Lecture - 2

Deep Learning Intuition

1. Problem with one hot encoding?

If there are multiple objects in image, it will not work. We will need multihot encoding then.
The way you choose label is important & you should choose before starting anything.

2. Encoding:

Slower layer see pixel level information.
Deeper layer see more complex, high level features.
Thus every layer encodes.

3. Given an image classify as taken "during the day" (0) or "during the night" (1).

1. Data - How many images? Depends upon the complexity of task to learn. If need to classify more complex like dawn images, indoor-night images. You need higher

Around 10,000 sufficient.

Split? → 80% 20%.

Bias? → Correct balance b/w classes.

2. Input: Image. Resolution? → In terms of computation, higher pixels, higher parameters.

~~256 x 256~~ Compare with human accuracy

fixe (64,64,3) and also in terms of complexity of task for human.

3. Output: $y = 0$ or $y = 1$ Last activation? ↳ sigmoid.

4. Architecture. Shallow network should do the job pretty well.

5. Loss: $\mathcal{L} = -[y \log(\hat{y}) + (1-y) \log(1-\hat{y})]$

4. <u>Face Verification</u>: A school wants to use face verification for its facilities. i.e Gym, (for validating the ID)

i) Data? - picture of every student labelled with their name.

ii) Input - picture of person standing

Resolution - for (64,64) it's harder to decide the deep features like nose, eyes. etc.

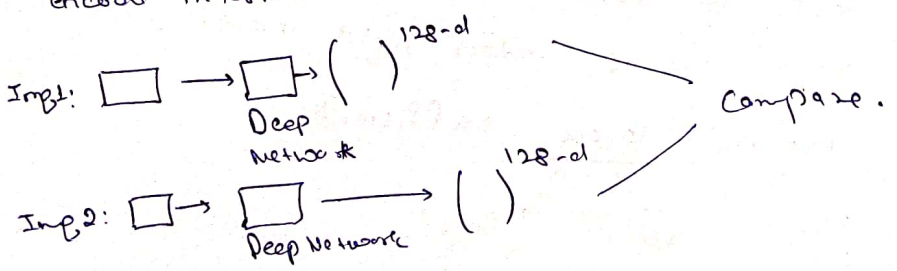
So try over (412, 412, 3)

iii) Output: $y = 1$ it's you

or $y = 0$ it's not you.

iv) Architecture:

encode information about a picture in a vector



we gather all student faces encoding in a database

Given a new image, we compute its distance with the encoding vector.

v) Loss? Training? - first we need more data so that algo model understands how to encode. Use public face datasets.

what we really want?

Similar faces
⇓
Similar encoding

Different faces
⇓
Different Encoding

Thus: Anchor Positive Negative.

Loss function:

$$L = \|Enc(A) - Enc(P)\|_2^2 - \|Enc(A) - Enc(N)\|_2^2$$

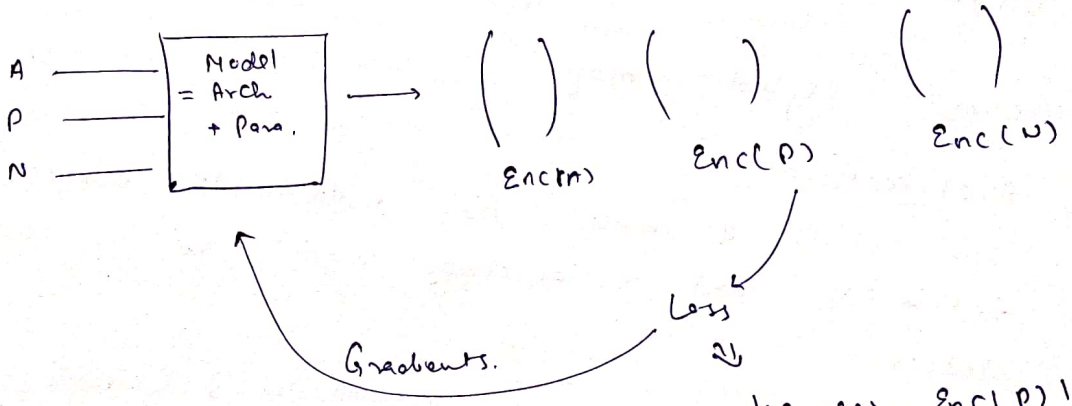
(A)

$$L = \|Enc(A) - Enc(N)\|^2 - \|Enc(A) - Enc(P)\|_2^2$$

(B)

$$L = \|Enc(P) - Enc(N)\|_2^2 - \|Enc(P) - Enc(A)\|^2$$

(C)



⇓

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

α ↓ margin to push the network to learn more.

5. <u>Face Recognition</u>:

Goal: A school wants to use face identification to recognise students.

$\boxed{\text{K-Nearest neighbours}}$ on encodings.

Goal: You want to use face clustering to group pictures of same people on your smartphone.

$\boxed{\text{K-Means Algorithm}}$ on encodings ~~from~~

6. <u>Art Generation</u>: Neural style transfer.

Goal: Given a picture, make it beautiful.

1. Data: Let's say we have 9y data
2. Input: content Image + style Image.
3. Output: styled Image.
4. Architecture:

We want a model that understands images very well. We load an existing model trained on imagenet.

$\boxed{\text{Image}} \longrightarrow \boxed{\text{Deep Network}} \longrightarrow$ Classification

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

5. <u>Loss</u>?

$$L = |Content_C - Content_G|_2^2 + |Style_S - Style_G|_2^2 \quad (A)$$

$$L = ||Content_C - Content_G||_2^2 + ||Style_S - Style_G||_2^2 \quad (B)$$

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

7. Trigger word detection.

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

Data: A bunch of 10s clips

Distribution?
↙
all types of noises, temp, accents

Input: x = A 10sec audio clip

Resolution? (sample rate).
↳ asks researcher.

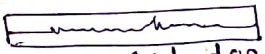
Output: $y = 0$ or 1

pomerejo
⇓
Afternoon in Italian

Output. — 0000000001 00000

Should word 1 come before or after?
— After. You need to hear to label.

Hacks: Add several ones after first 1.



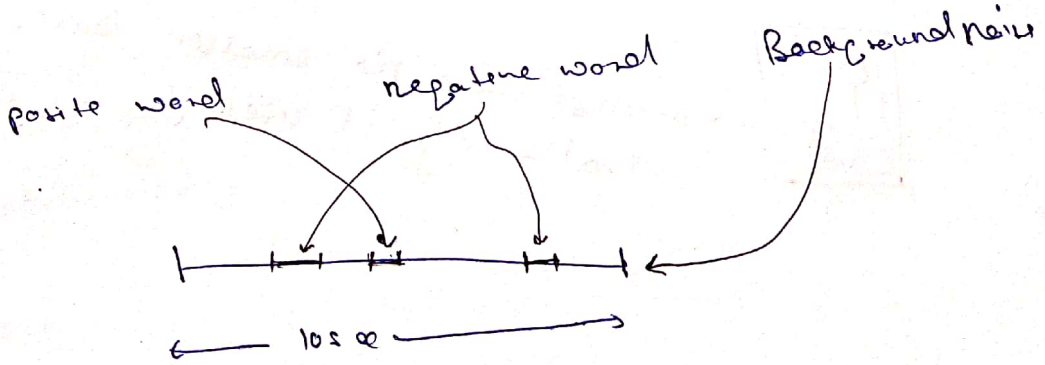
0000000 1 100

Last activation: Sigmoid (sequential)

Architecture: Sounds like it should be RNN.

Loss: $L = -(y \log \hat{y} + (1-y) \log(1-\hat{y}))$

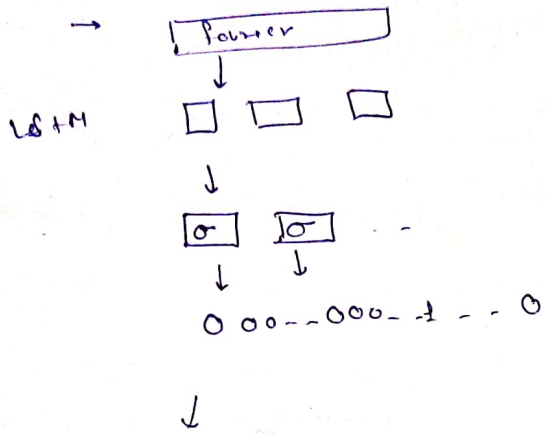
What is critical to the success of this project?
i) Strategic data collection | labelling process.

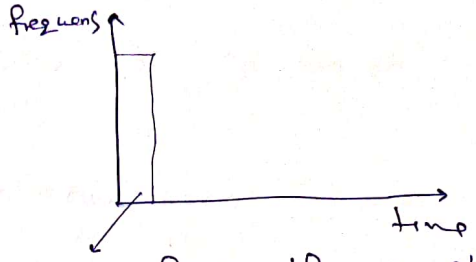


By this way programmatic labelling & generation.

ii) Architecture Search & Hyperparameter tuning.

→ Started with Fourier transform.

→ 

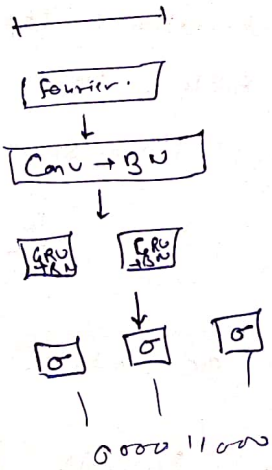


value of all the amplitudes of this frequency for every time step.
This is spectogram.

But it doesn't work.

So an expert friend guided & finally.

Find an expert for your problem



The most beautiful loss functions of 2015.

↓

YOLO. Bounding boxes.

→ penalize more the smaller boxes.
That's why it used $\sqrt{w_i}$, $\sqrt{h_i}$