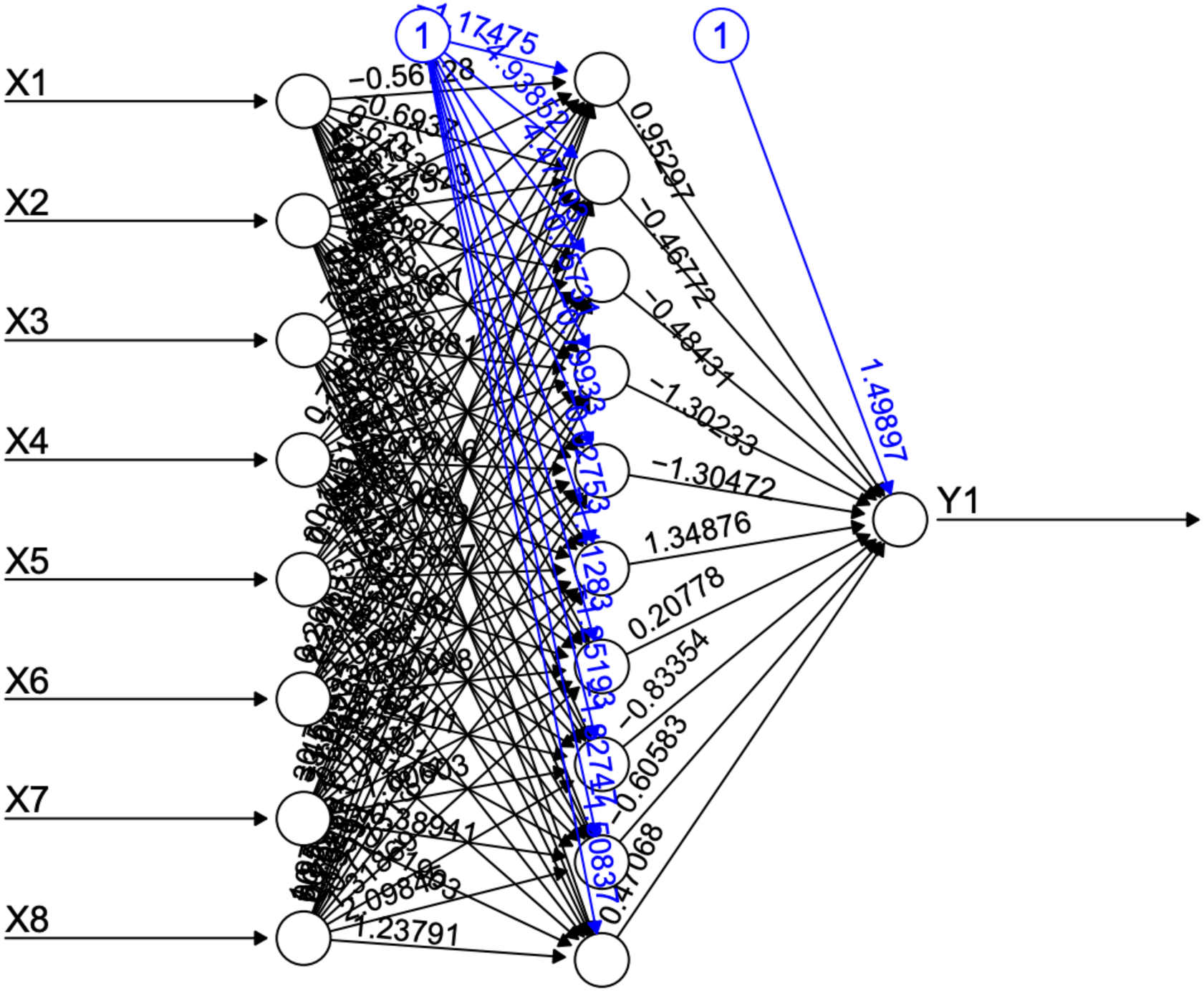
Models

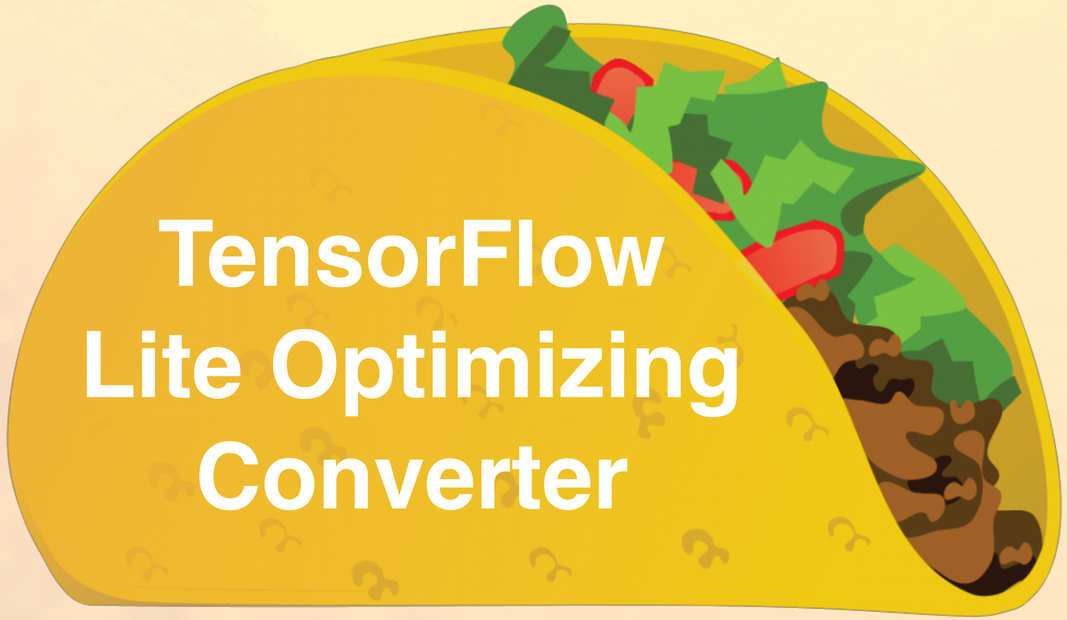



X1
X2
X3
X4
X5
X6
X7
X8
1
1
-0.56128
0.95297
-0.46772
-0.48434
-1.30233
-1.30472
1.34876
0.20778
-0.83354
-0.60583
0.47068
1.23791
1.49897
Y1


X1
X2
X3
X4
X5
X6
X7
X8
1
1
-0.56128
0.95297
-0.46772
-0.48434
-1.30233
-1.30472
1.34876
0.20778
-0.83354
-0.60583
0.47068
1.23791
1.49897
Y1


X1
X2
X3
X4
X5
X6
X7
X8
1
1
-0.56128
0.95297
-0.46772
-0.48434
-1.30233
-1.30472
1.34876
0.20778
-0.83354
-0.60583
0.47068
1.23791
1.49897
Y1


X1
X2
X3
X4
X5
X6
X7
X8
1
1
-0.56128
0.95297
-0.46772
-0.48434
-1.30233
-1.30472
1.34876
0.20778
-0.83354
-0.60583
0.47068
1.23791
1.49897
Y1


TensorFlow
Lite Optimizing
Converter

Models

Float Models:

- AutoML mobile image classification
- Image classification

Quantized Models

- Image classification

Float Models

- 25 available
- 1.9 Mb - 355.3 Mb
- 13-7940 ms
- Top 5 accuracy: 70.2%-96.1%
- e.g. NasNet, MobileNet

Quantized Models

- 18 available
- 0.5 Mb - 23 Mb
- 3.7-637 ms
- Top 5 accuracy: 64.4%-93.7%
- e.g. MobileNet, Inception

AutoML mobile - Float

- Float models
- 9 available
- 8.5 Mb - 24 Mb
- 37-152 ms
- Top 5 accuracy: 87.79%-92.55%
- Only MnasNet

	Float	Quantized	AutoML Float
Available Models	25	18	9
Size [Mb]	2 - 355	0.5 - 23	8 - 24
Latency [ms]	13 ms - 8 sec	4 - 637	37 - 152
Accuracy [%]	70 - 96	64-95	88-93

	Float	Quantized	AutoML Float
Available Models	25	18	9
Size [Mb]	2 - 355	0.5 - 23	8 - 24
Latency [ms]	13 ms - 8 sec	4 - 637	37 - 152
Accuracy [%]	70 - 96	64-95	88-93

	Float	Quantized	AutoML Float
Available Models	25	18	9
Size [Mb]	2 - 355	0.5 - 23	8 - 24
Latency [ms]	13 ms - 8 sec	4 - 637	37 - 152
Accuracy [%]	70 - 96	64-95	88-93

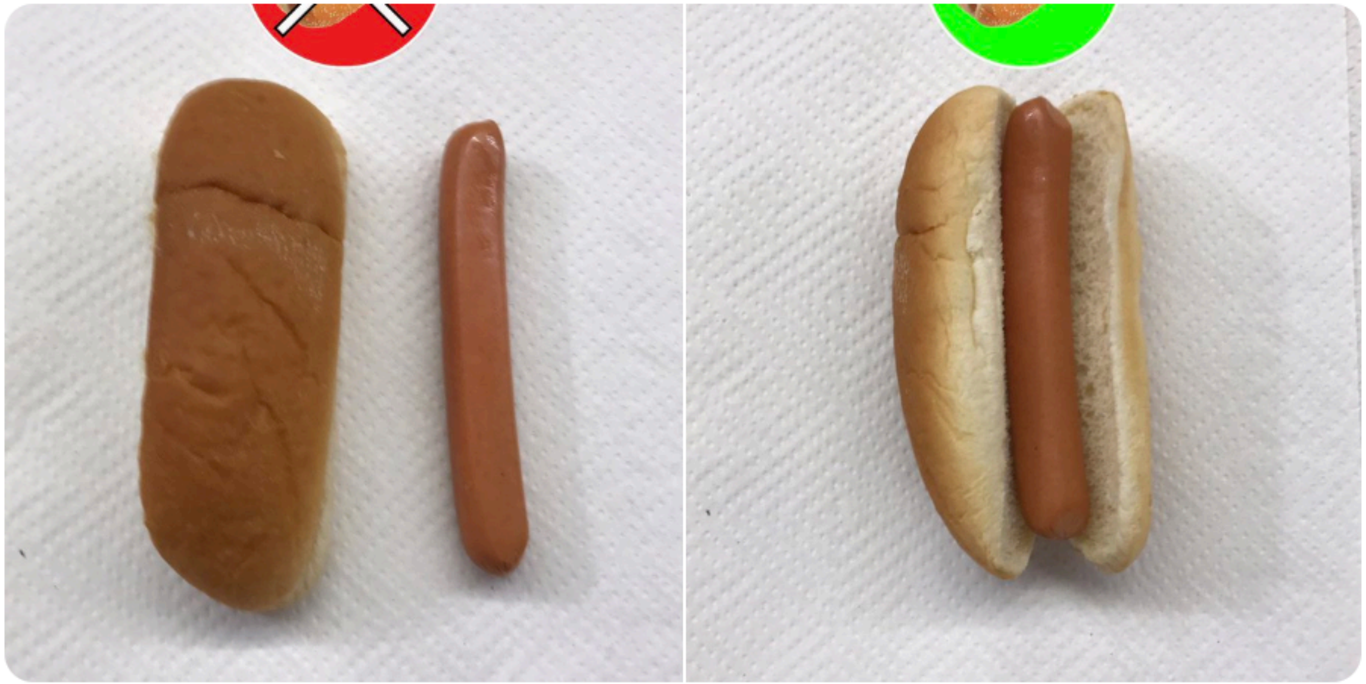
Hotdog
DELL
HBO

Hotdog
DELL
HBO

paul
@paulstachniak

Follow

Can confirm that the 'Not Hotdog' app is not easily fooled. #SiliconValleyHBO @SiliconHBO



1:17 PM - 19 May 2017

7 Retweets **30** Likes

7 30

Apps Using TFLite

Photos

GBoard

Gmail

Nest

Assistant

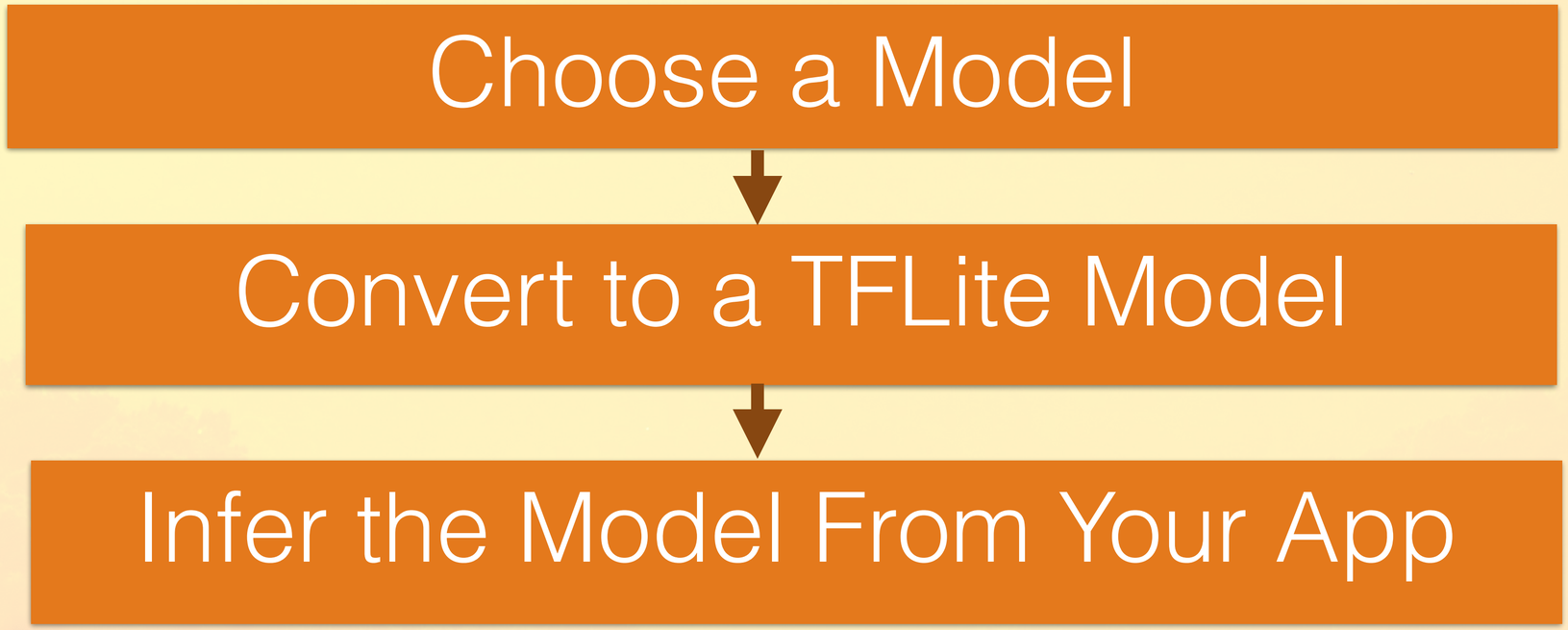
PicsArt

VSCO

Shazam

ML Kit

And many more...


Choose a Model
Convert to a TFLite Model
Infer the Model From Your App

Run a Demo

Run a Pretrained Model

Build a Custom Model

Optimize a Model

Run a Demo

Run a Pretrained Model

Build a Custom Model

Optimize a Model

Overview
Developer guide
Android demo app
iOS demo app

TensorFlow Lite APIs
Custom operators
TensorFlow Lite ops versioning
TensorFlow Lite compatibility guide
List of hosted models
TensorFlow Lite for iOS
TensorFlow Lite for Raspberry Pi

Android Demo App

Contents

Download the pre-built binary
Build in Android Studio with TensorFlow Lite AAR from JCenter
Using other models
Build TensorFlow Lite and the demo app from source
...

An example Android application using TensorFLow Lite is available on GitHub. The demo is a sample camera app that classifies images continuously using either a quantized Mobilenet model or a floating point Inception-v3 model. To run the demo, a device running Android 5.0 (API 21) or higher is required.

https://www.tensorflow.org/lite/demo_android

Overview
Developer guide
Android demo app
iOS demo app

TensorFlow Lite APIs
Custom operators
TensorFlow Lite ops versioning
TensorFlow Lite compatibility guide
List of hosted models
TensorFlow Lite for iOS
TensorFlow Lite for Raspberry Pi

iOS Demo App

Contents

Prerequisites
Building the iOS Demo App

The TensorFlow Lite demo is a camera app that continuously classifies whatever it sees from your device's back camera, using a quantized MobileNet model. These instructions walk you through building and running the demo on an iOS device.

Prerequisites

https://www.tensorflow.org/lite/demo_ios

Run a Demo

Run a Pretrained Model

Build a Custom Model

Optimize a Model

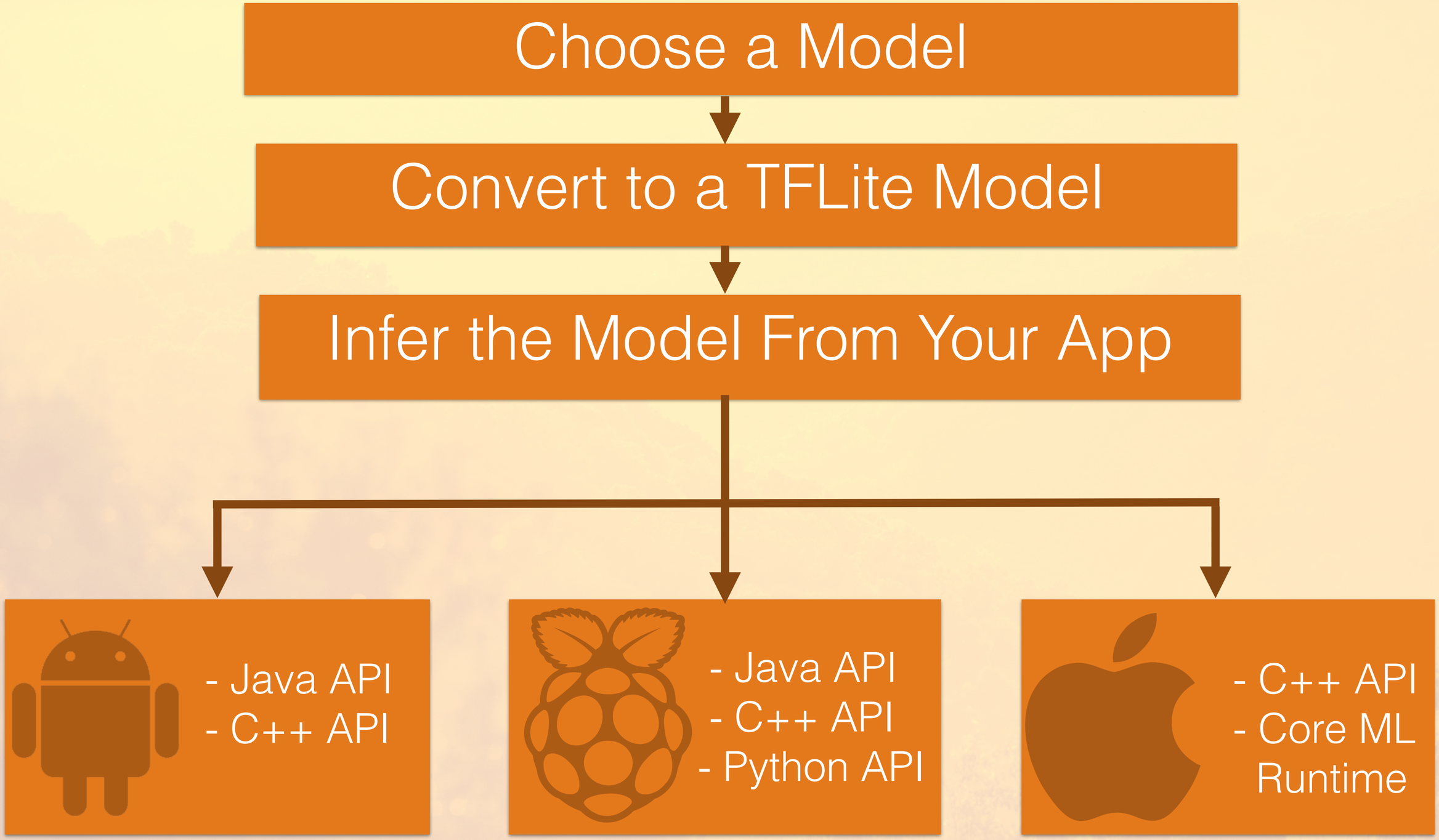
Choose a Model

↓

Convert to a TFLite Model

↓

Infer the Model From Your App


Choose a Model
Convert to a TFLite Model
Infer the Model From Your App
- Java API
- C++ API
- Java API
- C++ API
- Python API
- C++ API
- Core ML Runtime

Overview
Developer guide
Android demo app
iOS demo app

TensorFlow Lite APIs
Custom operators
TensorFlow Lite ops versioning
TensorFlow Lite compatibility guide
List of hosted models
TensorFlow Lite for iOS
TensorFlow Lite for Raspberry Pi

Build TensorFlow Lite for iOS

Contents

Building
Using in your own application

This document describes how to build TensorFlow Lite iOS library. If you just want to use it, the easiest way is using the TensorFlow Lite CocoaPod releases. See TensorFlow Lite iOS Demo for examples.

Building

To create a universal iOS library for TensorFlow Lite, you need to build it using Xcode's command line tools on a MacOS

https://www.tensorflow.org/lite/ios

Overview
Developer guide
Android demo app
iOS demo app

TensorFlow Lite APIs
Custom operators
TensorFlow Lite ops versioning
TensorFlow Lite compatibility guide
List of hosted models
TensorFlow Lite for iOS
TensorFlow Lite for Raspberry Pi

TensorFlow Lite for Raspberry Pi

Contents

Cross compiling
Installing the toolchain
Building
Native compiling

Cross compiling

Installing the toolchain

https://www.tensorflow.org/lite/rpi

Run a Demo

Run a Pretrained Model

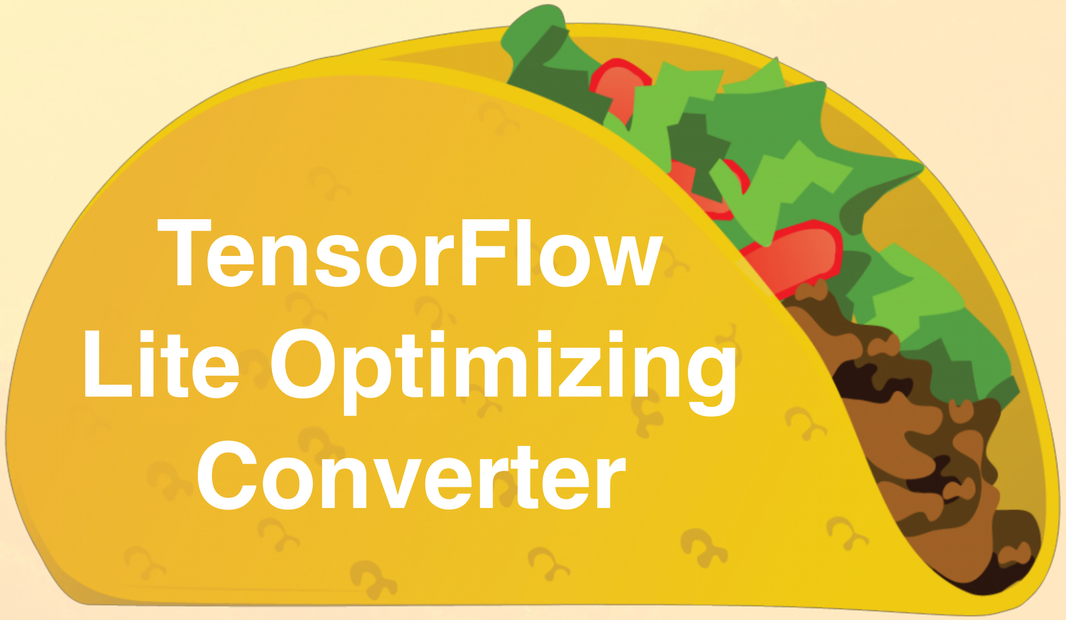
Build a Custom Model

Optimize a Model

Resources

How to ML

+



Run a Demo

Run a Pretrained Model

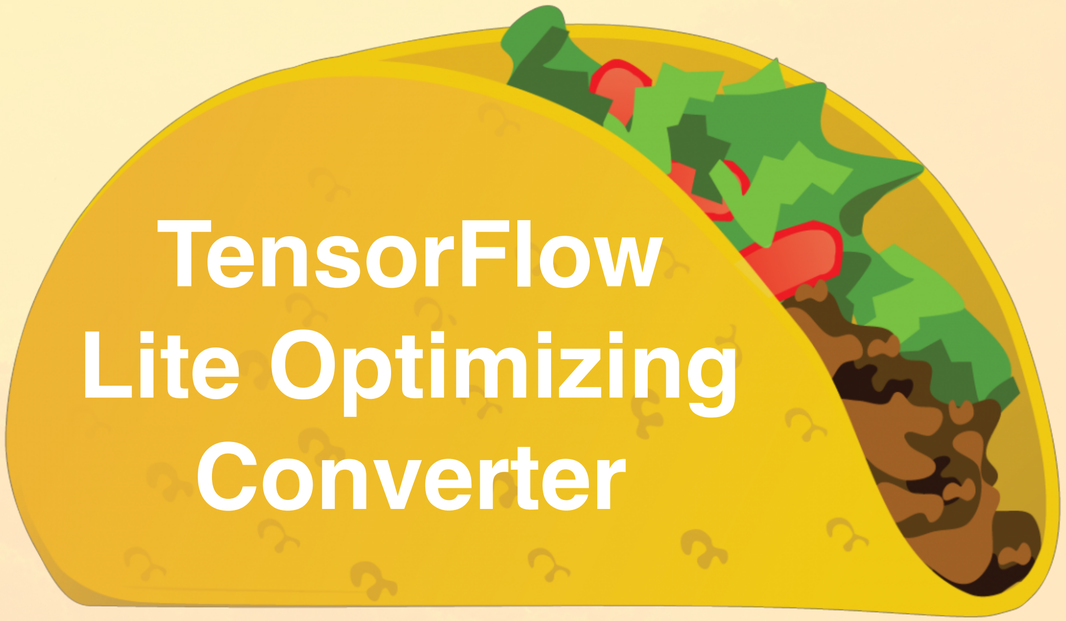
Build a Custom Model

Optimize a Model

Resources

Your Model

+



Resources

github.com/tensorflow/tensorflow
tensorflow.org/lite

Dev guide: tensorflow.org/lite/devguide
Hosted models: tensorflow.org/lite/models
Slack: tensor-flow-talk-invite.herokuapp.com/invite
Twitter: twitter.com/tensorflow
Mailing List: tflite@tensorflow.org

TFLite Timeline

On 31/10 it moved on GitHub!
//tensorflow/contrib/lite → //tensorflow/lite

Comming soon: OpenGL-based GPU acceleration for Android

Future plans: Edge-TPU and generally hardware acceleration

Upcoming API support: C, C#, C++

09/11/2016 - TensorFlow

14/11/2017 - TensorFlow Lite

19/11/2018 - ?

Thank you!

S

@NataliePis