Lecture 2:
Visual Object Detection, Recognition and Visual-Language Models

Visual Object Recognition

1. Fundamental tasks
   - Image classification
   - Object detection
   - Semantic segmentation
   - Instance segmentation

2. Application-driven tasks
   - Face recognition
   - Fashion item recognition
   - Content Retrieval
   - Content filtering
   - ……

Image Classification

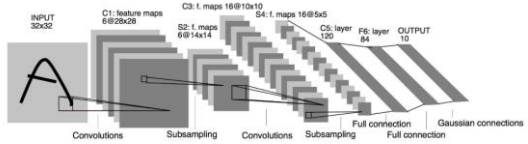
- Given an image, predict its class from a set of candidate classes

f(image)

Cat + Grass/ field

LeNet-5

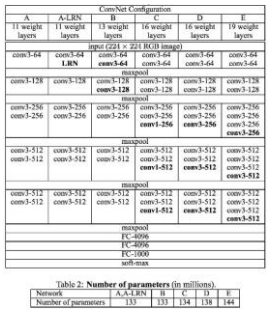
Proposed by LeCun. It's one of the very first convolutional neural networks.



1998

VGG

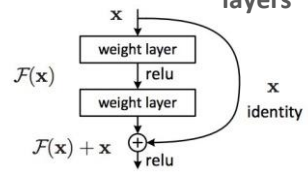
The VGG networks from Oxford were the first to use much smaller 3×3 filters in each convolutional layers



2014

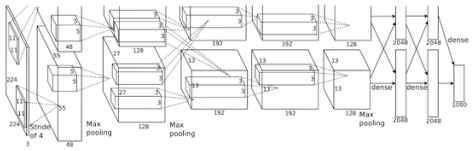
ResNet

Kaiming He proposed to feed the output of two successive convolutional layer AND also bypass the input to the next layers



2015

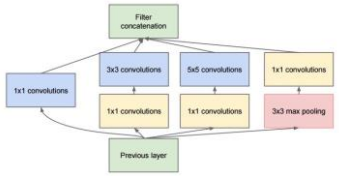
.....



2012

AlexNet

Alex Krizhevsky released a deeper and much wider version of the LeNet.



2014

Inception

Google aimed at reducing the computational burden of deep neural networks.

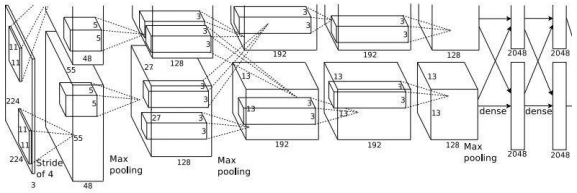
.....

Xception (2016)
DenseNet (2016)
SENet (2017)

Figure 1. Several popular Neural Network Architectures for image classification

Object Detection

- Classification + Localization



Encoder network

- usually pretrained on ImageNet
- e.g., VGG16, ResNet, Inception

Visual Feature

FC-layer

Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...

Softmax Loss

Correct label:
Cat

Multitask Loss

Loss

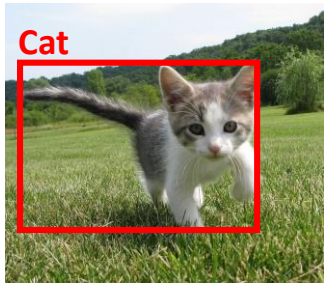
FC-layer

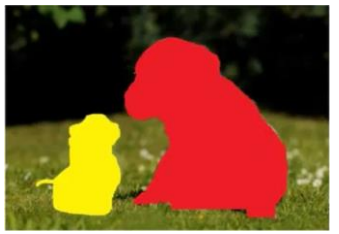
Box Coordinates
(x, y, w, h)

L2 Loss

Correct box:
(x', y', w', h')

Treat localization as a **regression** problem!



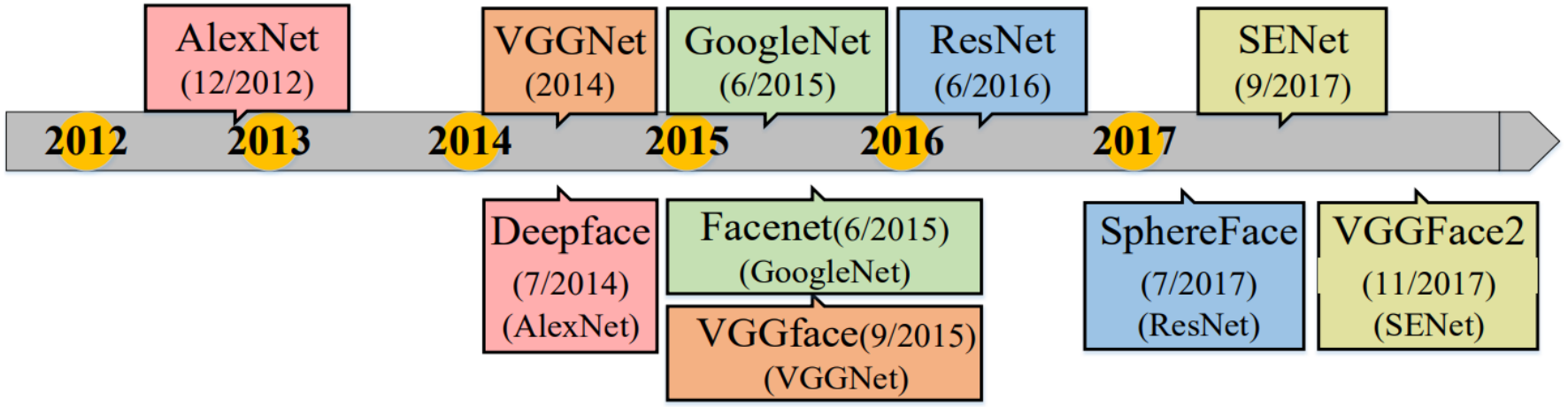


Object Segmentation

Image from CS231n: Convolutional Neural Networks for Visual Recognition. Fei-Fei Li et al.

Application to Face Recognition

- **Face Detection:** detect the locations of the faces in an image.
- **Face Verification**: A one-to-one mapping of a given face against a known identity (e.g. is this the person?).
- **Face Identification:** A one-to-many mapping for a given face against a database of known faces (e.g. who is this person?).



- **The top row** presents the typical network architectures in object classification
- **The bottom row** describes the well-known algorithms of deep face recognition (FR) that use the typical architectures and achieve good performance.
- **The same color rectangles** mean the same architecture. It is easy to find that the architectures of deep FR have always followed those of deep object classification.

Image from Deep Face Recognition: A Survey. Mei Wang et al.

List of Papers to present for Lecture 2 -a

L2: Visual Object Recognition, Detection & Vision-Language Model:

(to be presented by Dr. Ji Wei)

P2-1: Image Recognition:

(Must-Read) A Dosovitskiy, L Beyer, A Kolesnikov et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

(To-Read) Z Liu, Y Lin, Y Cao et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV 2021

(Must-Read, Best Paper) K He, X Zhang, S Ren & J Sun. Deep Residual Learning for Image Recognition. CVPR 2016

P2-2: Object Detection:

(Must-Read) Z Liu, H Hu, Y Lin, Z Yao, Z Xie, Y Wei, J Ning, Y Cao, Z Zhang, L Dong, F Wei &B Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. CVPR 2022.

(Must-Read) S Ren, K He, R Girshick, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. TPAMI 2016.

(To-Read) N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov & S Zagoruyko. End-to-End Object Detection with Transformers. ECCV 2020.

List of Papers to present for Lecture 2 -b

L2: Visual Object Recognition, Detection & Vision-Language Model:

(to be presented by Ji Wei)

P2-3: Vision-Language Models:

(Must-Read) X Chen, X Wang, S Changpinyo, et al. PaLI: A Jointly-Scaled Multilingual Language-Image Model. ICLR 2023.

(Must-Read) W Wang, H Bao, L Dong et al. Image as a Foreign Language: BEiT Pre-training for Vision and Vision-Language Tasks. CVPR 2023

(To-Read) L Ma, ZD Lu, LF Shang & H Li. Multimodal Convolutional Neural Networks for Matching Image and Sentence. ICCV 2015.

Slight Change of Requirements for Paper Presenters and Askers

- **Change of Requirements:** Because the number of students is bigger than expected, we will revert to presentation of papers on a sub-topic by a student, instead of group presentation.
- **Presenter:** The presentation of a sub-topic should cover (20 mins):
  - Objectives of paper
  - Clear literature reviews
  - Limitations, design/ implementation and results
  - Highlight key innovations, answer the how and why questions, such as How it works and Why it works
  - Future work.
  - Presenter Report: the presenter needs to submit a report within 2 weeks time (≤ 2 pages, Single-Spaced Times font 12)
- **Asker:**
  - You will need to pose 2-3 questions
  - Questions should have good depth and help to uncover insight of paper

Papers for Lecture 3: Semantic & Temporal Segmentation, & Relation Grounding

P3-1: Semantic Segmentation: (Presenter: Cheng Yi) (Asker:)

(Must-Read) A Kirillov, E Mintun, N Ravi, et al. Segment anything. arXiv 2023.
(To-Read) K He, G Gkioxari, P Dollár & R Girshick (2017). Mask R-CNN. ICCV 2017.

P3-2: Temporal Segmentation: (Presenter: Nguyen Thong Thanh) (Asker:)

(Must-Read) Z Hou, W Zhong, L Ji, D Gao, K Yan, et al. CONE: An Efficient COarse-to-fiNE Alignment Framework for Long Video Temporal Grounding. ACL 2023.
(Must-Read) LA Hendricks, O Wang, E Shechtman, J Sivic, T Darrell & B Russell. Localizing Moments in Video with Temporal Language. EMNLP 2018.

P3-3: Relation Grounding: (Presenter: xx) (Asker:)

(Must-Read) Y Cong, MY Yang & B Rosenhahn. RelTR: Relation Transformer for Scene Graph Generation. TPAMI. 2023.
(To-Read) B Dai, Y Zhang & D Lin. Detecting Visual Relationships with Deep Relational Networks. CVPR 2017

- **Volunteer Presenters**: Cheng Yi & Nguyen Thong Thanh
- Need one more volunteer the present the third topic
- Will randomly assign the Askers