Large Scale Landmark Recognition via Triplet Network

Jifu Zhao

Statistics @ UIUC

04/26/2018, Urbana

Object Recognition

- A well developed area in computer vision
- With deep learning, machine can beat human

Object Recognition

- A well developed area in computer vision
- With deep learning, machine can beat human

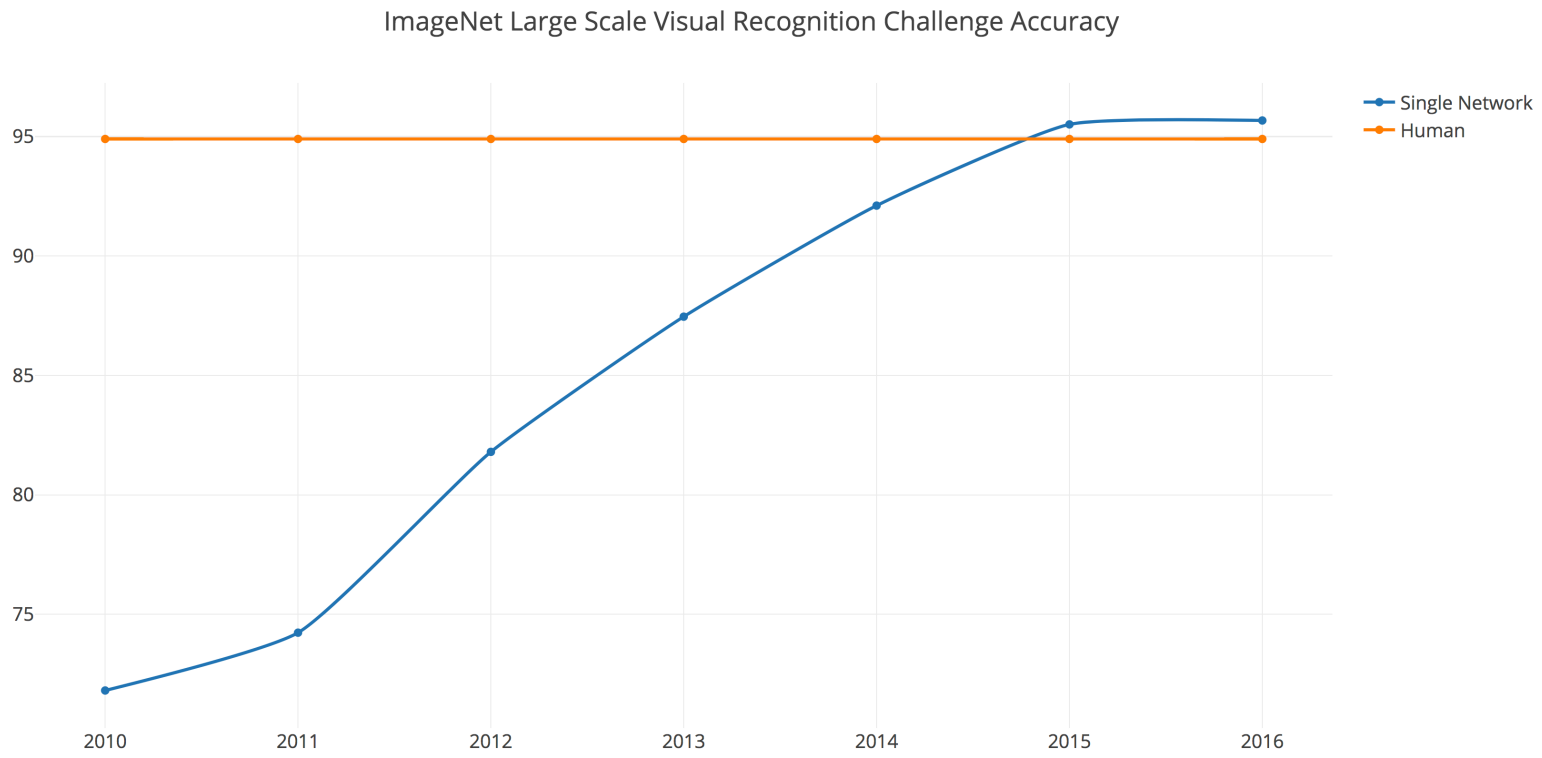


Image: https://plot.ly/~botevmg/5.embed

Object Recognition

- A well developed area in computer vision
- With deep learning, machine can beat human

- Large amount of training images are required
- Example
  - CIFAR-10: 60,000 images in 10 classes.
  - ImageNet: 14,197,122 images in 1,000 classes.

Object Recognition

- A well developed area in computer vision
- With deep learning, machine can beat human

- Large amount of training images are required
- Example
  - CIFAR-10: 60,000 images in 10 classes.
  - ImageNet: 14,197,122 images in 1,000 classes.

- What if limited amount of images are available?
  - One-Shot Learning

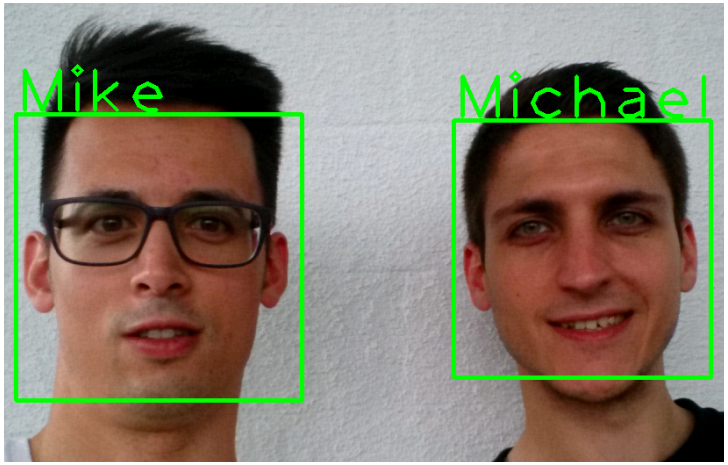
One-Shot Learning

- **One-shot learning**
  - An object categorization problem in computer vision. Whereas most machine learning based object categorization algorithms require training on hundreds or thousands of images and very large datasets, one-shot learning aims to learn information about object categories from one, or only a few, training images.

- Characteristics:
  - A few training images for each class
  - Potentially large amount of classes

One-Shot Learning

- Characteristics:
  - A few training images for each class
  - Potentially large amount of classes
- Applications
  - Face/Item Verification
  - Street-to-Shop



Landmark Recognition

- Given a photo, can we recognize the correct landmarks it contains?

Landmark Recognition

- Given a photo, can we recognize the correct landmarks it contains?

Reference image 1

?

?

Test image

Reference image 2

?

?

Reference image 3

Reference image 4

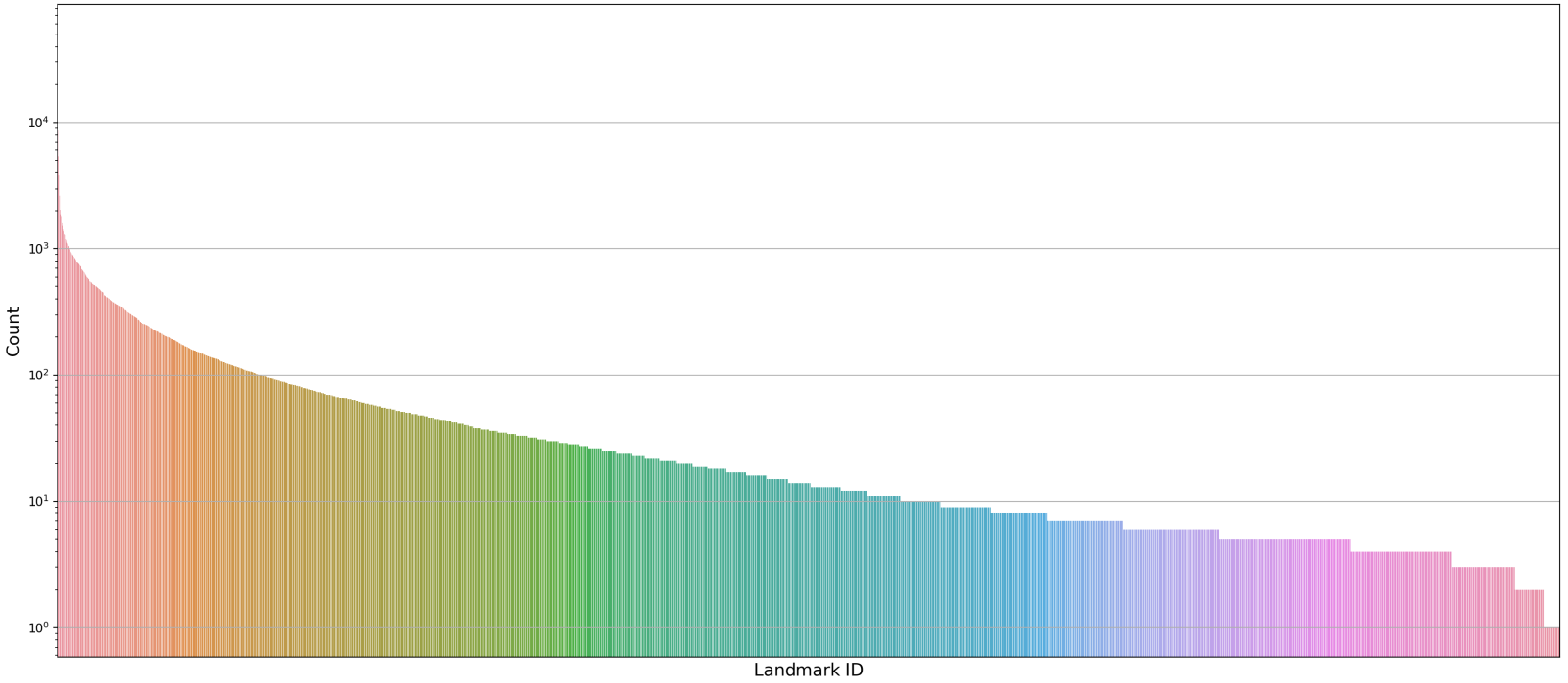
Data

- On March 01 2018, Google released a new dataset for landmark recognition
  - 1,225,029 training images with 14,951 landmarks
  - 117,703 test images

Data

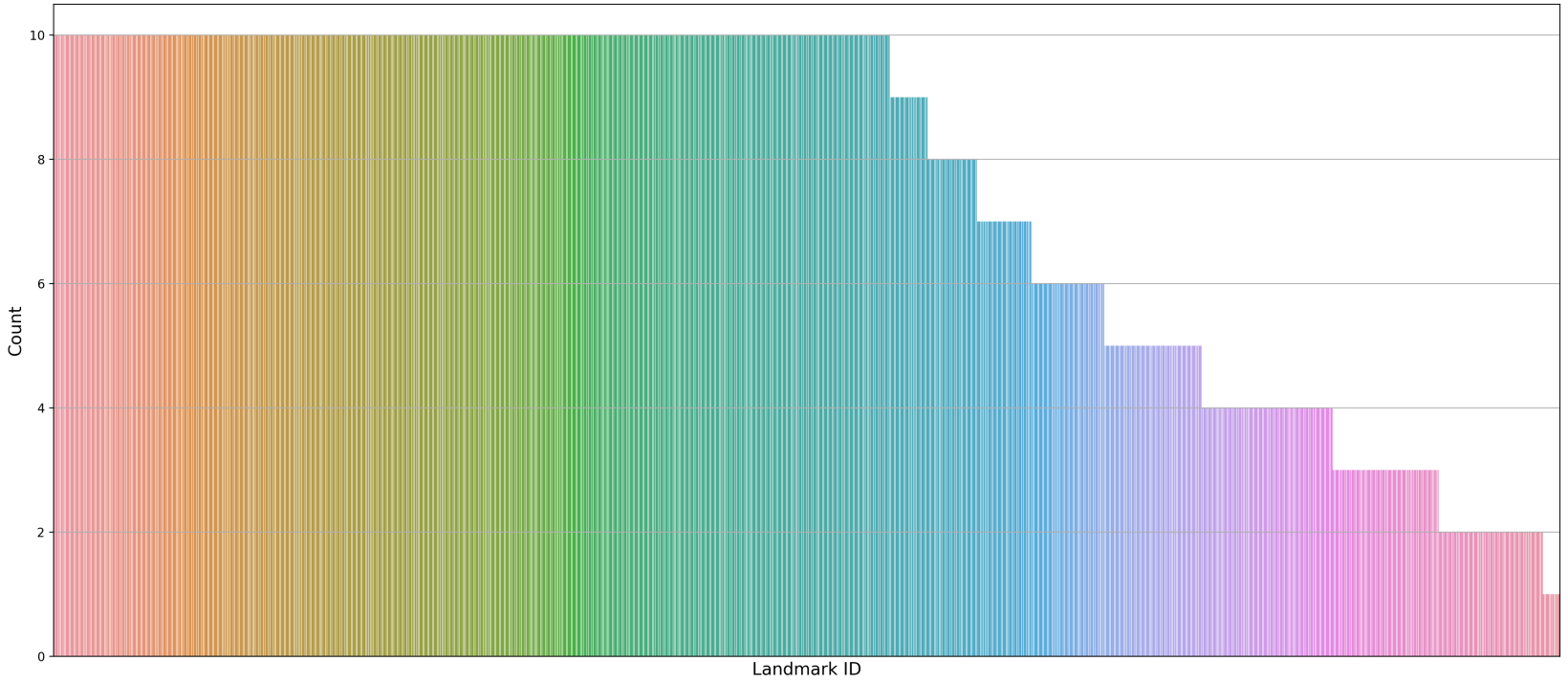
Data

- On March 01 2018, Google release a new dataset for landmark recognition
  - 1,225,029 training images with 14,951 landmarks
  - 117,703 test images



Data

- In our case
  - 113,783 training images with 14,943 different landmarks
  - 22,255 validation images with 7,675 different landmarks
  - 22,391 test images with 14,436 different landmarks



Methodology

- Idea 1: Random Guess
  - Extremely low accuracy

Methodology

- Idea 1: Random Guess
  - Extremely low accuracy
- Idea 2: Classical CNN for Multiclass Classification

Methodology

- Idea 1: Random Guess
  - Extremely low accuracy
- Idea 2: Classical CNN for Multiclass Classification

Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.715	0.901	138,357,544	23
VGG19	549 MB	0.727	0.910	143,667,240	26
ResNet50	99 MB	0.759	0.929	25,636,712	168
InceptionV3	92 MB	0.788	0.944	23,851,784	159
InceptionResNetV2	215 MB	0.804	0.953	55,873,736	572
MobileNet	17 MB	0.665	0.871	4,253,864	88
DenseNet121	33 MB	0.745	0.918	8,062,504	121
DenseNet169	57 MB	0.759	0.928	14,307,880	169
DenseNet201	80 MB	0.770	0.933	20,242,984	201

Methodology

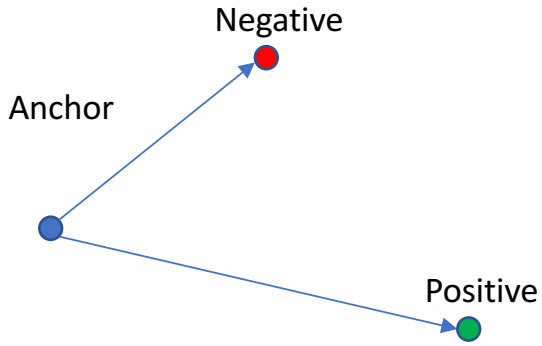
- Idea 1: Random Guess
  - Extremely low accuracy
- Idea 2: Classical CNN for Multiclass Classification
  - Large amount of landmarks: 14,951 different classes
  - Imbalanced data: Some landmarks only have 1 training images

Methodology

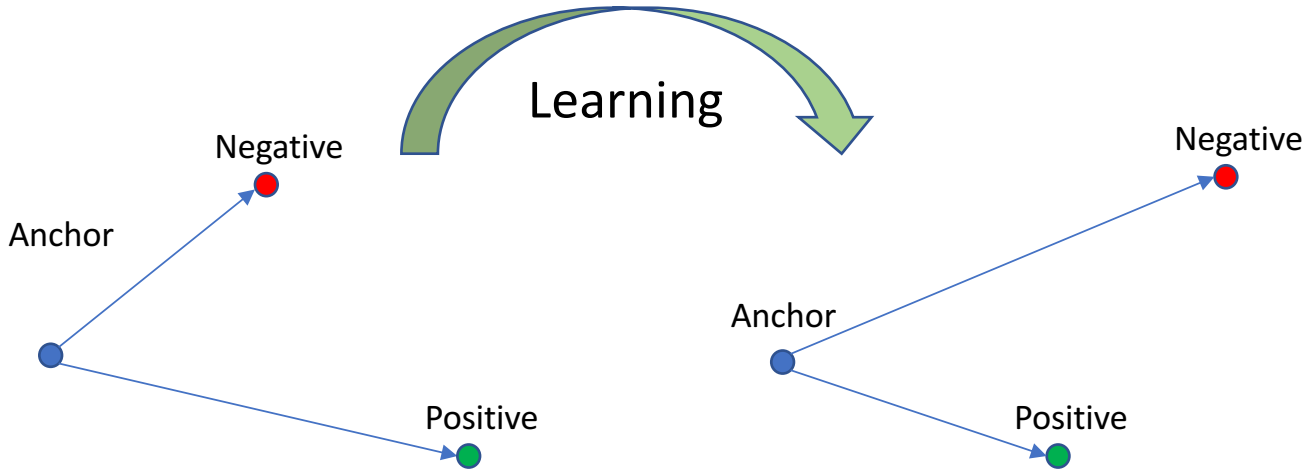
- Idea 1: Random Guess
  - Extremely low accuracy
- Idea 2: Classical CNN for Multiclass Classification
  - Large amount of landmarks: 14,951 different classes
  - Imbalanced data: Some landmarks only have 1 training images

- What can we do?
  - Try to learn a metric for similarity/distance

Goal



Goal


Learning
Negative
Anchor
Positive
Negative
Anchor
Positive

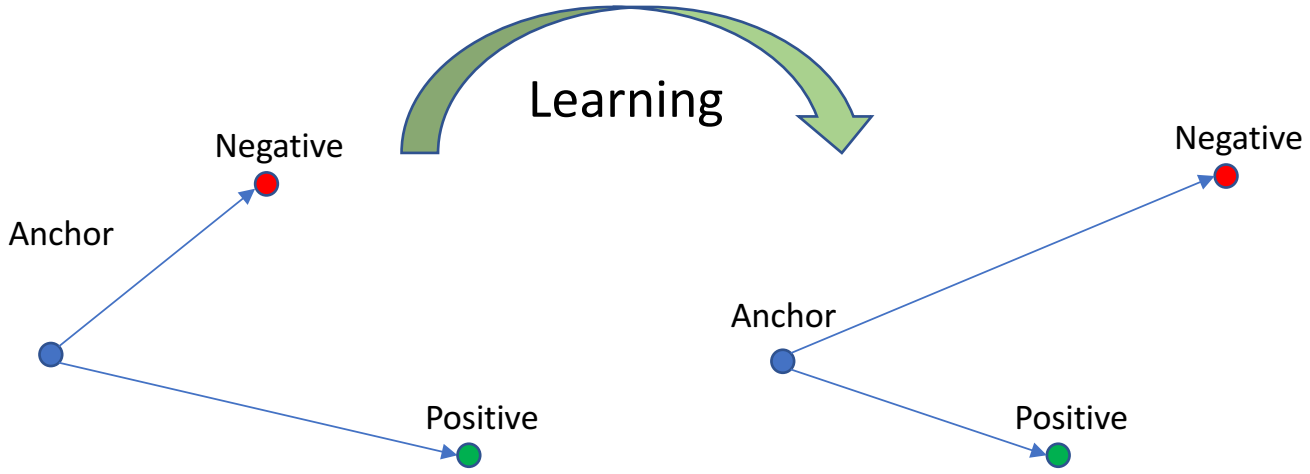
Triplet Loss

- Goal:
  - Learn new representation $f(\cdot)$ of the original image such that

$$\|f(a) - f(p)\|_2^2 \leq \|f(a) - f(n)\|_2^2$$

  - Add some margin α

$$\|f(a) - f(p)\|_2^2 + \alpha \leq \|f(a) - f(n)\|_2^2$$



Triplet Loss

- Goal:
  - Learn new representation $f(\cdot)$ of the original image such that

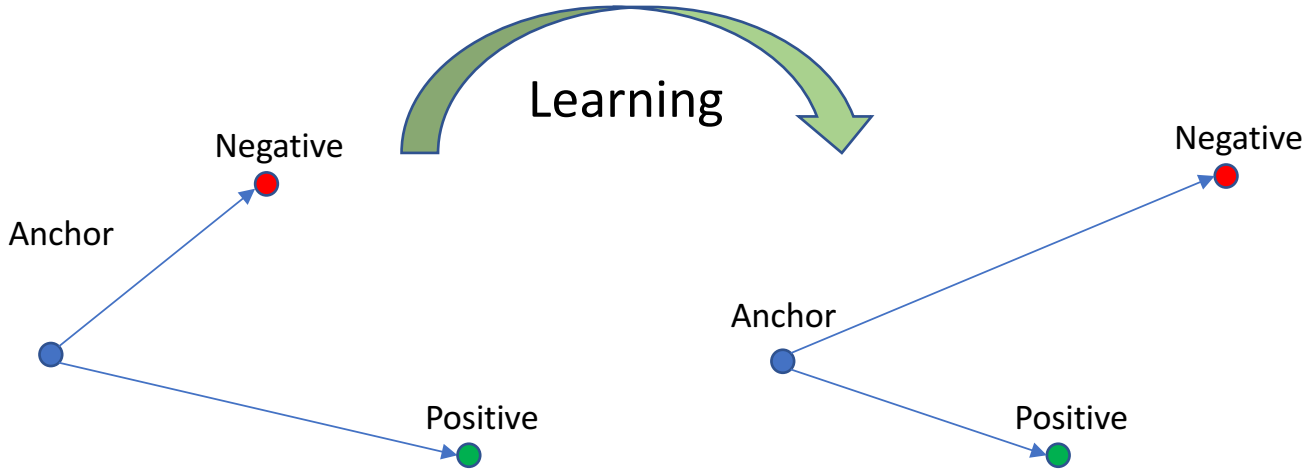
$$\|f(a) - f(p)\|_2^2 \leq \|f(a) - f(n)\|_2^2$$

  - Add some margin α

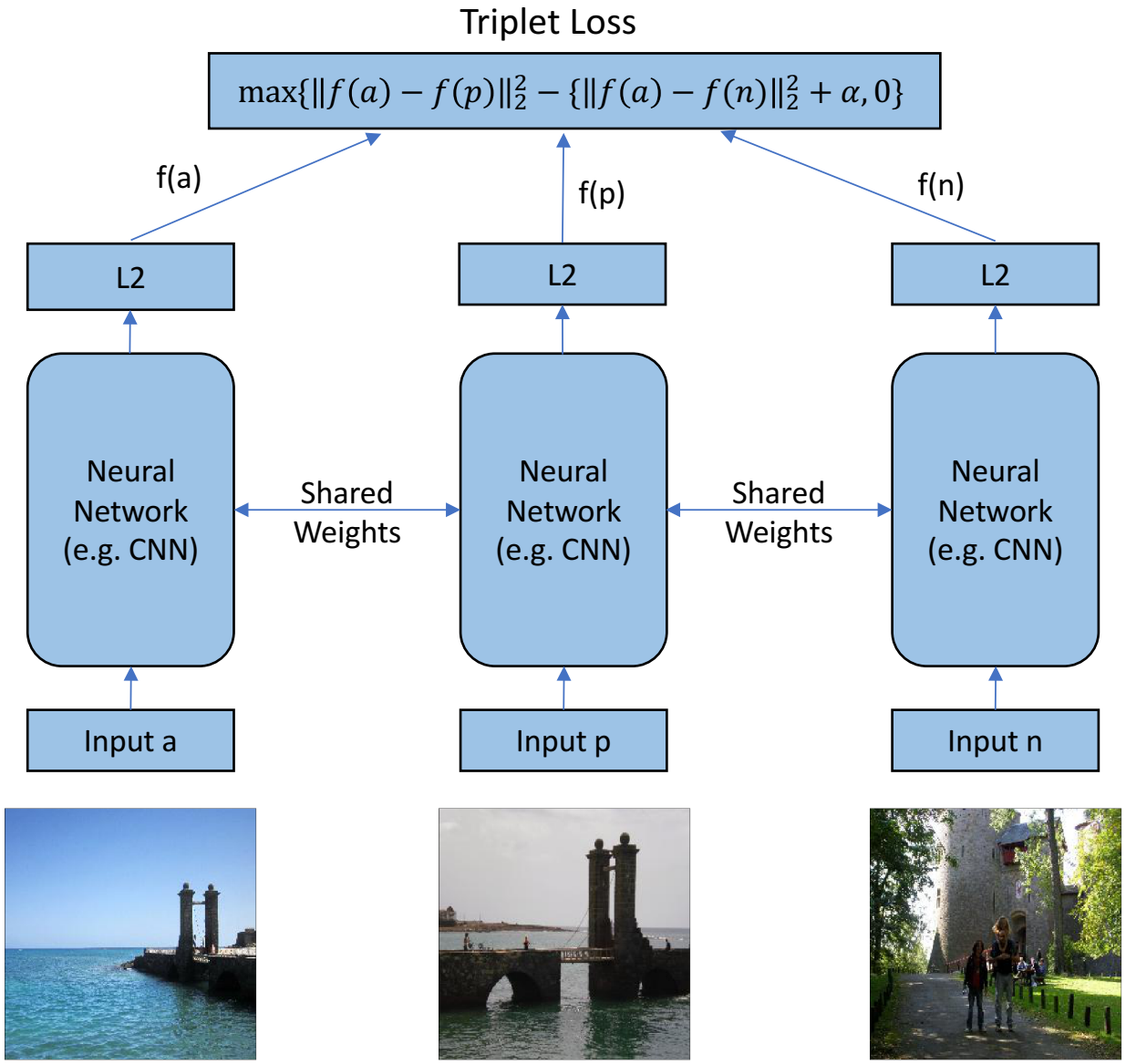
$$\|f(a) - f(p)\|_2^2 + \alpha \leq \|f(a) - f(n)\|_2^2$$

- Triplet Loss:

$$L(a, p, n) = \max\{\|f(a) - f(p)\|_2^2 + \alpha - \|f(a) - f(n)\|_2^2, 0\}$$



Triplet Network



Training

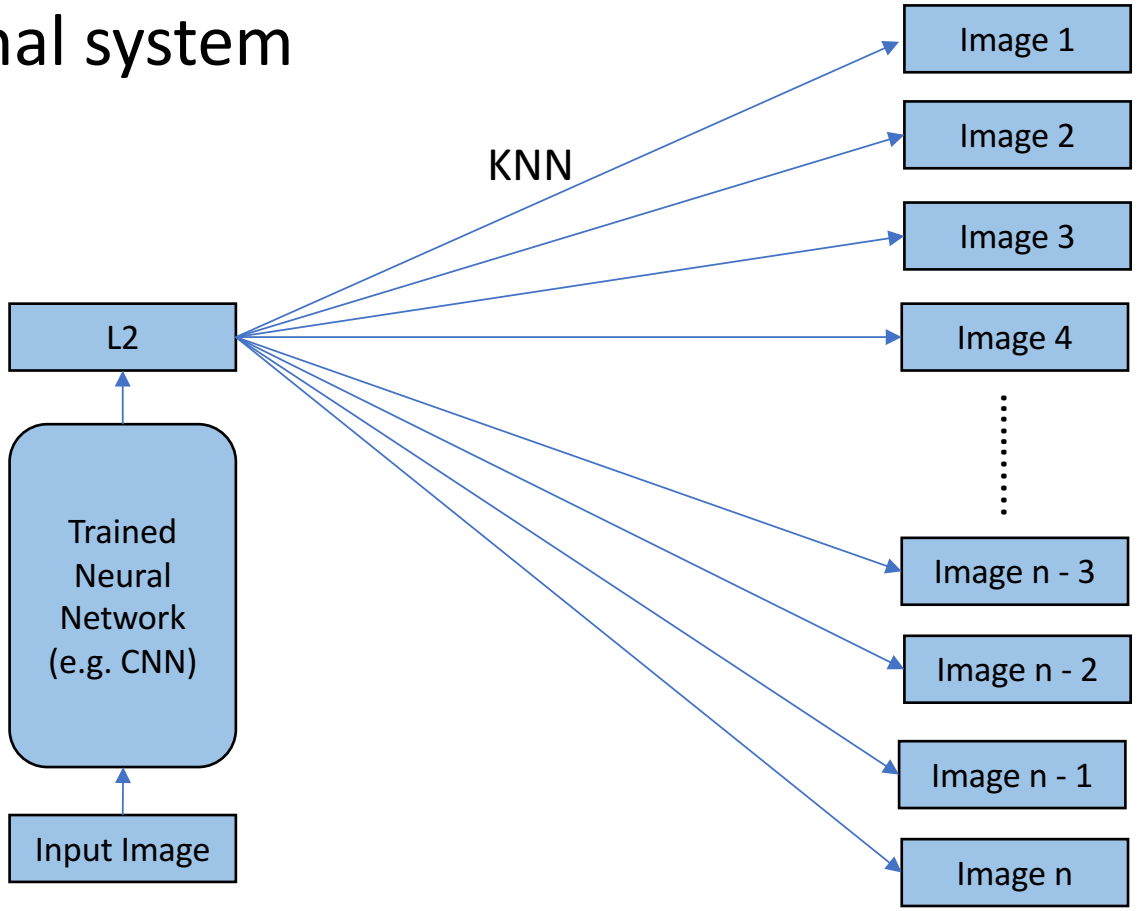
- Build CNN from scratch
  - Works with large training set, enough training time and advanced hardware

Training

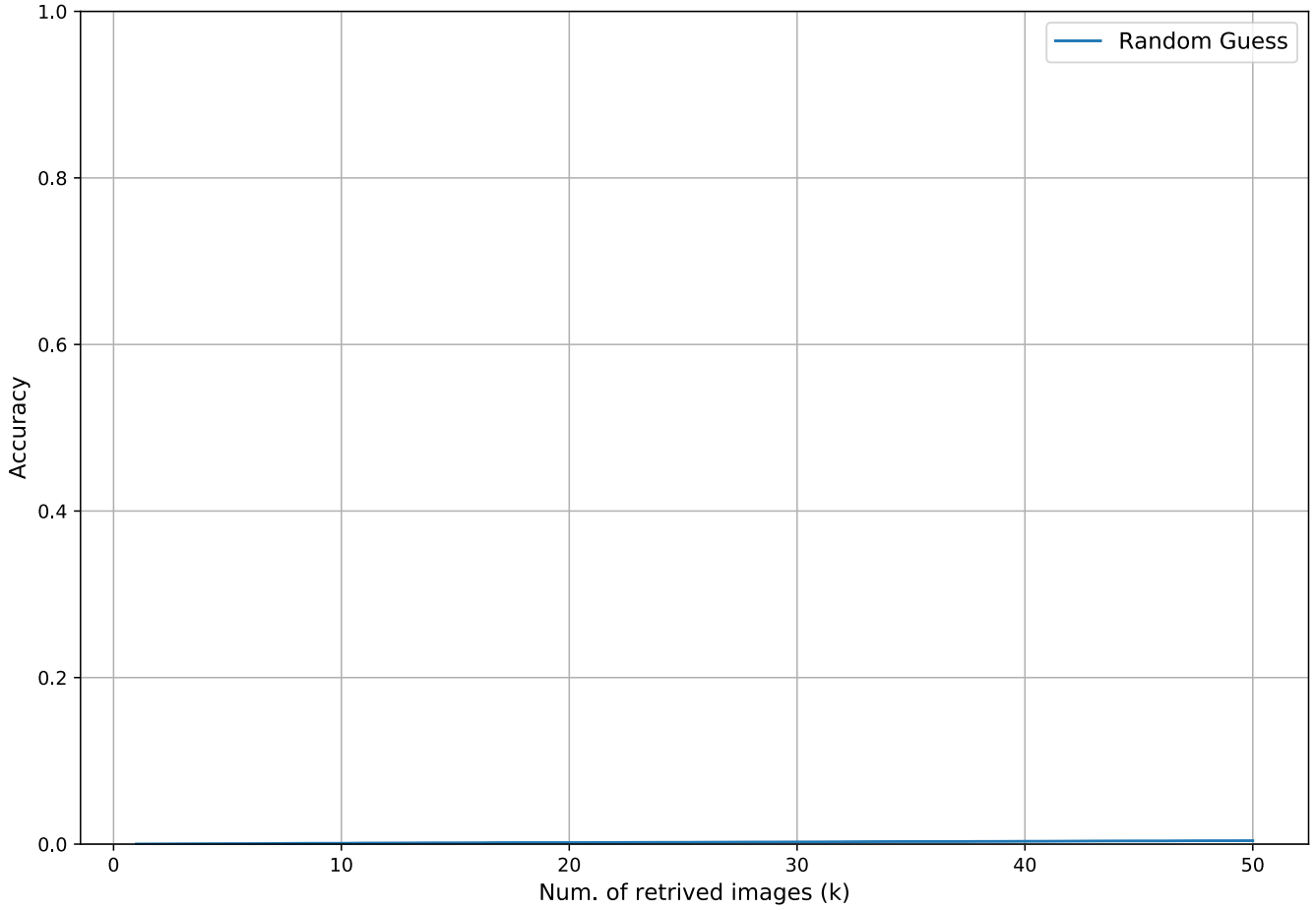
- Build CNN from scratch
  - Works with large training set, enough training time and advanced hardware
- Fine-tuning with pre-trained models
  - VGG16, InceptionV3, and ResNet have been well trained
  - Lower layers usually encode more generic, reusable features
  - Higher layers encode more specialized features
  - Freeze lower layers and only train the top several layers

Prediction

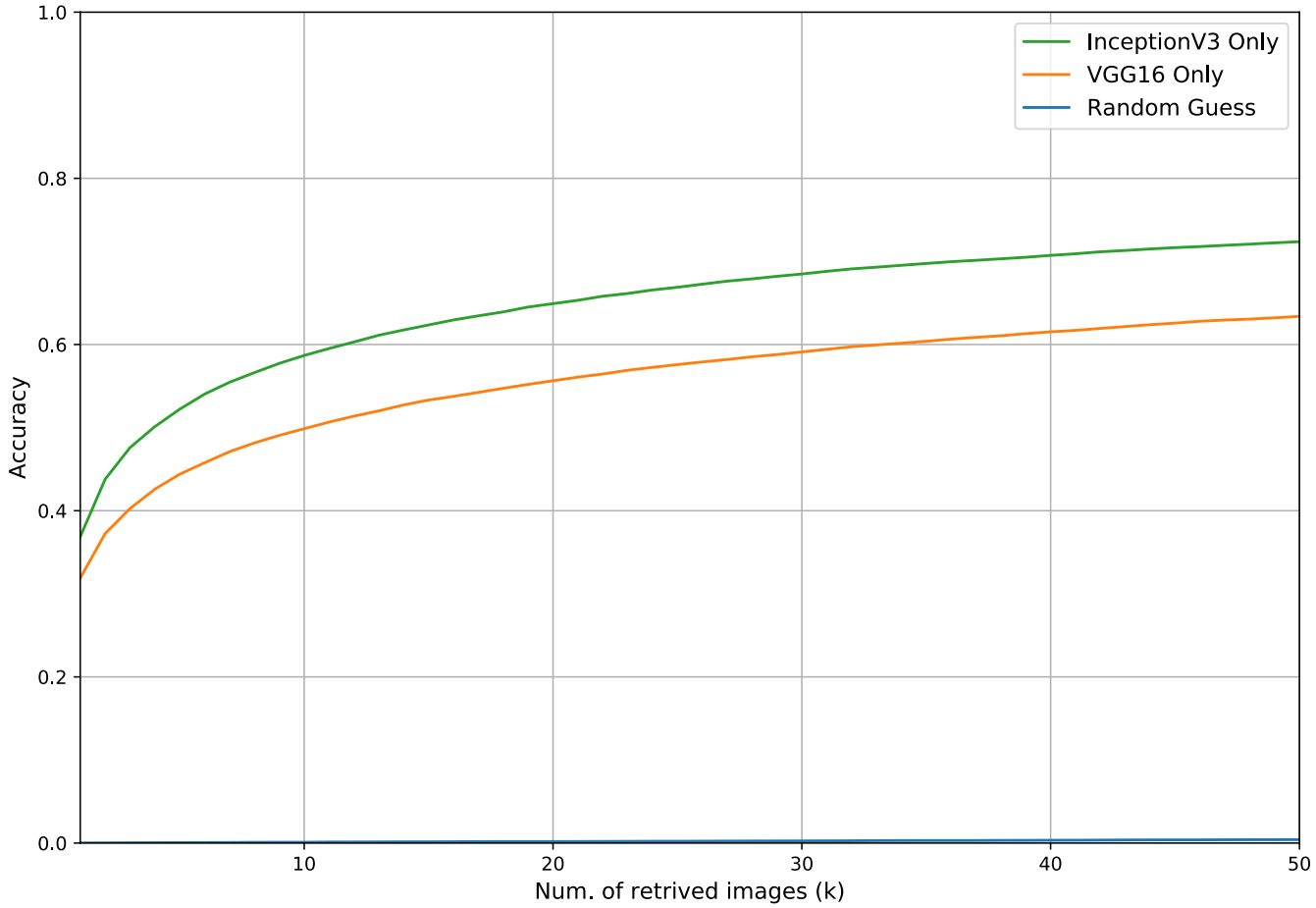
- Given a new image, how to predict its landmark?
  - K-Nearest Neighbors (KNN)
- Final system



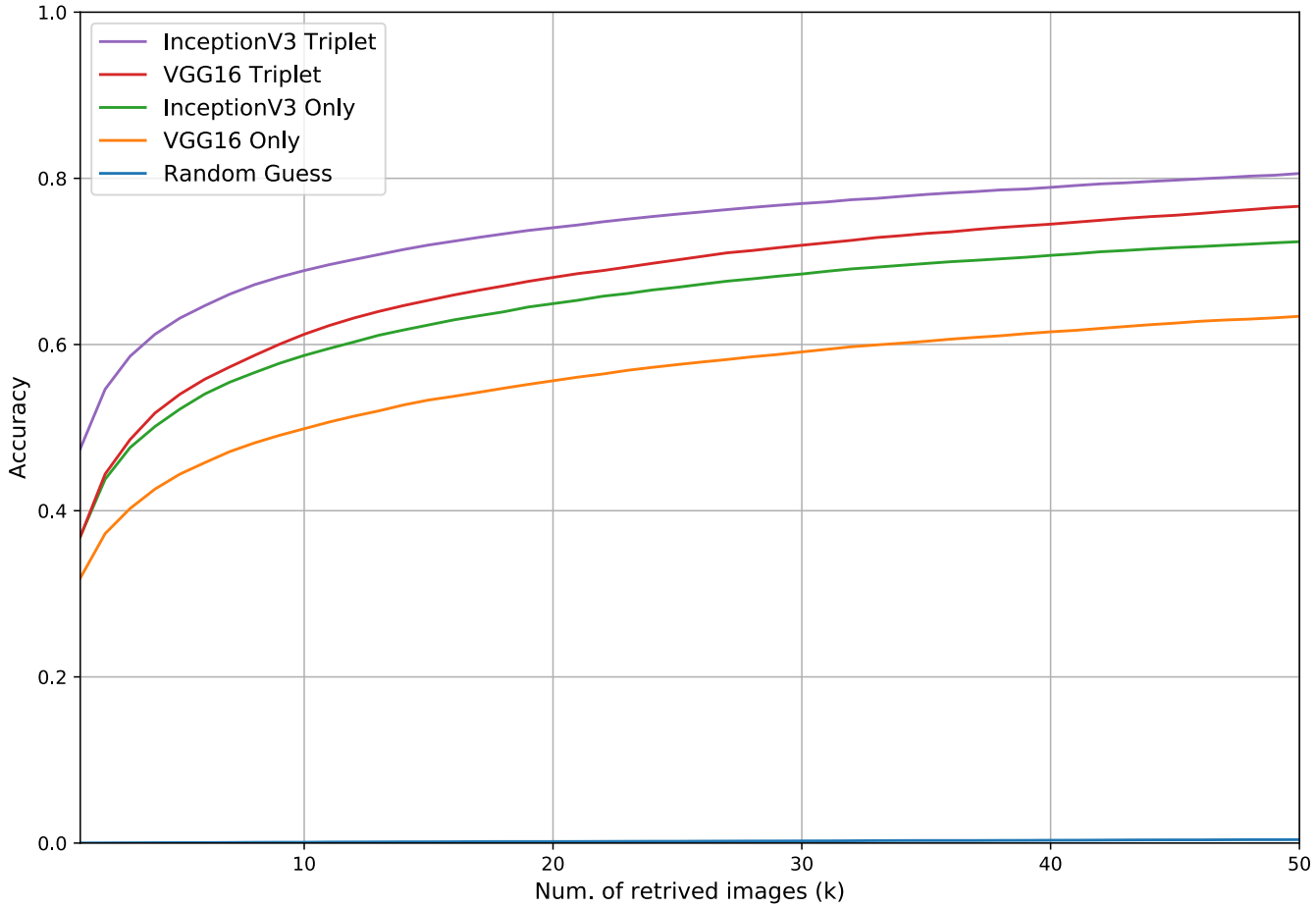
Results


Random Guess
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0
10
20
30
40
50
Num. of retrived images (k)

Results


InceptionV3 Only
VGG16 Only
Random Guess
Accuracy
Num. of retrived images (k)
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50

Results

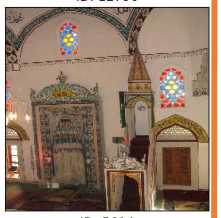


- With fine-tuned InceptionV3 Triplet Network, top 1 accuracy is 47%

Results

Query

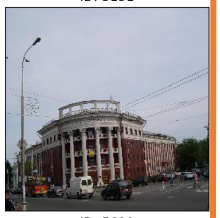
ID: 12790



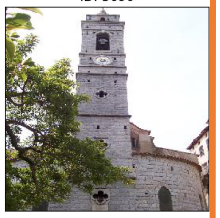
ID: 5614

ID: 6827

ID: 3151



ID: 5690



Results


Query
Similar Images
ID: 12790
ID: 12790
ID: 12790
ID: 9260
ID: 11439
ID: 6554
ID: 13034
ID: 5614
ID: 5614
ID: 5614
ID: 5614
ID: 7565
ID: 5614
ID: 13200
ID: 6827
ID: 6827
ID: 4874
ID: 4526
ID: 3354
ID: 5520
ID: 2370
ID: 3151
ID: 3151
ID: 2361
ID: 3151
ID: 11729
ID: 6423
ID: 8665
ID: 5690
ID: 1309
ID: 59
ID: 5690
ID: 9349
ID: 1309
ID: 59

Results


Query
Similar Images
ID: 12790
ID: 12790
ID: 12790
ID: 9260
ID: 11439
ID: 6554
ID: 13034
ID: 5614
ID: 5614
ID: 5614
ID: 5614
ID: 7565
ID: 5614
ID: 13200
ID: 6827
ID: 6827
ID: 4874
ID: 4526
ID: 3354
ID: 5520
ID: 2370
ID: 3151
ID: 3151
ID: 2361
ID: 3151
ID: 11729
ID: 6423
ID: 8665
ID: 5690
ID: 1309
ID: 59
ID: 5690
ID: 9349
ID: 1309
ID: 59

Summary

- landmark recognition with large amount of real-world images
- Triplet network for metric learning
- Fine-tuning to improve accuracy

- However
  - A subset of images are used => Better performance with all images?
  - Top 1 accuracy 47% => Modern systems (FaceNet) > 95% accuracy

- Future
  - Pre-process images
  - Train specific base network for landmark recognition

* Source code: https://github.com/JifuZhao/Landmark-Recognition

Questions

Thank you !