# What is face recognition?

Face Verification vs Face Recognition:

- Verification:
  - Input image, name/ID
  - Output weather the input image is that of the claimed person
- Recognition
  - Has a database of K persons
  - Get an input image
  - Output ID if the image is any of the K persons

# One Shot Learning

In the one shot problem you need to train your algorithm in a single example (from a single person) to classify it.

For this you want to learn a “similarity” function. It works as $d(i_1, i_2)$ and for a given $\theta$ threshold it decides if both inputs are the same person or not.

# Siamese Network

Given an encoding of a given input the siamese network takes multiple inputs and perform further calculations to express how close the inputs are.

Paper: https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf

# Triplet loss

Given 3 input images: Anchor (A), Positive (P) and Negative (N) we want to train our network using the following loss function:

1. $L(A, P, N) = ||f(A) - f(P)||^2 \leq ||f(A) - f(N)||^2$
2. $L(A, P, N) = d(A, P) \leq d(A, N)$
3. $L(A, P, N) = ||f(A) - f(P)||^2 - ||f(A) - f(N)||^2 + \alpha \leq 0$

In plain text what does it means? It’s trying to minimize the encoding distance of the Anchor and Positive images ($d(A, P)$) at the same time is trying to maximize the distance from the Anchor image to the Negative image ($d(A, N)$).

Given that one of the easiest ways for the network to obtain low loss would be mapping all the inputs to $\vec{0}$ or to map every input to the same output, causing all the squared differences to turn to 0. To avoid that we can adjust the equation adding the $\alpha$ in the equation 3 which is a hyperparameter that prevents the network from learning the trivial solutions. The $\alpha$ is called *margin* which is a similar terminology as on SVMs literature. The $\alpha$ dictates how far away the distances are being determined. Higher alphas means that the network will try to map and perform the subtractions as big as alpha whereas small alphas allow the mappings to be as closer as desired.

One of the things we must take care is choosing "hard" triplets to train the network on. Is easy to realize that choosing randomly any triplet (A, P, N) could satisfy the loss equation. But to ensure the network is focusing on learn we must pick the most similar distances ($d(A,P) \approx d(A,N)$). To do this we can implement a generator that decides if a given triplet should keep being fed to the network or not. Given an initial set of triplets combinations we can keep track of the ones that already perform well and not feeding those to the network anymore, just the ones that don't perform well.

Paper: https://arxiv.org/pdf/1503.03832

# Face Verification and Binary Classification

To address the same problem one could use both encodings of a given image pair and perform logistic regression with binary output to classify if it's the same person.

To the given inputs there are some available options:

- Use both inputs as features. So if the encoding has length n the input for the logistic regression has size 2n
- Use the element-wise difference between the given inputs defined as $\hat{y} = \sigma(\sum_{k=1}^{n} |f(x_k^{(i)}) - f(x_k^{(j)})| + b)$
- Use the squared element-wise difference between the given inputs defined as $\hat{y} = \sigma(\sum_{k=1}^{n} (f(x_k^{(i)}) - f(x_k^{(j)}))^2 + b)$ also referred as chi-squared ($\chi^2$) difference

# What is neural style transfer?

Given an image with content and another image with a style you want to transfer the style to the content.

# What are deep ConvNets learning?

Given an specific filter in a convnet to understand what it's trying to learn you need to pick a set of the patches that activated it the most. Given a small set of activated neurons you can grasp better on what that specific neuron is doing.

Paper: https://arxiv.org/pdf/1503.03832

# Cost Function

Given an image with the content C and an image with the style S we want to generate image G which keeps the content C but with the style S. The loss function is defined as $L(G) = \alpha L_{content}(C,G) + \beta L_{style}(S,G)$

1. Initialize G randomly. $G.shape = 100 \times 100 \times 3$, for ex
2. Use gradient descent to minimize $L(G)$. Where $G := G - \frac{\partial}{\partial G} L(G)$

Paper: https://arxiv.org/abs/1508.06576

# Content Cost Function

If we use the hidden layer l to compute the content cost given a pre-trained ConvNet (e.g. VGG Network) we would have $a^{[l](C)}$ and $a^{[l](G)}$ to be the activation layer l on the images C and G, respectively. If $a^{[l](C)}$ and $a^{[l](G)}$ are similar, both images have similar content. One possible given function to compute this can be given by $L(C,G) = \frac{1}{2}||a^{[l](C)} - a^{[l](G)}||^2$

# Style Cost Function

If we use the hidden layer l to measure the style. We would like to define style as the correlation between activations across channels in the computed image G and the style reference S.

1. Let $a^{[l]}_{i,j,k}$ be the activation at the $(i,j,k)$. G is $n_c^{[l]} \times n_c^{[l]}$
2. $G^{[l](S)}_{kk'} = \sum_{i=1}^{n_H^{[l]}} \sum_{j=1}^{n_W^{[l]}} a^{[l](S)}_{i,j,k} a^{[l](S)}_{i,j,k'}$ if the activations are correlated then the internal multiplication tends to be high. The correlation being calculated here is in reality the unnormalized cross of the areas.
3. $G^{[l](G)}_{kk'} = \sum_{i=1}^{n_H^{[l]}} \sum_{j=1}^{n_W^{[l]}} a^{[l](G)}_{i,j,k} a^{[l](G)}_{i,j,k'}$
4. $L_{style}(S,G) = ||G^{[l](S)} - G^{[l](G)}||^2_F$
5. $L^{[l]}_{style}(S,G) = \frac{1}{(2n_H^{[l]} n_W^{[l]} n_C^{[l]})^2} \sum_k \sum_{k'} (G^{[l](S)}_{kk'} - G^{[l](G)}_{kk'})^2$
6. You can get better results using activations from multiple layers, therefore defining the loss as $L_{style}(S,G) = \sum_l \lambda^{[l]} L^{[l]}_{style}(S,G)$

# 1D and 3D Generalizations

1D Examples: EKG Info 3D Examples: Whole Body Scan

# Quiz

1.Face verification requires comparing a new picture against one person's face, whereas face recognition requires comparing a new picture against K person's faces.

- [x] True
- [ ] False

2.Why do we learn a function d(img1, img2) for face verification? (Select all that apply.)

- [ ] Given how few images we have per person, we need to apply transfer learning.
- [x] We need to solve a one-shot learning problem.
- [x] This allows us to learn to recognize a new person given just a single image of that person.
- [ ] This allows us to learn to predict a person's identity using a softmax output unit, where the number of classes equals the number of persons in the database plus 1 (for the final "not in database" class).

3.In order to train the parameters of a face recognition system, it would be reasonable to use a training set comprising 100,000 pictures of 100,000 different persons.

- [ ] True
- [x] False

4.Which of the following is a correct definition of the triplet loss? Consider that $\alpha > 0$. (We encourage you to figure out the answer from first principles, rather than just refer to the lecture.)

- ☐ $\max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 - \alpha, 0)$
- ☐ $\max(\|f(A) - f(N)\|^2 - \|f(A) - f(P)\|^2 - \alpha, 0)$
- ☑ $\max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0)$
- ☐ $\max(\|f(A) - f(N)\|^2 - \|f(A) - f(P)\|^2 + \alpha, 0)$

5.Consider the following Siamese network architecture: The upper and lower neural networks have different input images, but have exactly the same parameters.

- ☑ True
- ☐ False

6.You train a ConvNet on a dataset with 100 different classes. You wonder if you can find a hidden unit which responds strongly to pictures of cats. (I.e., a neuron so that, of all the input/training images that strongly activate that neuron, the majority are cat pictures.) You are more likely to find this unit in layer 4 of the network than in layer 1.

- ☑ True
- ☐ False

7.Neural style transfer is trained as a supervised learning task in which the goal is to input two images (x), and train a network to output a new, synthesized image (y).

- ☐ True
- ☑ False

8.In the deeper layers of a ConvNet, each channel corresponds to a different feature detector. The style matrix $G^{[l]}$ measures the degree to which the activations of different feature detectors in layer l vary (or correlate) together with each other.

- ☑ True
- ☐ False

9.In neural style transfer, what is updated in each iteration of the optimization algorithm?

- ☐ The regularization parameters
- ☐ The neural network parameters
- ☑ The pixel values of the generated image G
- ☐ The pixel values of the content image C

10.You are working with 3D data. You are building a network layer whose input volume has size 32x32x32x16 (this volume has 16 channels), and applies convolutions with 32 filters of dimension 3x3x3 (no padding, stride 1). What is the resulting output volume?

- ☐ Undefined: This convolution step is impossible and cannot be performed because the dimensions specified don't match up.
- ☐ 30x30x30x16
- ☑ 30x30x30x32