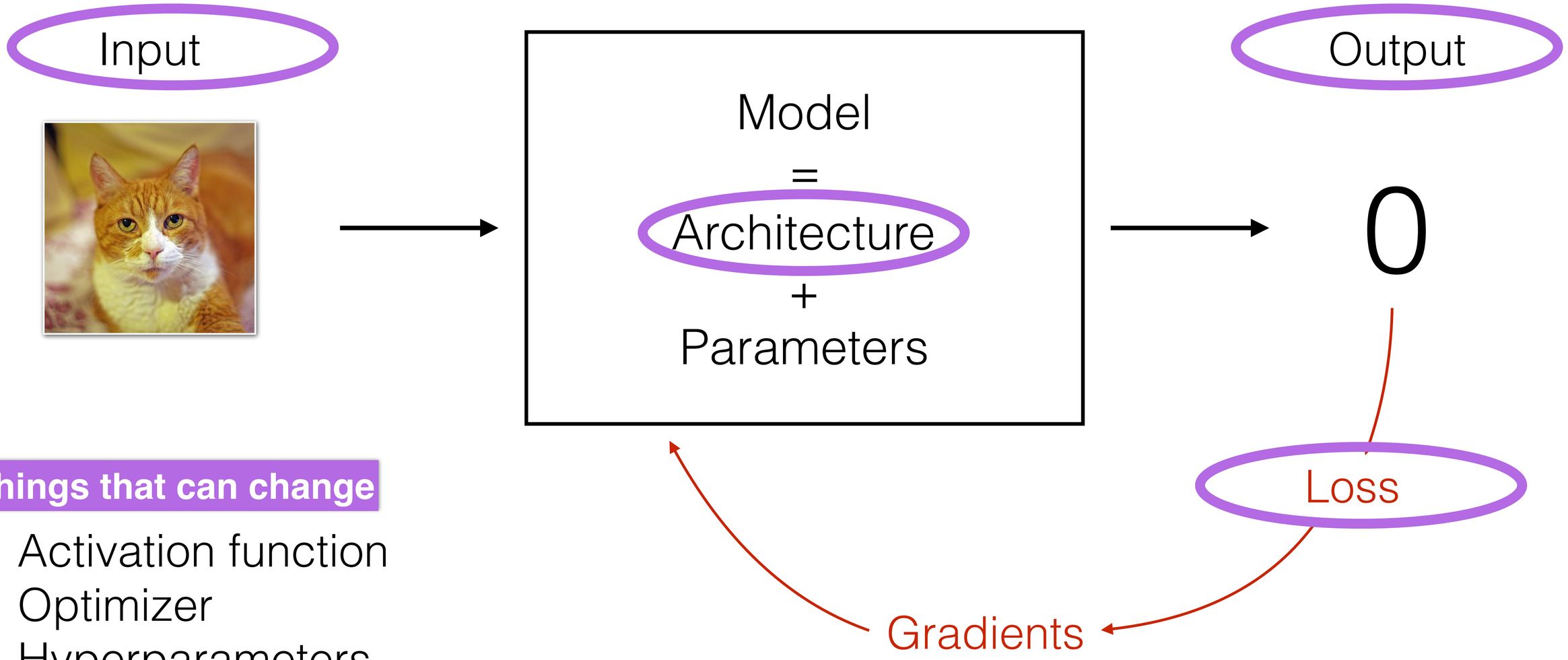
CS230: Lecture 3

Deep Learning Intuition

Kian Katanforoosh

Recap

Learning Process

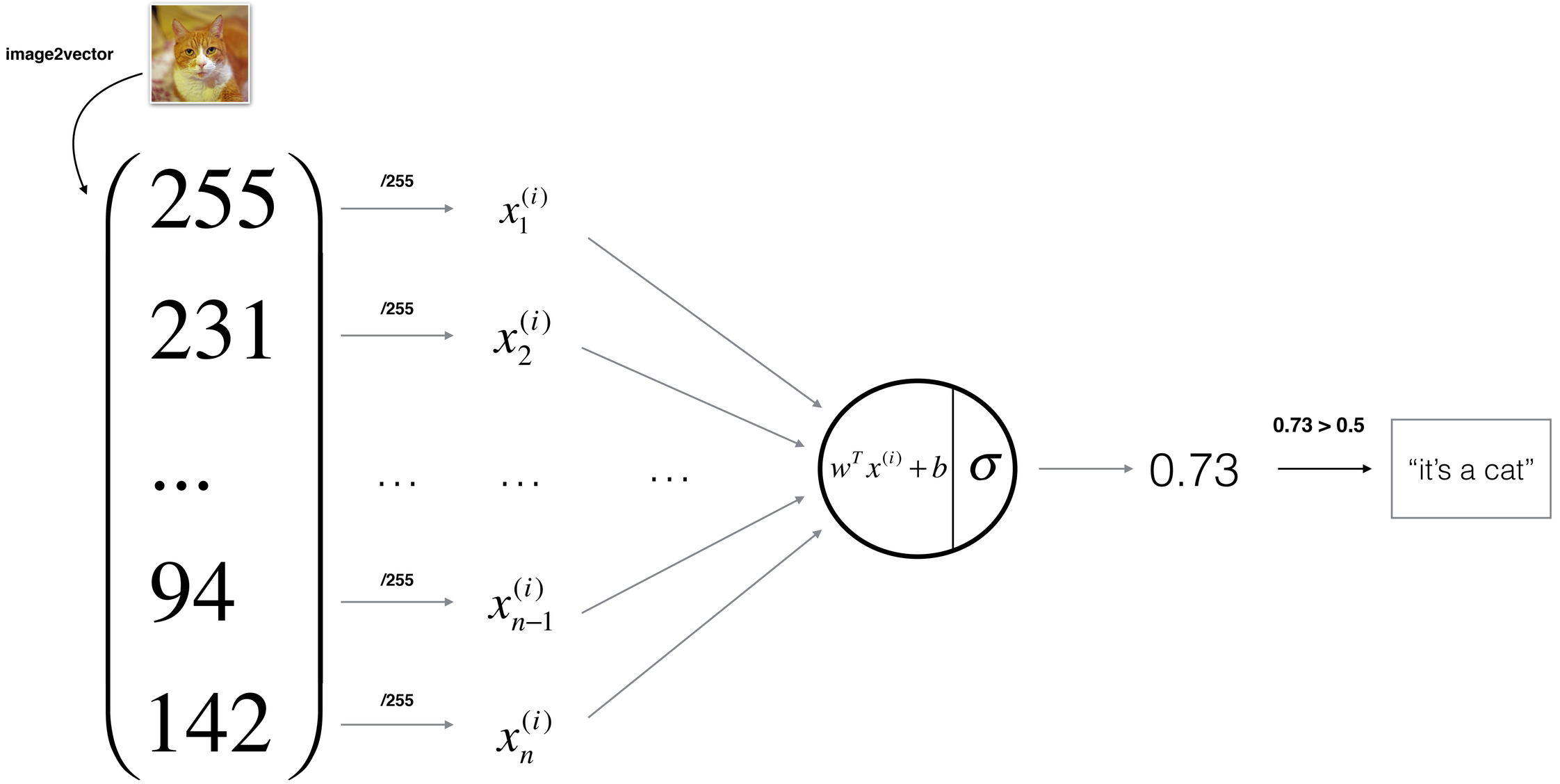


Things that can change

- Activation function
- Optimizer
- Hyperparameters
- …

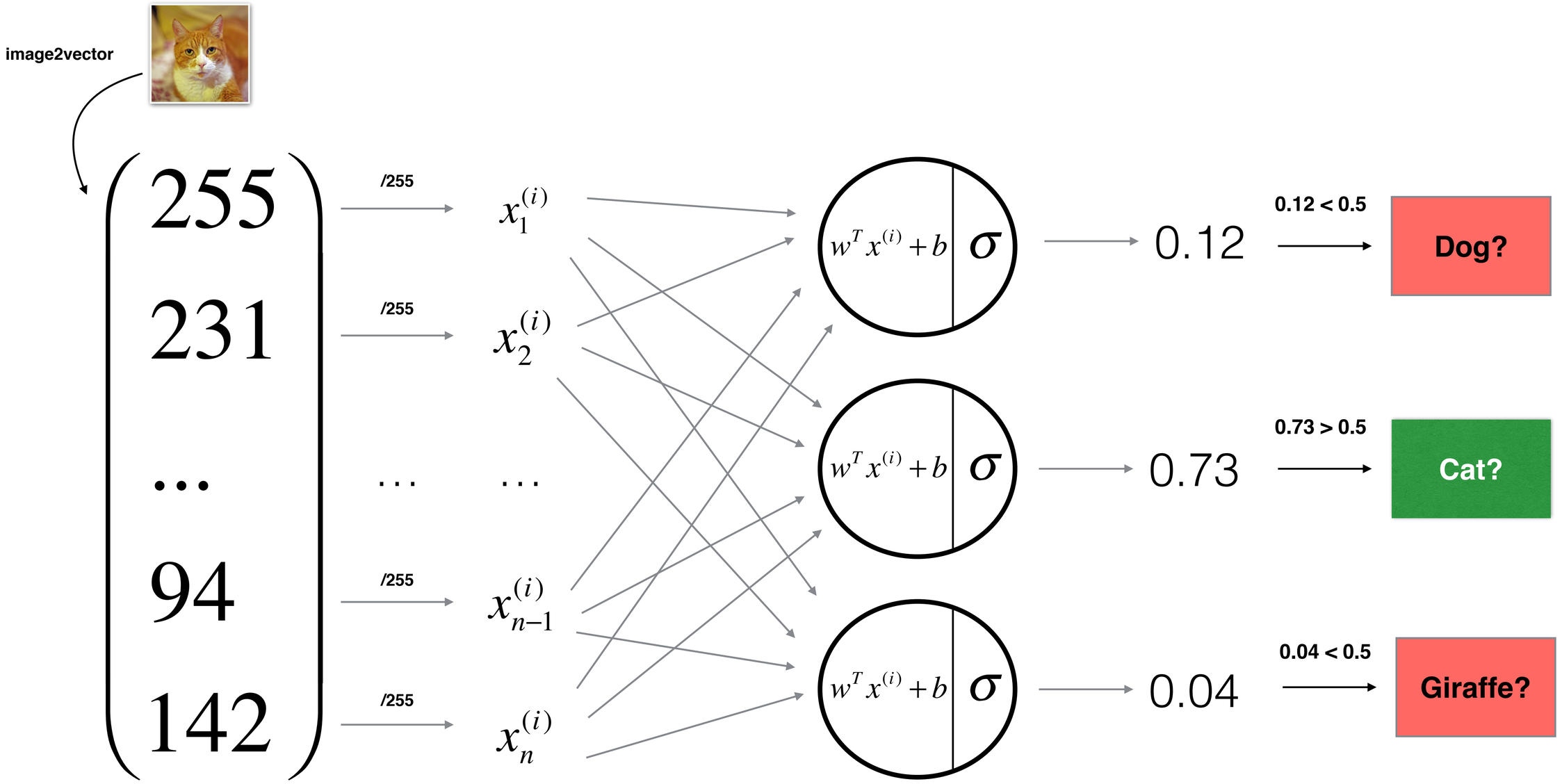
Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Logistic Regression as a Neural Network



Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Multi-class



Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Neural Network (Multi-class)

image2vector

$$\begin{pmatrix} 255 \\ 231 \\ \dots \\ 94 \\ 142 \end{pmatrix}$$

/255 → $x_1^{(i)}$

/255 → $x_2^{(i)}$

$\dots$ $\dots$

/255 → $x_{n-1}^{(i)}$

/255 → $x_n^{(i)}$

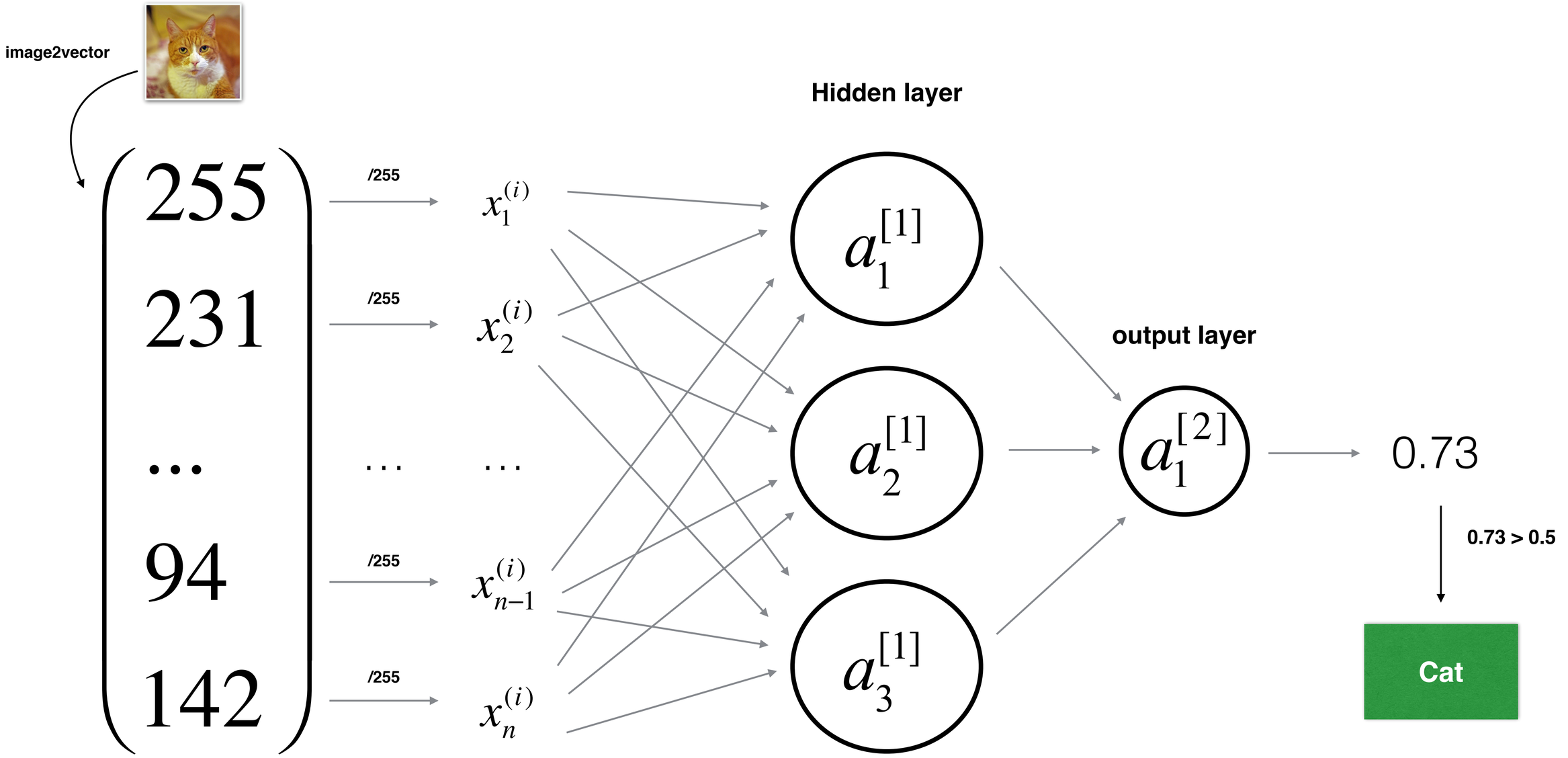
$w^T x^{(i)} + b \mid \sigma$

$w^T x^{(i)} + b \mid \sigma$

$w^T x^{(i)} + b \mid \sigma$

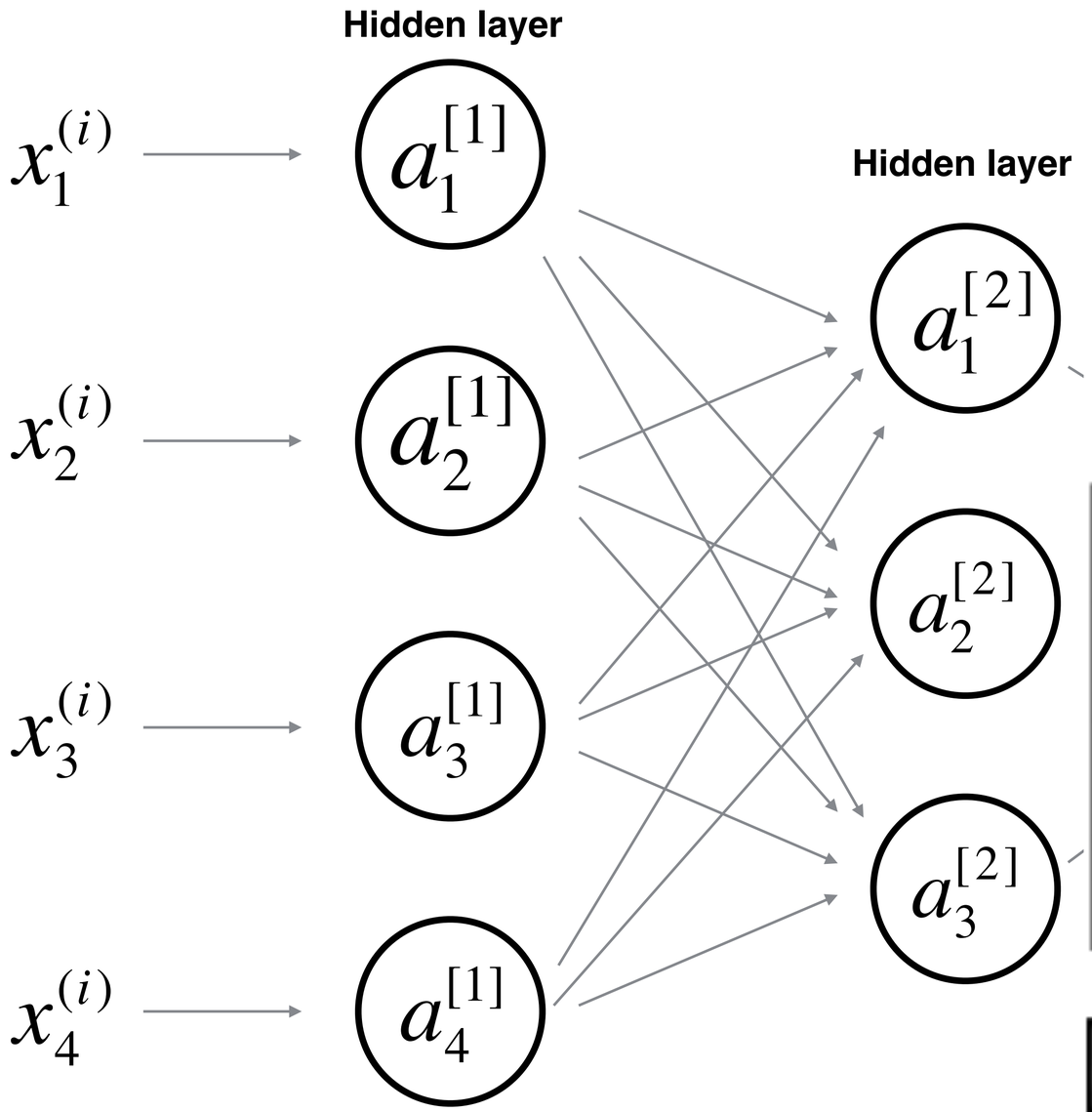
Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

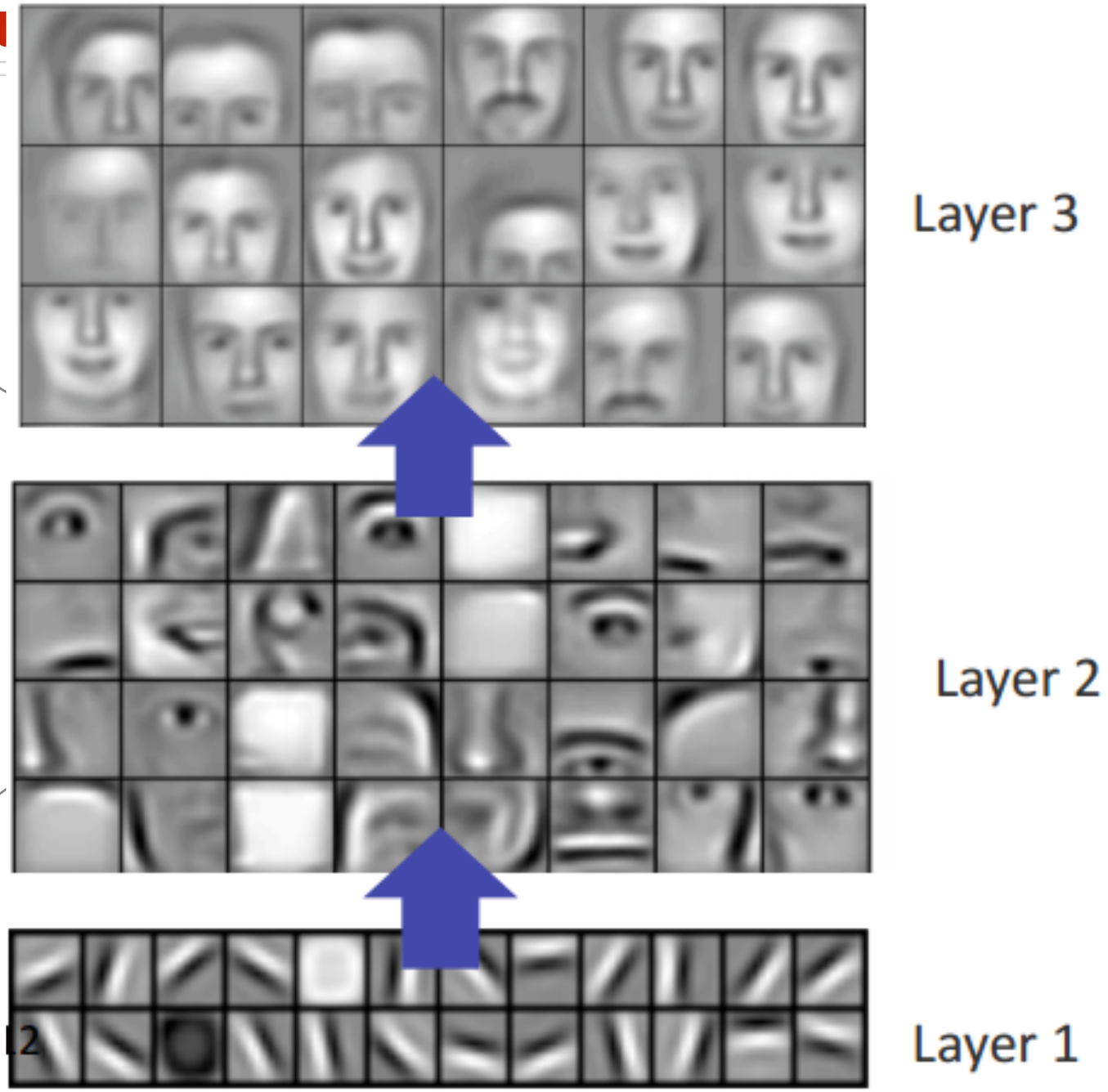
Neural Network (1 hidden layer)



Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Deeper net





Technique called “encoding”

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Let’s build intuition on concrete applications

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Today's outline

We will learn tips and tricks to:

- Analyze a problem from a deep learning approach
- Choose an **architecture**
- Choose a **loss** and a **training strategy**

I. Day'n'Night classification

II. Face verification and recognition

III. Neural style transfer (Art generation)

IV. Trigger-word detection

V. Shipping model

Day'n'Night classification

Goal: Given an image, classify as taken "during the day" (0) or "during the night" (1)

1. **Data?** 10,000 images Split? Bias?
2. **Input?**  Resolution? (64, 64, 3)
3. **Output?** y = 0 or y = 1 Last Activation? sigmoid
4. **Architecture ?** Shallow network should do the job pretty well
5. **Loss?** $$L = -[y \log(\hat{y}) + (1-y)\log(1-\hat{y})]$$

Easy warm up

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Face Verification

Goal: A school wants to use Face Verification for validating student IDs in facilities (dinning halls, gym, pool …)

1. Data?

Picture of every student labelled with their name

Bertrand

2. Input?

Resolution?

(412, 412, 3)

3. Output?

y = 1 (it's you)

or

y = 0 (it's not you)

Face Verification

Goal: A school wants to use Face Verification for validating student IDs in facilities (dinning halls, gym, pool …)

4. What architecture?

Simple solution:

database image

compute distance pixel per pixel

if less than threshold then y=1

input image

Issues:

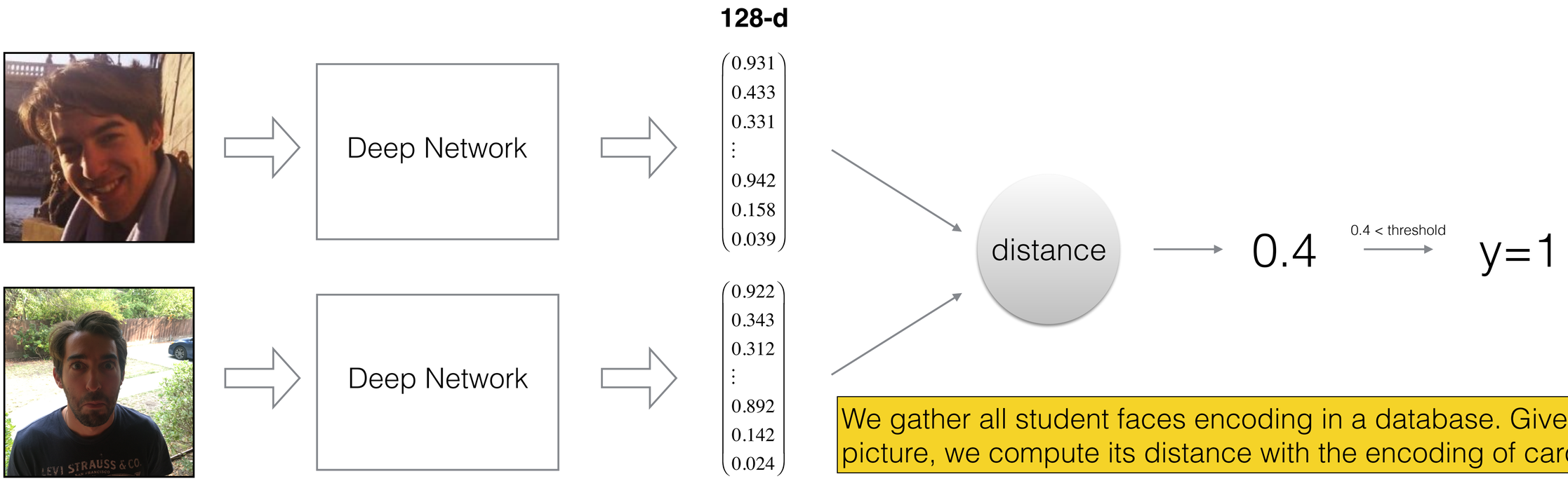
- Background lighting differences
- A person can wear make-up, grow a beard…
- ID photo can be outdated

Face Verification

Goal: A school wants to use Face Verification for validating student IDs in facilities (dinning halls, gym, pool …)

4. What architecture?

Our solution: encode information about a picture in a vector



We gather all student faces encoding in a database. Given a new picture, we compute its distance with the encoding of card holder

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Face Recognition

Goal: A school wants to use Face Verification for validating student IDs in facilities (dinning hall, gym, pool …)

4. Loss? Training?

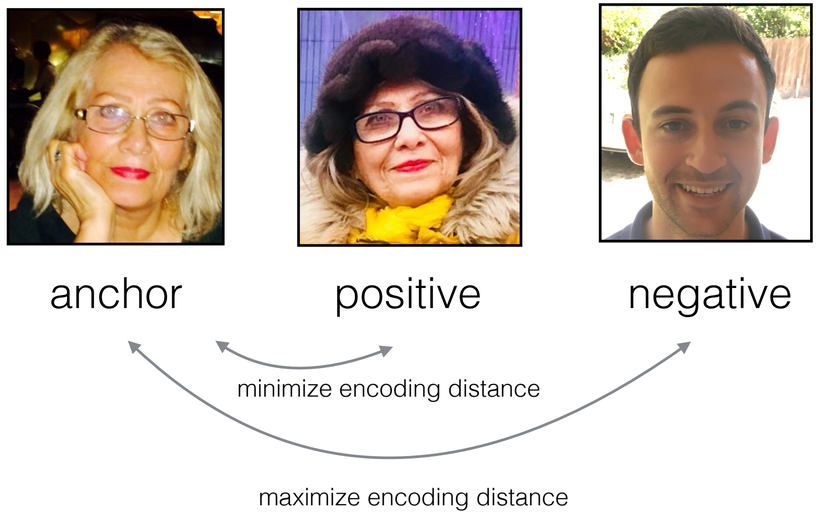
We need more data so that our model understands how to encode:
Use public face datasets

What we really want:

similar encoding

different encoding

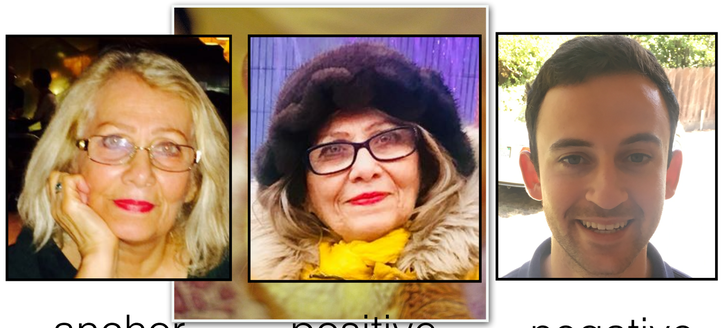
So let's generate triplets:



Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Recap: Learning Process

Input



anchor positive negative

Model
=
Architecture
+
Parameters

Output

$$\begin{pmatrix} 0.13 \\ 0.42 \\ .. \\ 0.10 \\ 0.31 \\ 0.73 \\ .. \\ 0.43 \\ 0.33 \end{pmatrix} \quad \begin{pmatrix} 0.01 \\ 0.54 \\ .. \\ 0.45 \\ 0.01 \\ 0.89 \\ .. \\ 0.12 \\ 0.01 \end{pmatrix} \quad \begin{pmatrix} 0.95 \\ 0.45 \\ .. \\ 0.20 \\ 0.41 \\ 0.89 \\ .. \\ 0.31 \\ 0.34 \end{pmatrix}$$

Enc(A) Enc(P) Enc(N)

Loss

$$L = \left\|Enc(A) - Enc(P)\right\|_2^2 - \left\|Enc(A) - Enc(N)\right\|_2^2 + \alpha$$

Gradients

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Face Recognition

Goal: A school wants to use Face Identification for recognize students in facilities (dinning hall, gym, pool …)

K-Nearest Neighbors

Goal: You want to use Face Clustering to group pictures of the same people on your smartphone

K-Means Algorithm

Maybe we need to detect the faces first?

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Art generation (Neural Style Transfer)

Goal: Given a picture, make it look beautiful

1. Data?

Let's say we have any data

2. Input?

content image

style image

3. Output?

generated image

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge: A Neural Algorithm of Artistic Style, 2015

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Art generation (Neural Style Transfer)

4. Architecture?

We want a model that **understands images** very well
We load an **existing model trained on ImageNet** for example

Deep Network

classification

When this image forward propagates, we can get information about its content & its style by inspecting the layers.

$Content_C$
$Style_S$

5. Loss?

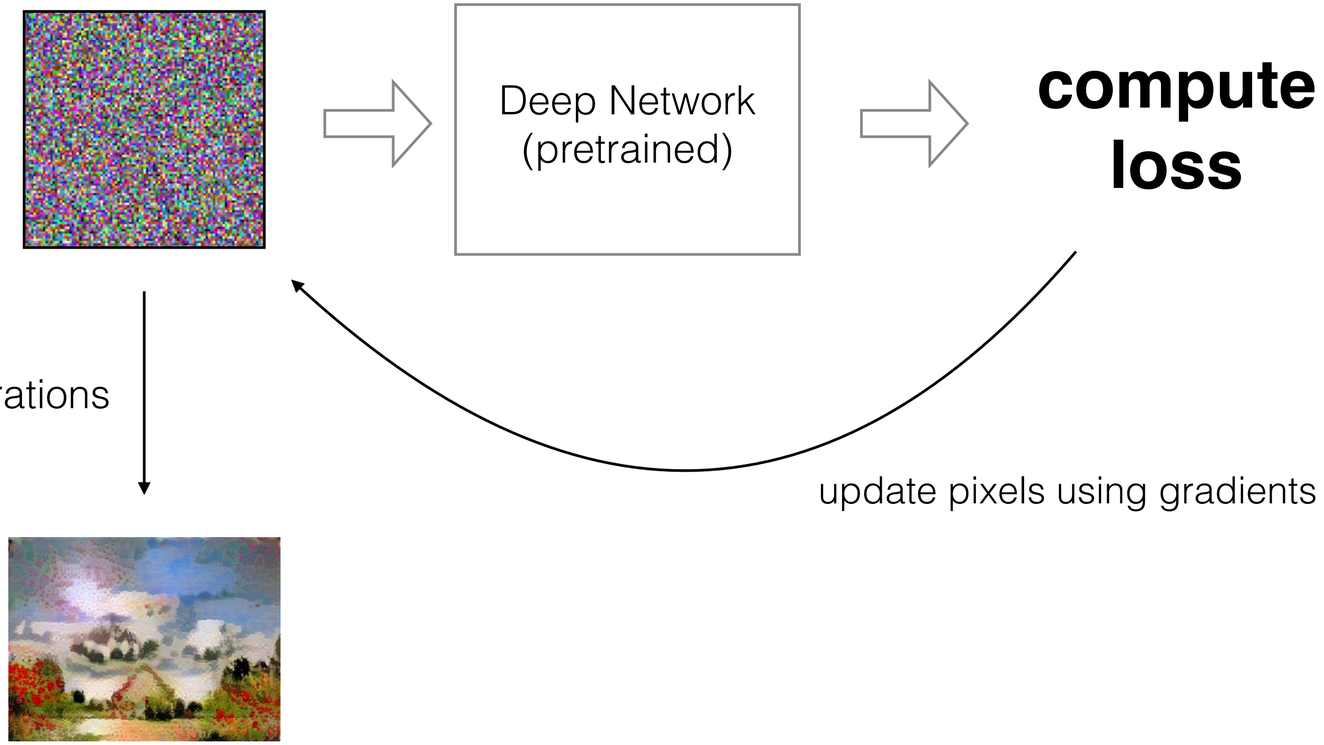
Leon A. Gatys, Alexander S. Ecker, Matthias Bethge: A Neural Algorithm of Artistic Style, 2015

Art generation (Neural Style Transfer)

Correct Approach

$$L = \left\| Content_C - Content_G \right\|_2^2 + \left\| Style_S - Style_G \right\|_2^2$$

We are not learning parameters by minimizing L. We are learning an image!



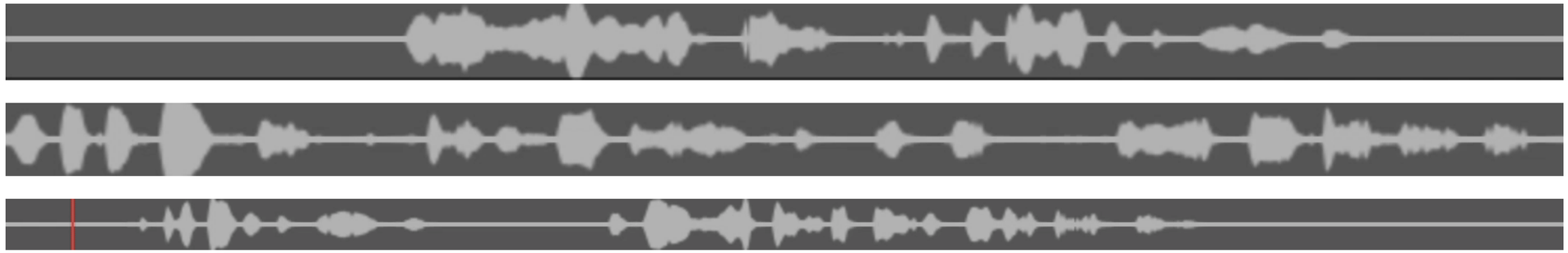
Leon A. Gatys, Alexander S. Ecker, Matthias Bethge: A Neural Algorithm of Artistic Style, 2015

Trigger word detection

Goal: Given a 10sec audio speech, detect the word “activate”.

1. **Data?** A bunch of 10s audio clips — Distribution?
2. **Input?** x = A 10sec audio clip  — Resolution? (sample rate)
3. **Output?** y = 0 or y = 1

Let’s have an experiment!



y = 1

y = 0

y = 1

y = 000000000000000000000000000000000000**1**0000000000

y = 00000000000000000000000000000000000000000000000000

y = 0000000000**1**0000000000000000000000000000000000000

Trigger word detection

Goal: Given a 10sec audio speech, detect the word “activate”.

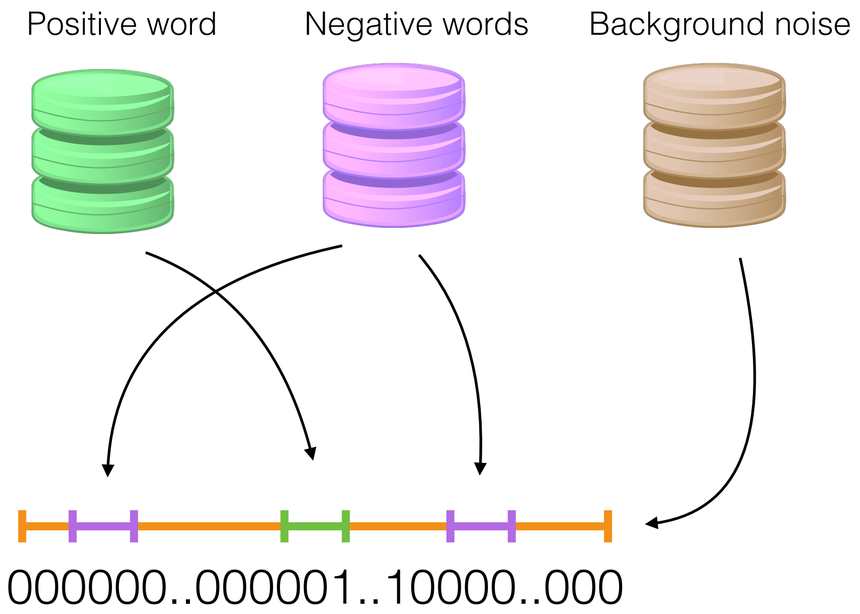
1. **Data?** A bunch of 10s audio clips — Distribution?
2. **Input?** x = A 10sec audio clip  — Resolution? (sample rate)
3. **Output?** y = 0 or y = 1
   y = 00..0000**1**00000..000
   y = 00..0000**1..1**000..000
   — Last Activation? sigmoid (sequential)
4. **Architecture ?** Sounds like it should be a RNN
5. **Loss?** $$L = -(y\log(\hat{y}) + (1-y)\log(1-\hat{y}))$$ (sequential)

Kian Katanforoosh

Trigger word detection

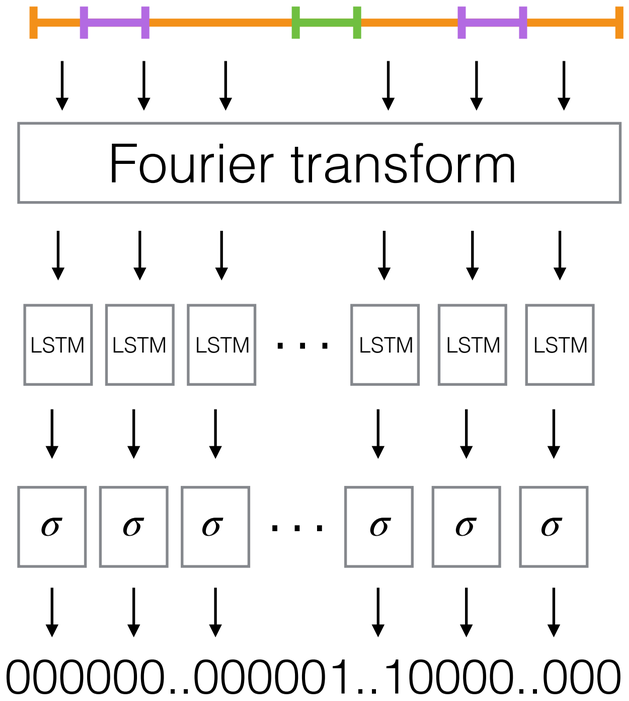
What is critical to the success of this project?

1. Strategic data collection/ labelling process

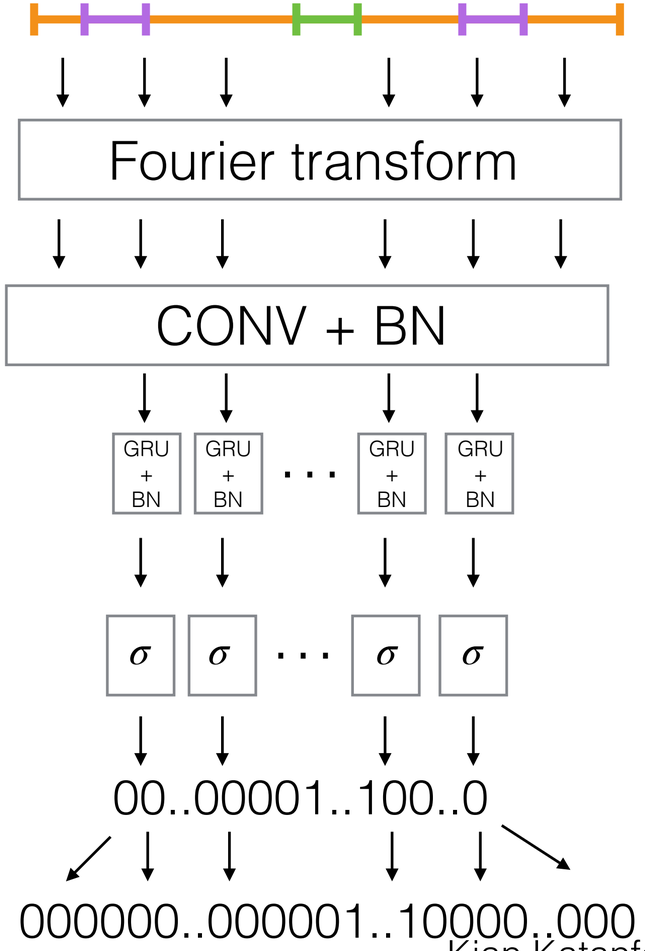


Automated labelling

2. Architecture search & Hyperparameter tuning



Never give up



+ Error analysis

Kian Katanforoosh

Another way of solving the TWD problem?

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Trigger word detection (other method)

Goal: Given an audio speech, detect the word “lion”.

$$L = \left\|Enc(A) - Enc(P)\right\|_2^2 - \left\|Enc(A) - Enc(N)\right\|_2^2 + \alpha$$

4. What architecture?

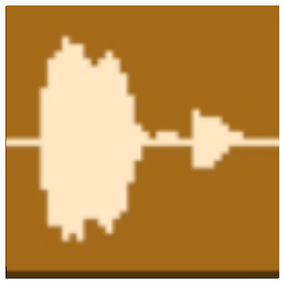
y = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,….,0,0,0,0,0,0,1,0,0, ….,0,0,0,0,0,0,0,0,0,0,0)

0.12 0.01 0.27 … … … … … … … 0.21 **0.92** 0.43 … … … … … …

Threshold: 0.6

Deep Network

$\begin{pmatrix} 0.931 \\ 0.433 \\ 0.331 \\ \vdots \\ 0.942 \\ 0.158 \\ 0.039 \end{pmatrix}$



Deep Network

$\begin{pmatrix} 0.922 \\ 0.343 \\ 0.312 \\ \vdots \\ 0.892 \\ 0.142 \\ 0.024 \end{pmatrix}$

distance → 0.4 → (0.4 < threshold) → y=1

[For more on query-by-example trigger word detection, check: Guoguo Chen et al.: Query-by-example keyword spotting using long short-term memory networks (2015)]

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Featured in the Magazine “the Most Beautiful Loss functions of 2015”

$$\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

$$+ \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2$$

$$+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2$$

$$+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left(p_i(c) - \hat{p}_i(c) \right)^2$$

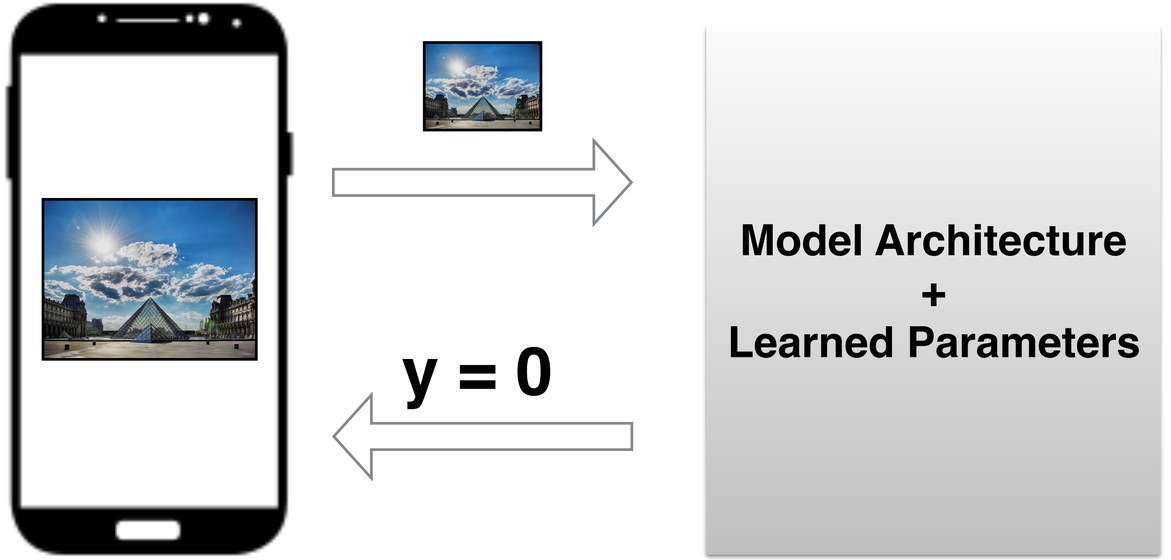
Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi: You Only Look Once: Unified, Real-Time Object Detection

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

App implementation

Server-based or on-device?

Server-based



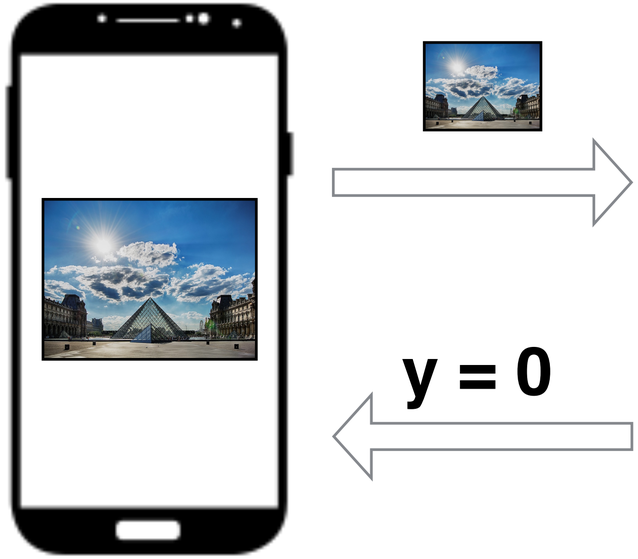
On-device

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Server-based or on-device?

Server-based

- App is light-weight
- App is easy to update



On-device

- Faster predictions
- Works offline

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri

Duties for next week

For Tuesday 29/01, 10am:

C1M3

- **Quiz: Shallow Neural Networks**
- **Programming Assignment: Planar data classification with one-hidden layer**

C1M4

- **Quiz: Deep Neural Networks**
- **Programming Assignment: Building a deep neural network - Step by Step**
- **Programming Assignment: Deep Neural Network Application**

Others:

- **TA project mentorship (mandatory this week)**
- **Friday TA section (01/25)**
- **Fill-in AWS Form to get GPU credits for your projects**