Computer Vision: Looking Back to Look Forward

Svetlana Lazebnik
IRIM Short Course
Spring 2020
Outline of lectures

**Tues. 1/28:** Where are we going, where have we been?

**Wed. 1/29:** Stepping back: What is the nature of visual perception?

**Thurs. 1/30:** History of ideas in recognition – Part I

**Tues. 2/4:** History of ideas in recognition – Part II

**Wed. 2/5:** Emerging trends – or, what should we be working on?

**Thurs. 2/6:** Ethical issues for computer vision researchers
Where are we now?

• Glass is half full?
  • Current methods *actually work* and are better than older methods in most ways
  • Good methodologies: standard benchmarks, quantitative comparisons, ablation studies
  • Good infrastructure: deep learning packages, cloud services, etc.
  • Culture of code sharing and reproducibility

• Glass is half empty?
  • People are too focused on the benchmarks
  • Number of papers is exploding but diversity of topics is not increasing
  • Cutting-edge research is prohibitively resource-intensive
  • Core benchmarks (ImageNet, COCO) are likely saturating but bigger datasets (or larger amounts of computing power) are not yet in sight
What’s wrong with ImageNet classification?

- Performance has saturated – further advances in architectures by looking at ImageNet unlikely.
What’s wrong with ImageNet classification?

• Nobody believes that attaching labels to images is the right task

Favorite cartoon of Jan Koenderink  Favorite cartoon of David Forsyth
What’s wrong with ImageNet classification?

• Image classification models are too easy to “hack”

A. Nguyen, J. Yosinski, J. Clune, Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, CVPR 2015
What’s wrong with ImageNet classification?

• Image classification models are too easy to “hack”

How can we move forward?

• Focus on tasks beyond image classification
  • “Rich” prediction tasks
Figure 2: Our model (MarrNet) has three major components: (a) 2.5D sketch estimation, (b) 3D shape estimation, and (c) a loss function for reprojection consistency. MarrNet first recovers object normal, depth, and silhouette images from an RGB image. It then regresses the 3D shape from the 2.5D sketches. In both steps, it uses an encoding-decoding network. It finally employs a reprojection consistency loss to ensure the estimated 3D shape aligns with the 2.5D sketches. The entire framework can be trained end-to-end.

J. Wu, Y. Wang, T. Xue, X. Sun, W. Freeman, J. Tenenbaum, **MarrNet: 3D Shape Reconstruction via 2.5D Sketches**, NeurIPS 2017
Learning 3D structure


Learning controllers

Fig. 1. Simulated characters performing highly dynamic skills learned by imitating video clips of human demonstrations. Left: Humanoid performing cartwheel B on irregular terrain. Right: Backflip A retargeted to a simulated Atlas robot.

Video

X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, S. Levine, SFV: Reinforcement Learning of Physical Skills from Videos, SIGGRAPH Asia 2018
S. Zuffi, A. Kanazawa, M. Black, Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape from Images, CVPR 2018
Image-to-image translation

Challenge: Diverse prediction

Toward Multimodal Image-to-Image Translation, NIPS 2017
How can we move forward?

• Focus on tasks beyond image classification
  • “Rich” prediction tasks
  • Generation
State-of-the-art GANs

• Faces: 1024x1024 resolution, CelebA-HQ dataset

T. Karras, T. Aila, S. Laine, and J. Lehtinen, Progressive Growing of GANs for Improved Quality, Stability, and Variation, ICLR 2018

Follow-up work
State-of-the-art GANs

- May shed light on important open questions, such as:
  - Do we need explicitly compositional (part-based) representations?

A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, ICLR 2019
State-of-the-art GANs

• May shed light on important open questions, such as:
  • Do we need explicitly compositional (part-based) representations?

A. Brock, J. Donahue, K. Simonyan, *Large scale GAN training for high fidelity natural image synthesis*, ICLR 2019
State-of-the-art GANs

• May shed light on important open questions, such as:
  • Do we need explicitly compositional (part-based) representations?

State-of-the-art GANs

• May shed light on important open questions, such as:
  • Do we need explicit 3D representations?

https://twitter.com/phillip_isola/status/1066574964793458689
State-of-the-art GANs

• May shed light on important open questions, such as:
  • Do we need explicit 3D representations?

T. Nguyen-Phuoc et al., *HoloGAN: Unsupervised Learning of 3D Representations From Natural Images*, 2019
Integrating discriminative and generative models?

Encoder-decoder architecture
Integrating discriminative and generative models?

• Can “hallucination” help with recognition?
  • Low-shot learning, incremental learning, transfer learning

Figure 1. Given a single image of a novel visual concept, such as a blue heron, a person can visualize what the heron would look like in other poses and different surroundings. If computer recognition systems could do such hallucination, they might be able to learn novel visual concepts from less data.

Wang et al. (2018)

High-quality images synthesized by “inverting” an ImageNet-pretrained network

Yin et al. (2019)
Integrating discriminative and generative models?

- Can attribute-based manipulation of training examples help to train more accurate or less biased models?

Figure 1: Counterfactual attribute sensitivity measures the effect of manipulating a specific property of an image on the output of a trained classifier. In this example, we consider the effect of adding facial hair on the output of a smiling classifier. If the classifier’s output systematically changes as a result of adding or removing facial hair then a potentially undesirable bias has been detected since, all else being equal, facial hair should be irrelevant to the classification task.

Denton et al. (2019)

Figure 1: Our approach learns to actively create training images for fine-grained attribute comparisons, focusing attention on informative pairs of images unavailable in the real image training pool. Here we show two example pairs it generated to improve learning of open (left) and masculine (right).

Yu and Grauman (2019)
Possible ways forward

• Focus on tasks beyond image classification
  • “Rich” prediction tasks
  • Generation

• Move away from full supervision
  • Self-supervised learning
Self-supervised learning

- Key idea: Use one part of the data to predict another part of the data
  - Contrast with unsupervised learning, where you use data to predict itself
- Motivation (in traditional recognition settings): learn feature representations that could match the accuracy of ImageNet-trained networks without requiring large amounts of labeled training data

G. Larsson, M. Maire, and G. Shakhnarovich, Colorization as a Proxy Task for Visual Understanding, CVPR 2017
Self-supervision for traditional recognition tasks

• Carefully designed self-supervised representations may be able to beat fully supervised ones for “non-semantic” tasks (e.g., surface normal estimation) and even PASCAL detection
  • P. Goyal, D. Mahajan, A. Gupta, I. Misra, Scaling and Benchmarking Self-Supervised Visual Representation Learning, ICCV 2019
  • I. Misra and L. van der Maaten, Self-Supervised Learning of Pretext-Invariant Representations, arXiv 2019

• If there is enough supervised data available for the target task (e.g., COCO) one may not need to pre-train at all:
  • K. He, R. Girshick, P. Dollar, Rethinking ImageNet Pre-training, ICCV 2019
Robot grasping

L. Pinto and A. Gupta, *Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours*, ICRA 2016

[YouTube video](#)
Future prediction

Sound prediction

(a) Images grouped by audio cluster   (b) Clustered audio stats.   (c) CNN model

Self-supervision for exploration

- Learning to play a video game with sparse or non-existent rewards

(a) learn to explore on Level-1  
(b) explore faster on Level-2

Video

Self-supervision: Discussion

• For traditional recognition benchmarks, self-supervision is not as effective (so far) as getting enough supervised training data

• However, self-supervision is very attractive (and possibly unavoidable) in scenarios that go beyond still images: video, video + audio, and sensorimotor learning
Possible ways forward

• Focus on tasks beyond image classification
  • “Rich” prediction tasks
  • Generation

• Move away from full supervision
  • Self-supervised learning

• Focus on embodied vision, sensorimotor learning
Embodied vision

• A cross-section of topics from one representative researcher:

• See also: Abhinav Gupta, Pieter Abbeel, Sergey Levine, Chelsea Finn
Embodied vision platforms

- Simulation: AI2Thor, Habitat
  ![AI2Thor](image1) ![Habitat](image2)

- Real robots: PyRobot
  ![PyRobot](image3)

What can you do with PyRobot?

- Manipulation
- Navigation
- Demonstrations
Simulation vs. reality

“Our main finding is that computer vision does matter. Models equipped with intermediate representations train faster, achieve higher task performance, and generalize better to previously unseen environments.”

Challenges for embodied vision

• Speed and cost of doing experiments
• Training faster, generalizing from one task to another
• Multi-agent interactions, communication
• Planning and reasoning
• Integrating memory, knowledge
Parting thoughts

• The next breakthroughs are not likely to come cheaply
• Access to data, computation, and platforms will be key
• The next few years will make it clearer which problems have been truly solved and which ones have been underestimated
• The hard problems are getting into “AI-complete” territory
What wasn’t on my list?

• Adversarial examples
• Few-shot learning
• Vision and language, vision and knowledge
• Explainability
• Reasoning
• Lifelong learning