Deep learning trends
(a non-comprehensive, non-objective view)

William Blake, Jacob's Ladder (1805)
Outline

• Vision
• Embodied learning (RL, robotics)
• Language
Deep learning in vision: Where are we now?

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

• Glass is half empty?
  • Too much focus on benchmarks and numbers
  • Number of papers is exploding but diversity of topics is not necessarily increasing
  • Cutting-edge research is becoming prohibitively resource-intensive
  • Core benchmarks (ImageNet, COCO) are likely saturating but bigger datasets (or larger amounts of computing power) are not yet generally accessible
ImageNet: Asset or liability?

- Performance on the basic classification task has saturated
ImageNet: Asset or liability?

• Performance on the basic classification task has saturated

Figure 1: When presented with a model’s prediction and the original ImageNet label, human annotators now prefer model predictions on average (Section 4). Nevertheless, there remains considerable progress to be made before fully capturing human preferences.

ImageNet: Asset or liability?

• Attaching labels to images is not very meaningful in the first place
ImageNet: Asset or liability?

• Attaching labels to images is not very meaningful in the first place
How can we move forward?

• Develop the next generation of architectures
Beyond convolutional networks?

• Transformers for images

A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020
Beyond convolutional networks?

- Transformers for images

A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020

BiT: ResNet (Big Transfer)
ViT: Vision Transformer (Base/Large/Huge)

JFT-300M: internal Google dataset (not public)
How can we move forward?

• Develop the next generation of architectures
• Go beyond classification
  • “Rich” prediction tasks
“Rich” prediction tasks

Class label to image

Zhang et al. (2019)

Image to image

Isola et al. (2017)

Text to image

Qiao et al. (2019)

Image to 3D

Hane et al. (2019)
Learning 3D structure

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
Historical context: Marr’s visual processing pipeline

David Marr (1945-1980)

Input Image

Perceived intensities

Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries.

Primal Sketch

2 1/2-D Sketch

Local surface orientation and discontinuities in depth and in surface orientation

3-D Model Representation

3-D models hierarchically organised in terms of surface and volumetric primitives

Image source
Learning 3D structure

F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero and M. Black, 
*Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image*, ECCV 2016

A. Kanazawa, J. Zhang, P. Felsen, J. Malik, 
*Learning 3D Human Dynamics from Video*, CVPR 2019
Learning skills from video

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
Learning skills from video

X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, S. Levine, SFV: Reinforcement Learning of Physical Skills from Videos, SIGGRAPH Asia 2018
How can we move forward?

• Develop the next generation of architectures
• Go beyond classification
  • “Rich” prediction tasks
  • Generation
Generation

- State of the art: StyleGAN

Generation

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

Source: D. Marr, Vision, 1982
Generation

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

• BigGAN proves that we don’t?

A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, ICLR 2019
Generation

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

• Or maybe we do?

A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, ICLR 2019
Generation

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

Generation

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

Maybe we don’t?

A. Jahanian, L. Chai, and P. Isola. On the "steerability" of generative adversarial networks. ICLR 2020
Generation

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

  Or maybe we do?

T. Nguyen-Phuoc et al., *HoloGAN: Unsupervised Learning of 3D Representations From Natural Images*, 2019
Possible ways forward

• Develop the next generation of architectures
• Go beyond classification
  • “Rich” prediction tasks
  • Generation
• Work on integrating discriminative and generative models
Integrating discriminative and generative models?

- **Discriminative model**
  - Label

- **Generative model**
  - Noise/latent variable

**Encoder-decoder architecture**
Integrating discriminative and generative models?

Discriminative model

Label

Noise/latent variable

Generative model

GAN
Integrating discriminative and generative models?

What else is there?
Integrating discriminative and generative models?

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

![Image of hallucination process](image1.png)

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)](https://example.com)

![Image of high-quality images](image2.png)

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

[Yin et al. (2019)](https://example.com)
Integrating discriminative and generative models?

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

**Denton et al. (2019)**

**Yu and Grauman (2019)**
Possible ways forward

• Develop the next generation of architectures
• Go beyond classification
  • “Rich” prediction tasks
  • Generation
• Work on integrating discriminative and generative models
• Move away from full supervision
Self-supervised learning

- For still image recognition tasks, self-supervision is more cumbersome (so far) as getting enough supervised training data.
- However, it is very attractive (and possibly unavoidable) for video, audio, and sensorimotor learning.

Figure source: G. Larsson, M. Maire, and G. Shakhnarovich, *Colorization as a Proxy Task for Visual Understanding*, CVPR 2017
Cross-modal self-supervision

Self-supervision from future prediction

L. Pinto and A. Gupta, [Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours](https://ieeexplore.ieee.org/document/7486949), ICRA 2016

C. Finn and S. Levine, [Deep Visual Foresight for Planning Robot Motion](https://ieeexplore.ieee.org/document/7981652), ICRA 2017
Self-supervision from future prediction

• Curiosity is also about predicting the effects of one’s actions!

Possible ways forward

• Develop the next generation of architectures
• Go beyond classification
  • “Rich” prediction tasks
  • Generation
• Work on integrating discriminative and generative models
• Move away from full supervision
• Focus on embodied vision, sensorimotor learning
Outline

• Vision

• Embodied learning (RL, robotics)
Embodied platforms

- Simulation: [AI2Thor](https://ai2thor.com), [Habitat](https://habitat-ai.github.io)

- Real robots: [PyRobot](https://pyrobot.github.io)

- Robot on your smartphone: [OpenBot](https://www.openbot.io)
Can we trust simulators?

Can we scale up robot learning?

Figure 1. Our large-scale data collection setup, consisting of 14 robotic manipulators. We collected over 800,000 grasp attempts to train the CNN grasp prediction model.

S. Levine et al., Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection, International Symposium on Experimental Robotics, 2017
What internal representations are needed for embodied tasks?

“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.”

Multiple agents and communication
Hide and seek

B. Baker et al., Emergent tool use from multi-agent autocurricula, ICLR 2020

https://openai.com/blog/emergent-tool-use/
Emergence of language

In the above step-by-step run, at t=0 the red agent says a word corresponding to the red landmark (center right), then at t=1 says a word that is equivalent to ‘Goto’, then in t=2 says ‘green-agent’. The green-agent hears its instructions and immediately moves to the red landmark.

[Video]

https://openai.com/blog/learning-to-communicate/

A. Mordatch and P. Abbeel, Emergence of Grounded Compositional Language in Multi-Agent Populations, AAAI 2018
RL for negotiation

Figure 1: A dialogue in our Mechanical Turk interface, which we used to collect a negotiation dataset.

M. Lewis et al., Deal or No Deal? End-to-End Learning for Negotiation Dialogues, EMNLP 2017
**Figure 5:** Example of a model stubbornly holding out for maximum points in a dialogue.

**Figure 6:** Example of model compromising.

**Figure 7:** Dialogue in which the model's initial interest in the valueless books allows it to compromise while achieving a maximum score.
Scary AI?

Results are shown in Table 3, and show that the simple LIKELIHOOD model produces the most human-like responses, and the alternative training and decoding strategies cause a divergence from human language. Note however, that this divergence may not necessarily correspond to lower quality language—it may also indicate different strategic decisions about what to say. Results in §6.4 show all models could converse with humans.

Bob: I can i everything else
Alice: balls have zero to me to me to me to me to me to me to me to me to me to me to me to me to me to me to me
Bob: you i everything else
Alice: balls have a ball to me to me to me to me to me to me to me to me to me
Outline

• Vision
• Embodied learning (RL, robotics)
• Language
## Large-scale language models

<table>
<thead>
<tr>
<th>Model Alias</th>
<th>Org.</th>
<th>Article Reference</th>
</tr>
</thead>
</table>
| ULMfit      | fast.ai  | Universal Language Model Fine-tuning for Text Classification  
|             |          | Howard and Ruder                                       |
| ELMo        | AllenNLP | Deep contextualized word representations               
|             |          | Peters et al.                                          |
| OpenAI GPT  | OpenAI   | Improving Language Understanding by Generative Pre-Training  
|             |          | Radford et al.                                         |
| BERT        | Google   | BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding  
|             |          | Devin et al.                                            |
| XLM         | Facebook | Cross-lingual Language Model Pretraining               
|             |          | Lample and Conneau                                       |
Should we be scared of GPT-3?

See also:

Language: The importance of grounding

• *Meaning* cannot be learned from *form* alone – it requires knowing about the relationship between language and the outside world
Further challenges

• Reasoning
• Memory
• Lifelong learning
Parting thoughts

• Embodied learning is the way to go! However...
• 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