

DSME 6635: Artificial Intelligence for Business Research

Deep-Learning-Based Object Detection and Video Analysis

Renyu (Philip) Zhang

1

Agenda

- Object Detection
- Video Analysis
- Video Analysis in Business/Econ Research

2

2

1

Data Augmentation

- Reference: https://www.d2l.ai/chapter_computer-vision/image-augmentation.html
- Data Augmentation is used to create additional training data to make training more robust.
 - Data wrapping: transforming data in particular ways.
- The idea is to randomly transform/generate data while we are training:
 - Scale Variation
 - Rotation/mirror symmetry
 - Color Variation

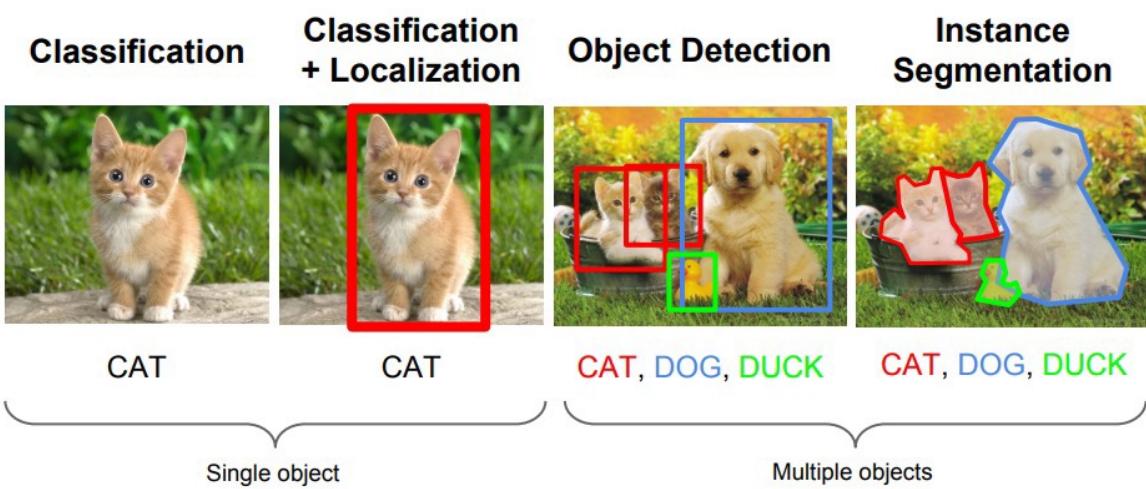


3

3

Beyond Image Classification

- So far, we focus on the task where we get one image and classify it into types of images.
- The next to do is object localization together with image classification.



4

4

Classification + Localization

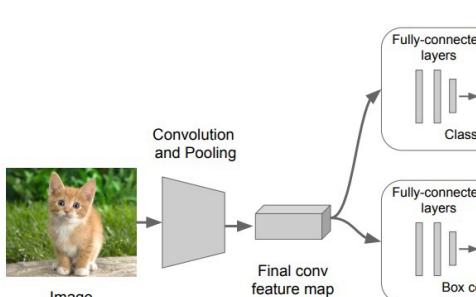
- Classification: C classes
 - Input: Image (x, y, z) matrix
 - Output: Class label c
 - Evaluation metric: Accuracy
- Localization:
 - Input: Image (x, y, z) matrix
 - Output: Box in the image (x, y, w, h)
 - Evaluation metric: Intersection area over the union area
- Object Detection: Classification + Localization

5

5

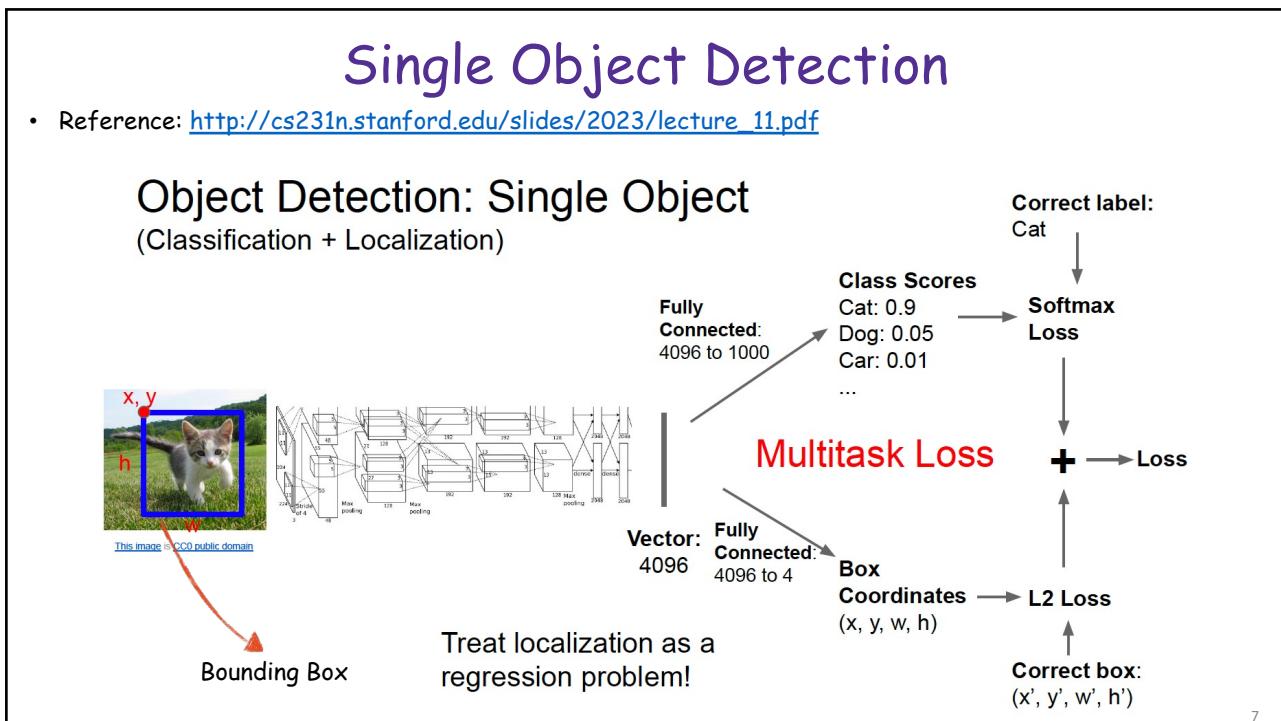
Classification + Localization

- Reference: http://cs231n.stanford.edu/slides/2023/lecture_11.pdf
- The simplest idea is to train a network that output both classes and coordinates (4 values, center of the box, width and height of the box).
- Train classification first, and then attach the regression head to only update the regression head parameters.



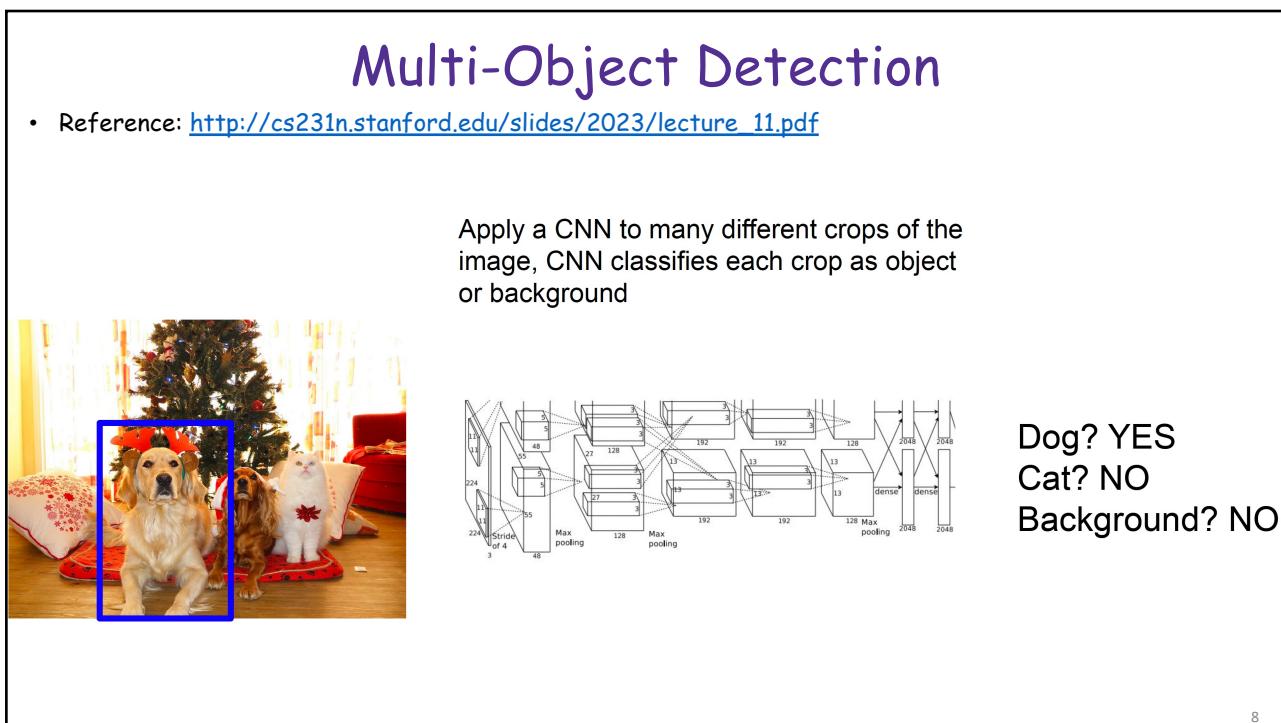
6

6



7

7



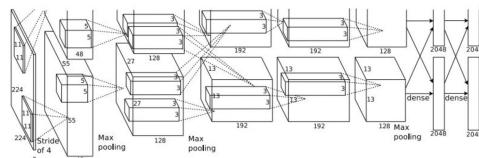
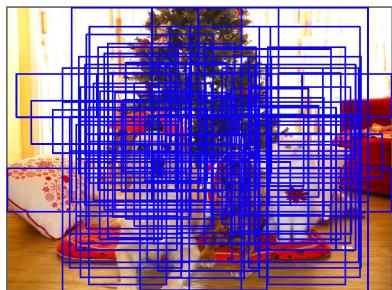
8

8

Multi-Object Detection

- Reference: http://cs231n.stanford.edu/slides/2023/lecture_11.pdf

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? NO
Cat? YES
Background? NO

Problem: Need to apply CNN to huge number of locations, scales, and aspect ratios, very computationally expensive!

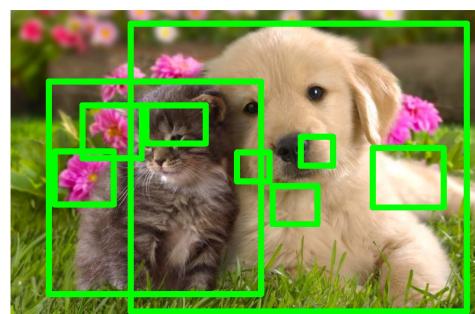
A lot of sliding windows.

9

9

Region Proposals: Selective Search

- Reference: http://cs231n.stanford.edu/slides/2023/lecture_11.pdf
- Selective Search starts by over-segmenting the image into a lot of initial regions and then merges these regions based on various similarity criteria such as color, texture, size, and shape compatibility. The process results in a set of region proposals that potentially contain objects.



Alexe et al, "Measuring the objectness of image windows", TPAMI 2012
Uijlings et al, "Selective Search for Object Recognition", IJCV 2013
Cheng et al, "BING: Binarized normed gradients for objectness estimation at 300fps", CVPR 2014
Zitnick and Dollar, "Edge boxes: Locating object proposals from edges", ECCV 2014

10

10

Region-Based CNN (R-CNN)

http://cs231n.stanford.edu/slides/2023/lecture_11.pdf
https://www.d2l.ai/chapter_computer-vision/rcnn.html

Rich feature hierarchies for accurate object detection and semantic segmentation
 R Girshick, J Donahue, T Darrell... - ... and pattern recognition, 2014 - openaccess.thecvf.com
 ... Object detection with R-CNN Our object detection system consists of three modules. The first generates category-independent region proposals. These proposals define the set of ...
 ☆ Save 99 Cite Cited by 35902 Related articles All 46 versions ☰

One CNN trained for each ROI.

Selective search

Class prediction
Bounding box prediction

CNN

Class prediction
Bounding box prediction

Linear Regression

SVM

~2K Regions of Interest (RoI)

Fig. 14.8.1 The R-CNN model.

R-CNN is computationally very expensive, even with pre-trained CNNs.

11

11

Fast R-CNN

http://cs231n.stanford.edu/slides/2023/lecture_11.pdf
https://www.d2l.ai/chapter_computer-vision/rcnn.html

Fast r-cnn
 R Girshick - ... of the IEEE international conference on ..., 2015 - openaccess.thecvf.com
 This paper proposes a Fast Region-based Convolutional Network method (Fast R-CNN) for object detection. Fast R-CNN builds on previous work to efficiently classify object proposals ...
 ☆ Save 99 Cite Cited by 33063 Related articles All 43 versions ☰

Linear Regression

Class prediction
Bounding box prediction

FC

RoI pooling

CNN

Selective search

Softmax

2 x 2 RoI Pooling

5 6
9 10

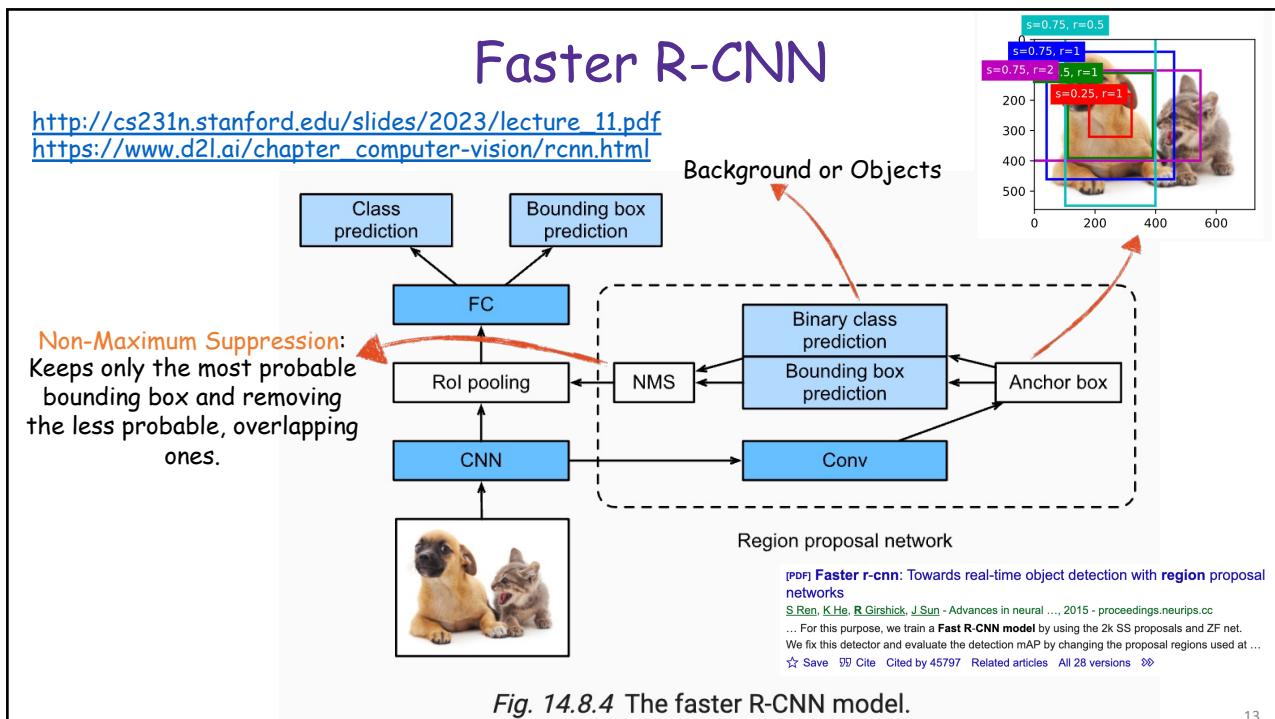
0 1 2 3
4 5 6 7
8 9 10 11
12 13 14 15

We only train one backbone CNN.

Fig. 14.8.2 The fast R-CNN model.

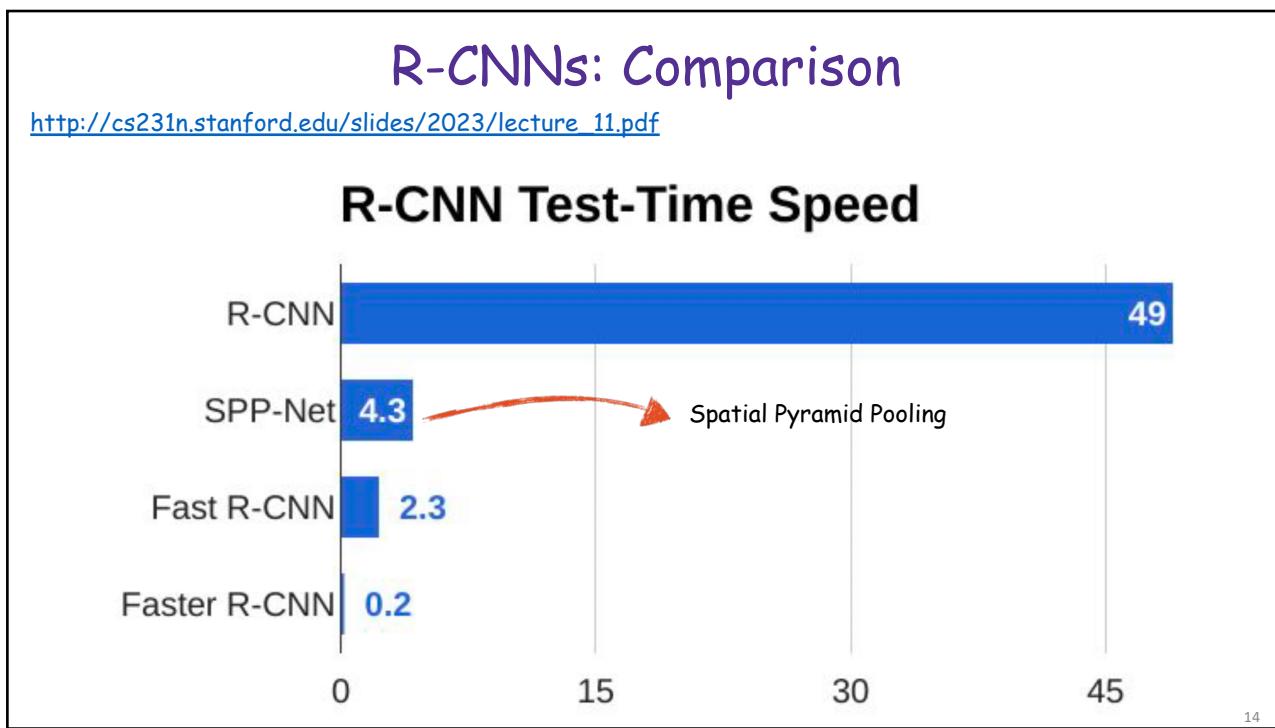
12

12



13

13



14

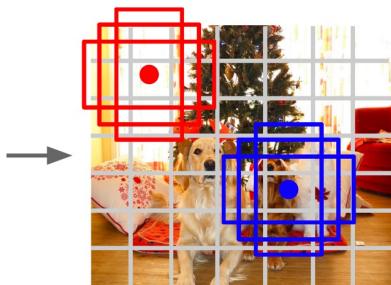
14

Single-Stage Detection: YOLO

http://cs231n.stanford.edu/slides/2023/lecture_11.pdf



Input image
3 x H x W



Divide image into grid
7 x 7
Image a set of **base boxes**
centered at each grid cell
Here B = 3

Within each grid cell:

- Regress from each of the B base boxes to a final box with 5 numbers: (dx, dy, dh, dw, confidence)
- Predict scores for each of C classes (including background as a class)
- Looks a lot like RPN, but category-specific!

Output:
7 x 7 x (5 * B + C)

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016
Lin et al, "Focal Loss for Dense Object Detection", ICCV 2017

YOLO has a lot of follow versions (v8 now): <https://huggingface.co/models?other=yolov8>

15

15

YOLO Loss Function

https://www.cs.utexas.edu/~yukez/cs391r_fall2021/slides/pre_09-02_Shivang.pdf

- ❖ For YOLO, we need to minimize the following loss
- ❖ Sum squared error is used

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

Coordinate Loss: Minimize the difference between x,y,w,h pred and x,y,w,h ground truth. ONLY IF object exists in grid box and if bounding box is resp for pred

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

Confidence Loss: Loss based on confidence ONLY IF there is object

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

No Object Loss based on confidence if there is no object

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

Class loss, minimize loss between true class of object in grid box

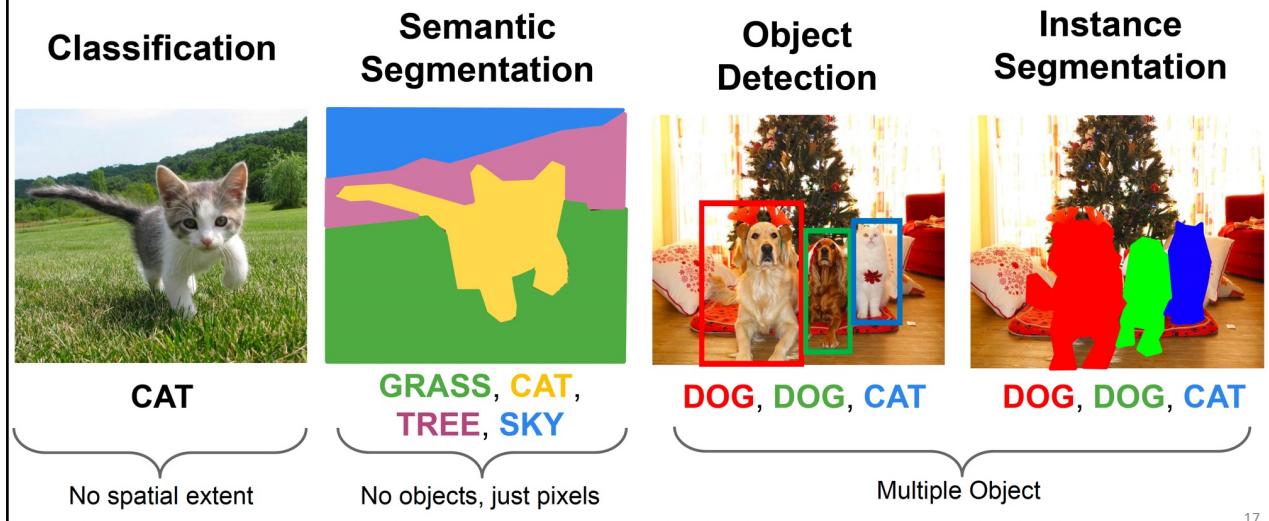
16

16

Beyond Boxes: Semantic Segmentation

http://cs231n.stanford.edu/slides/2023/lecture_11.pdf

- Semantic Segmentation: Predict the class of each pixel.



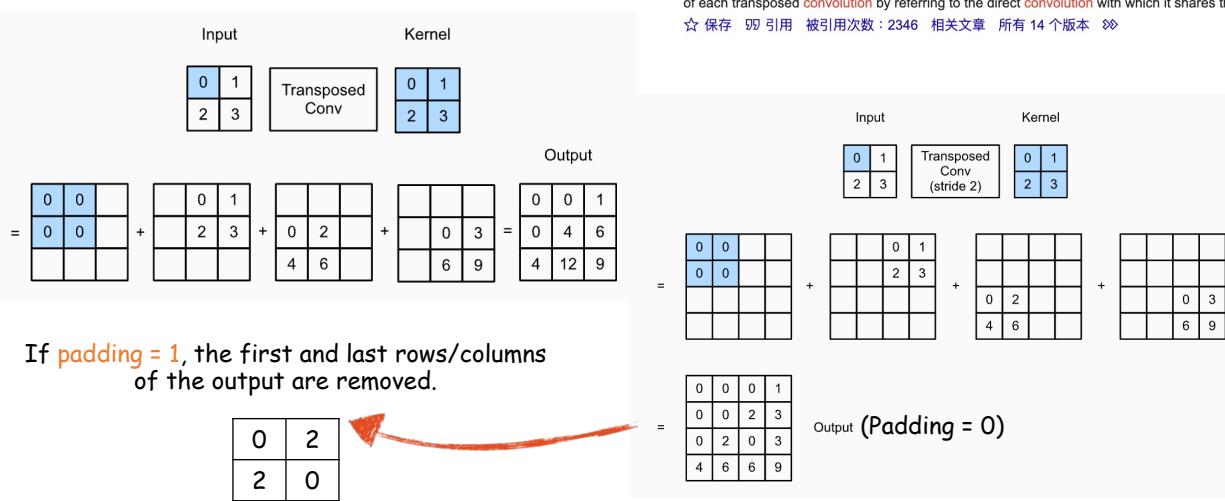
17

17

Transposed Convolution

https://www.d2l.ai/chapter_computer-vision/transposed-conv.html

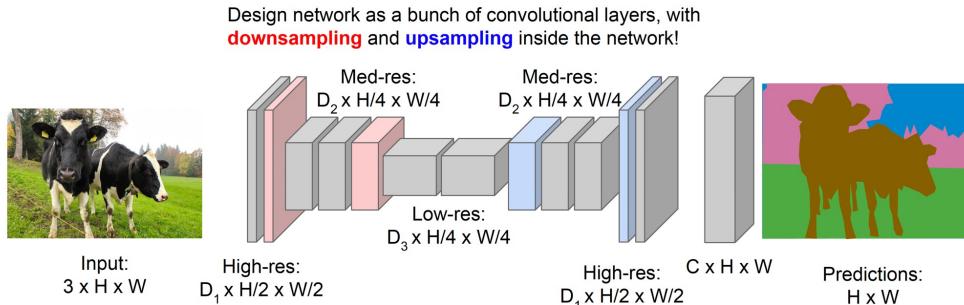
- Transposed Convolution: Increase/upsample the spatial dimensions of convolution.



18

Fully Convolutional Network (FCN)

https://www.d2l.ai/chapter_computer-vision/fcn.html
http://cs231n.stanford.edu/slides/2023/lecture_11.pdf

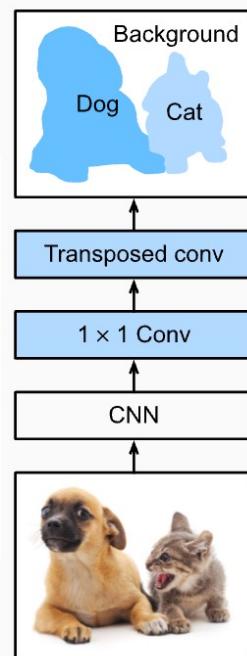


Fully convolutional networks for semantic segmentation

J Long, E Shelhamer, T Darrell - Proceedings of the IEEE ..., 2015 - openaccess.thecvf.com

... for per-pixel tasks like **semantic segmentation**. We show that a **fully convolutional network** (FCN) trained end-to-end, pixels-to-pixels on **semantic segmentation** exceeds the state-of-the-...

☆ 保存 99 引用 被引用次数 : 47601 相关文章 所有 55 个版本 Web of Science: 2429 ☺



19

19

Agenda

- Object Detection
- Video Analysis
- Video Analysis in Business/Econ Research

20

20

Video = Image x Time

http://cs231n.stanford.edu/slides/2023/lecture_10.pdf

A video is a **sequence** of images

4D tensor: $T \times 3 \times H \times W$
 (or $3 \times T \times H \times W$)



21

21

Videos Classification

http://cs231n.stanford.edu/slides/2023/lecture_10.pdf



**Swimming
Running
Jumping
Eating
Standing**

Input video:
 $T \times 3 \times H \times W$

22

22

Videos Are Big!

http://cs231n.stanford.edu/slides/2023/lecture_10.pdf

Videos are ~30 frames per second (fps)



Input video:
 $T \times 3 \times H \times W$

Size of uncompressed video
(3 bytes per pixel):

SD (640 x 480): **~1.5 GB per minute**
HD (1920 x 1080): **~10 GB per minute**

Solution: Train on short **clips**: low
fps and low spatial resolution
e.g. $T = 16$, $H=W=112$
(3.2 seconds at 5 fps, 588 KB)

23

23

Training on Clips

http://cs231n.stanford.edu/slides/2023/lecture_10.pdf

Raw video: Long, high FPS



Training: Train model to classify short **clips** with low FPS



Testing: Run model on different clips, average predictions



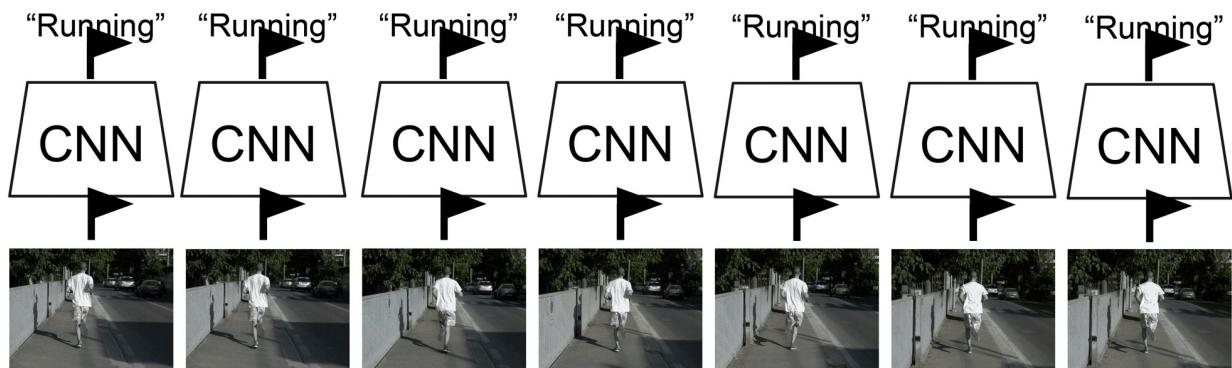
24

24

A Strong Baseline: Single Frame CNN

http://cs231n.stanford.edu/slides/2023/lecture_10.pdf

- Train a normal 2D CNN to classify video frames independently; average the predicted probs in testing.



25

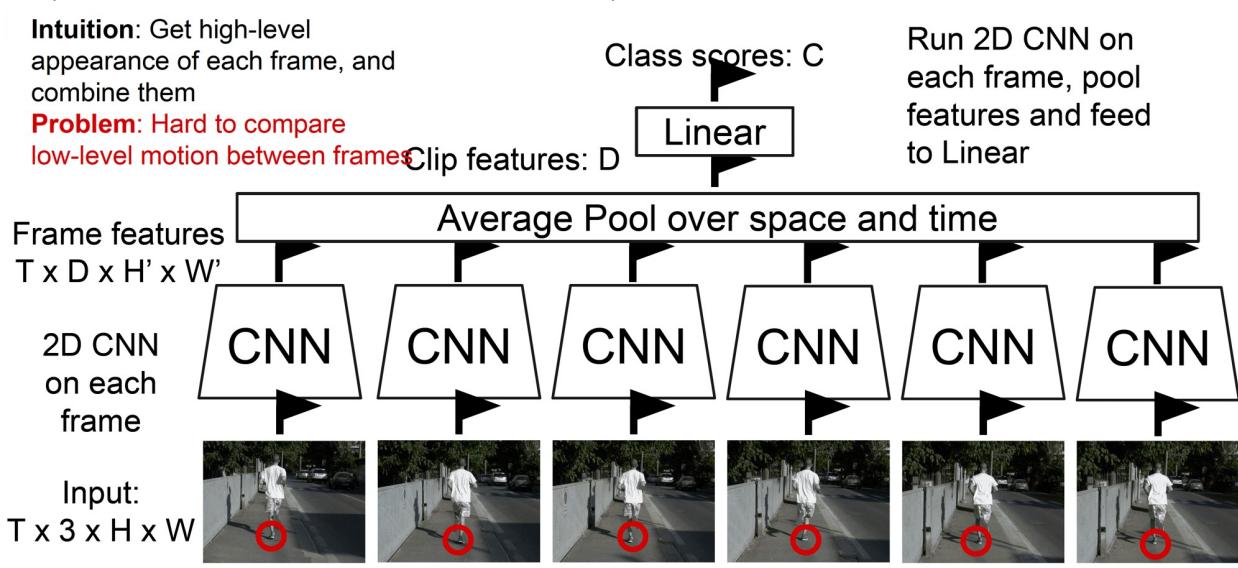
25

Late Fusion

http://cs231n.stanford.edu/slides/2023/lecture_10.pdf

Intuition: Get high-level appearance of each frame, and combine them

Problem: Hard to compare low-level motion between frames



26

26

Early Fusion

http://cs231n.stanford.edu/slides/2023/lecture_10.pdf

Intuition: Compare frames with very first conv layer, after that normal 2D CNN

Problem: One layer of temporal processing may not be enough!

First 2D convolution collapses all temporal information:
Input: $3T \times H \times W$
Output: $D \times H \times W$

Reshape:
Input: $T \times 3 \times H \times W$

Rest of the network is standard 2D CNN

Class scores: C

Large-scale video classification with convolutional neural networks
A Karpathy, G Toderici, S Shetty, T Leung... - Proceedings of the ..., 2014 - cv-foundation.org
... performance of CNNs in large-scale video classification, where the networks ... video classification benchmarks that match the scale and variety of existing image datasets because ...
☆ 保存 59 引用 被引用次数 : 8299 相关文章 所有 46 个版本 ☰

27

27

3D CNN

http://cs231n.stanford.edu/slides/2023/lecture_10.pdf

Intuition: Use 3D versions of convolution and pooling to slowly fuse temporal information over the course of the network

Each layer in the network is a 4D tensor: $D \times T \times H \times W$
Use 3D conv and 3D pooling operations

Input: $C \times T \times H \times W$

Class scores: C

3D CNN

Input: $3 \times T \times H \times W$

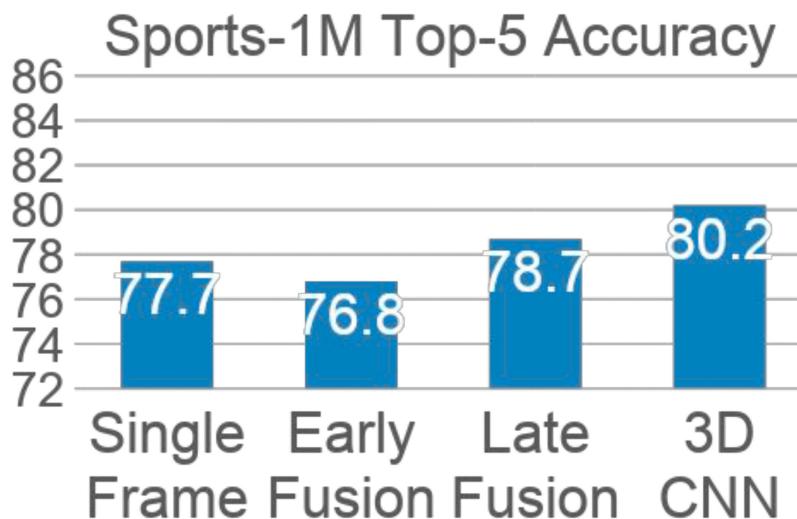
Class Scores

28

28

Comparison

http://cs231n.stanford.edu/slides/2023/lecture_10.pdf
Sports-1M dataset: <https://github.com/gtoderici/sports-1m-dataset>



Single Frame
model works well
– always try this
first!

3D CNNs have
improved a lot
since 2014!

29

29

Agenda

- Object Detection
- Video Analysis
- Video Analysis in Business/Econ Research

30

30

Frontiers: Unmasking Social Compliance Behavior During the Pandemic
 S Zhang, K Xu, K Srinivasan
 Marketing Science, 2023 · pubsonline.informs.org

In 2020, as the novel coronavirus spread globally, face masks were recommended in public settings to protect against and slow down viral transmission. People complied to varying extents, and their reactions may have been driven by a variety of psychological factors. Based on the literature on social influence and on mask-wearing, we define three customer segments: *Fully-Compliant* customers wear masks, and they seem motivated primarily by concerns about their own health risk. *Partially-Compliant* customers also wear masks, but with improper and ineffective coverage; our empirical analysis suggests that they are motivated primarily by a desire to comply with social norms. Finally, *Unmasked* customers do not wear masks. We examine changes in shopping behaviors with the onset of the pandemic to corroborate the conjectured mask-wearing motives. We find that the three groups made significantly different behavior changes: *Fully-Compliant* customers shopped significantly faster and practiced stricter social distancing with the onset of the pandemic, whereas the other two groups did not adjust their shopping duration or social distancing.

History: K. Sudhir served as the senior editor for this article. This paper was accepted through the *Marketing Science*: Frontiers review process.

Funding: Financial support from the National Natural Science Foundation of China [Grants 71622008 and 71832006] and National Social Science Foundation of China [No. 22VRC174] is gratefully acknowledged.

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/mksc.2022.1419>.



SHOW LESS ^

☆ Save 99 Cite Cited by 1 Related articles All 6 versions >>

Social Compliance

- Apply various CV technologies to video clips from security camera of a retail store to classify its customers into **fully-compliant**, **partially-compliant** and **non-compliant**.
- Fully-compliant customers shopped **significantly faster** and practiced **stricter social distancing** during the **COVID-19** pandemic, whereas the other two groups did not adjust their behaviors.

31

31

Introducing machine-learning-based data fusion methods for analyzing multimodal data: An application of measuring trustworthiness of microenterprises
 X Luo, N Jia, E Ouyang, Z Fang
 Strategic Management Journal, 2024 · Wiley Online Library

Research Summary

Multimodal data, comprising *interdependent* unstructured text, image, and audio data that collectively characterize the same source, with video being a prominent example, offer a wealth of information for strategy researchers. We emphasize the theoretical importance of capturing the interdependencies between different modalities when evaluating multimodal data. To automate the analysis of video data, we introduce advanced deep machine learning and data fusion methods that comprehensively account for all intra- and inter-modality interdependencies. Through an empirical demonstration focused on measuring the trustworthiness of grassroots sellers in live streaming commerce on Tik Tok, we highlight the crucial role of interpersonal interactions in the business success of microenterprises. We provide access to our data and algorithms to facilitate data fusion in strategy research that relies on multimodal data.

Managerial Summary

Our study highlights the vital role of both verbal and nonverbal communication in attaining strategic objectives. Through the analysis of multimodal data—incorporating text, images, and audio—we demonstrate the essential nature of interpersonal interactions in bolstering trustworthiness, thus facilitating the success of microenterprises. Leveraging advanced machine learning techniques, such as data fusion for multimodal data and explainable artificial intelligence, we notably enhance predictive accuracy and theoretical interpretability in assessing trustworthiness. By bridging strategic research with cutting-edge computational techniques, we provide practitioners with actionable strategies for enhancing communication effectiveness and fostering trust-based relationships. Access our data and code for further exploration.



SHOW LESS ^

☆ Save 99 Cite Cited by 1 Related articles All 2 versions >>

Data Fusion to Measure Trustworthiness

- Use DNN (CNN, Recurrent-CNN) to analyze the multi-modal video data (video, verbal, and audio) on **TikTok e-commerce**.
- Trustworthiness (of grassroots sellers) is measured using **buyer feedback** on **TikTok**.
- Multi-modal data improves the prediction accuracy of seller trustworthiness.

32

32