Unsupervised Learning (Learning in the Real World)

Prof. Adriana Kovashka
University of Pittsburgh
December 4, 2023
How to learn from weak supervision

The elephant are about to march through them. The spiders themselves have a span as wide as a

But the love serenade is over once a dog arrives.

Australian camels appear sick and emaciated.

Tigers are one of the few cats that actually enjoy swimming.

Male koalas play no role in parenting.

About 50 animals have died in just three months, including this adult orangutan on the day we

Unlike mechanics, langurs are the friends of spotted deer.

There's a turf war going on and the koalas are losing. (dog)

The mayor has declined offers of assistance and expert advice from animal welfare groups. (elephant)
How to recognize objects in new modalities

**Figure 6: Qualitative Results.** YOLO running on sample artwork and natural images from the internet. It is mostly accurate although it does think one person is an airplane.

How to use models across countries

(a) Geographic bias manifested in proposed GeoNet dataset

(b) Unsupervised domain adaptation does not suffice on GeoNet
How to integrate modalities (audio)

Prior work: strong supervision with object detectors

Category label: “Guitar”
Audio waveform
Video with detected objects

Audio separation
Separated audio for the detected object

Ours: self-supervised without object detectors

Natural language query: “person playing a guitar”
Audio waveform
Video

Audio separation and localization
Separated audio and localized regions

Figure 1. We propose to separate and localize audio sources based on a natural language query, by learning to align the modalities on completely unlabeled videos. In comparison, prior audio-visual sound separation approaches require object label supervision.

How to represent everyday activities

Figure 1. Ego4D is a massive-scale egocentric video dataset of daily life activity spanning 74 locations worldwide. Here we see a snapshot of the dataset (5% of the clips, randomly sampled) highlighting its diversity in geographic location, activities, and modalities. The data includes social videos where participants consented to remain unblurred. See https://ego4d-data.org/fig1.html for interactive figure.

How to perform high-level reasoning

Figure 1. VISPROG is a modular and interpretable neuro-symbolic system for compositional visual reasoning. Given a few examples of natural language instructions and the desired high-level programs, VISPROG generates a program for any new instruction using in-context learning in GPT-3 and then executes the program on the input image(s) to obtain the prediction. VISPROG also summarizes the intermediate outputs into an interpretable visual rationale (Fig. 4). We demonstrate VISPROG on tasks that require composing a diverse set of modules for image understanding and manipulation, knowledge retrieval, and arithmetic and logical operations.

How to understand media persuasion

Fig. 1: Example advertisements from our dataset that require challenging visual recognition and reasoning. Despite the potential applications of understanding the messages of ads, this problem has not been tackled in computer vision.

How to generate arbitrary content

Figure 1. Make-A-Scene: Samples of generated images from text inputs (a), and a text and scene input (b). Our method is able to both generate the scene (a, bottom left) and image, or generate the image from text and a simple sketch input (b, center).

How to reason and act

Das et al., CVPR 2018
How to use language models for robotics tasks

Figure 1: LLMs have not interacted with their environment and observed the outcome of their responses, and thus are not grounded in the world. SayCan grounds LLMs via value functions of pretrained skills, allowing them to execute real-world, abstract, long-horizon commands on robots.

Plan for this last lecture

• Weakly-supervised learning
• Language-supervised learning
• Self-supervised learning
• Domain adaptation
• Generation
• 3D vision
Weakly-supervised learning
Manual supervision for object recognition

Berman et al., What’s the Point: Semantic Segmentation with Point Supervision, ECCV 16

Weak supervision
Lower degree (or cheaper) annotation at train time than the required output at test time
Manual supervision for object recognition

- **target**
- **pixel label**
- **bounding box**
- **image label**

Regular/Standard supervision

**Weak supervision**

**Strong supervision**
Standard supervised object detection

Training images

Ground-truth labels

Object detection model
Weakly supervised object detection (WSOD)

Training images

Ground-truth labels

What can we say at minimum?
1- When image is positive, at least one object instance from target category is present
2- When image is negative, no object instance from target category is present

Assumptions
1- There exists a set of features present in positive images and absent in negative images
2- The same features are only present on the target object instances
Ambiguity in defining commonality

- Parts

Question: What is a person?

a) Face
b) Face + upper body
c) Face + whole body
Ambiguity in defining commonality

- **Context**

**Question:** What is a motorbike?

a) Motorbike + Person
b) Person
c) Motorbike + Motorbike
d) Motorbike 😊
Multiple-instance learning (MIL)

Dietterich et al. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence

Positive bags

Negative bags

bags = images
instances = $m$ windows per image

Finally, the aggregated image-level prediction is computed as follows, where greater values of $\hat{p}_c \in [0, 1]$ mean higher likelihood that $c$ is present in the image:

$$\hat{p}_c = \sigma \left( \sum_{i=1}^{m} p_{i,c}^{\text{det}} o_{i,c}^{\text{cls}} \right)$$  \hspace{1cm} (3)

Assuming the label $y_c = 1$ if and only if class $c$ is present, the multiple instance detection loss used for training the model is defined as:

$$L_{\text{mid}} = - \sum_{c=1}^{C} \left[ y_c \log \hat{p}_c + (1 - y_c) \log (1 - \hat{p}_c) \right]$$  \hspace{1cm} (4)

[Blaschko NIPS 10, Cinbis CVPR 14, Deselaers ECCV 10, Nguyen ICCV 09, Bilen BMVC 11, Russakovsky ECCV 12, Siva ICCV 11, Siva ECCV 12, Song NIPS 14, Song ICML 14, Bilen BMVC 14]

Adapted from Vitto Ferrari, Hakan Bilen, equations from Ye et al., Cap2Det
Class activation maps

Language-supervised learning
Learning object detectors from captions

Ye et al., ICCV 2019
Learning to amplify weak caption signal

- Use pseudo labels in free-form text as supervision
- Label inference module performs basic reasoning based on the textual context
- Multiple instance detection module predicts detection/classification scores from proposal features
Weakly supervised object detection results

- Caption supervision model comparable to one trained with image-level labels
- Other ways of obtaining pseudo labels are inferior
- Results (and text classifier) generalize to other datasets

Ye et al., ICCV 2019
Localization from sound

Harwath et al., “Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input”, ECCV 2018
Localization from sound

Fig. 7: On the left are shown two images and their speech signals. Each color corresponds to one connected component derived from two matchmaps from a fully random MISA network. The masks on the right display the segments that correspond to each speech segment. We show the caption words obtained from the ASR transcriptions below the masks. Note that those words were never used for learning, only for analysis.

Harwath et al., “Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input”, ECCV 2018
Detection from documentaries

Chen et al., “Discover and Learn New Objects from Documentaries”, CVPR 2017
Learning from Narrated Videos

Jean-Baptiste Alayrac
jbalayrac.com

3rd Workshop on YouTube-8M Large-Scale Video Understanding
28/10/2019
What are instructional videos?

- Depict complex, goal-oriented human activities (e.g. how to change a car tire)
- Multimodal: video and language
- Can be obtained at scale (e.g. on YouTube), without manual annotation
The HowTo100M dataset in numbers

- 23K human tasks scrapped from WikiHow
- 1.2M unique YouTube videos (duration 15 years)
- 136M clips with narration transcribed into text (mostly from ASR)
- Larger than any existing manually annotated captioning dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Clips</th>
<th>Captions</th>
<th>Videos</th>
<th>Duration</th>
<th>Source</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charades [48]</td>
<td>10k</td>
<td>16k</td>
<td>10,000</td>
<td>82h</td>
<td>Home</td>
<td>2016</td>
</tr>
<tr>
<td>MSR-VTT [58]</td>
<td>10k</td>
<td>200k</td>
<td>7,180</td>
<td>40h</td>
<td>Youtube</td>
<td>2016</td>
</tr>
<tr>
<td>YouCook2 [67]</td>
<td>14k</td>
<td>14k</td>
<td>2,000</td>
<td>176h</td>
<td>Youtube</td>
<td>2018</td>
</tr>
<tr>
<td>EPIC-KITCHENS [7]</td>
<td>40k</td>
<td>40k</td>
<td>432</td>
<td>55h</td>
<td>Home</td>
<td>2018</td>
</tr>
<tr>
<td>DiDeMo [15]</td>
<td>27k</td>
<td>41k</td>
<td>10,464</td>
<td>87h</td>
<td>Flickr</td>
<td>2017</td>
</tr>
<tr>
<td>M-VAD [52]</td>
<td>49k</td>
<td>56k</td>
<td>92</td>
<td>84h</td>
<td>Movies</td>
<td>2015</td>
</tr>
<tr>
<td>MPII-MD [43]</td>
<td>69k</td>
<td>68k</td>
<td>94</td>
<td>41h</td>
<td>Movies</td>
<td>2015</td>
</tr>
<tr>
<td>ANet Captions [26]</td>
<td>100k</td>
<td>100k</td>
<td>20,000</td>
<td>849h</td>
<td>Youtube</td>
<td>2017</td>
</tr>
<tr>
<td>TGIF [27]</td>
<td>102k</td>
<td>126k</td>
<td>102,068</td>
<td>103h</td>
<td>Tumblr</td>
<td>2016</td>
</tr>
<tr>
<td>LSMDC [44]</td>
<td>128k</td>
<td>128k</td>
<td>200</td>
<td>150h</td>
<td>Movies</td>
<td>2017</td>
</tr>
<tr>
<td>How2 [45]</td>
<td>185k</td>
<td>185k</td>
<td>13,168</td>
<td>298h</td>
<td>Youtube</td>
<td>2018</td>
</tr>
<tr>
<td><strong>HowTo100M</strong></td>
<td><strong>136M</strong></td>
<td><strong>136M</strong></td>
<td><strong>1.221M</strong></td>
<td><strong>134,472h</strong></td>
<td>Youtube</td>
<td><strong>2019</strong></td>
</tr>
</tbody>
</table>
How to collect HowTo100M?

Step 1: WikiHow

Result: list of 130k tasks

... How to be healthy
How to cook quinoa in a Rice Cooker
How to Sew an Apron
How to Break a Chain
How to April Fool your Girlfriend
...

Annotation cost: 0
How to collect HowTo100M?
Step 2: Filter task by verb to keep visual tasks

Result: list of 23k tasks

- How to Be healthy
- How to Cook quinoa in a Rice Cooker
- How to Sew an Apron
- How to Break a Chain
- How to April Fool your Girlfriend

Annotation cost: 8 hours for Antoine
How to collect HowTo100M?

Step 3: YouTube queries for videos with captions

Result: 1.2M unique videos

Annotation cost: 0
How to collect HowTo100M?
Step 4: Create clips

Result: 136M narrated clips

Annotation cost: 0
Learning a visual-text embedding on HowTo100M

Pre-trained word2vec word embeddings (dim=300) (No stop words)

fill the pot with water → fill → pot → water → FFL → MaxPool + FC → Text EMBD

Pre-trained visual features (frozen) → FC → Visual EMBD

S=Dot product

Learning a visual-text embedding on HowTo100M

\[ S_{i,j} = S(X_i, Y_j) \text{ (dot product)} \]

\[ \forall (i, j), \ j \neq i, S_{i,i} > S_{i,j}, S_{i,i} > S_{j,i} \]

\[ L = \frac{1}{B} \sum_{i=1}^{B} \sum_{j \neq i} \left[ \max(0, m + S_{i,j} - S_{i,i}) + \max(0, m + S_{j,i} - S_{i,i}) \right] \]

...fill pot water...

...fill pot water...

...these nice plants...
Evaluation procedure

- Text to video retrieval: YouCook2, MSRVTT, LSMDC

- Answering the phone

- Action localization: CrossTask
  - loose bolt
  - jack car
  - remove wheel
Within domain: YouCook2 retrieval (YouTube cooking videos)

YouCook2 (R@10)

- Random
- Trained on HowTo100M
- Trained on YouCook2
- PTHowTo100M + Finetune YouCook2

Jean-Baptiste Alayrac
Self-supervised learning
Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei Efros and Abhinav Gupta

ICCV 2015
ImageNet + Deep Learning

- Image Retrieval
- Detection (RCNN)
- Segmentation (FCN)
- Depth Estimation
- ...

Do we even need semantic labels?

Do we need this task?

Context as Supervision
[Collobert & Weston 2008; Mikolov et al. 2013]
Context Prediction for Images

Semantics from a non-semantic task

Relative Position Task

8 possible locations

What is learned?

<table>
<thead>
<tr>
<th>Input</th>
<th>Ours</th>
<th>ImageNet AlexNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Input Image" /></td>
<td><img src="image2" alt="Ours Image" /></td>
<td><img src="image3" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image4" alt="Input Image" /></td>
<td><img src="image5" alt="Ours Image" /></td>
<td><img src="image6" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image7" alt="Input Image" /></td>
<td><img src="image8" alt="Ours Image" /></td>
<td><img src="image9" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image10" alt="Input Image" /></td>
<td><img src="image11" alt="Ours Image" /></td>
<td><img src="image12" alt="ImageNet AlexNet Image" /></td>
</tr>
<tr>
<td><img src="image13" alt="Input Image" /></td>
<td><img src="image14" alt="Ours Image" /></td>
<td><img src="image15" alt="ImageNet AlexNet Image" /></td>
</tr>
</tbody>
</table>

Pre-Training for R-CNN

1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

Pre-train on relative-position task, w/o labels


[Girshick et al. 2014]
VOC 2007 Performance
(pretraining for R-CNN)

% Average Precision

- ImageNet Labels: 54.2
- Relative position: 46.3
- No Pretraining: 40.7

Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert
ECCV 2016
Fig. 1: (a) A video imposes a natural temporal structure for visual data. In many cases, one can easily verify whether frames are in the correct temporal order (shuffled or not). Such a simple sequential verification task captures important spatiotemporal signals in videos. We use this task for unsupervised pre-training of a Convolutional Neural Network (CNN). (b) Some examples of the automatically extracted positive and negative tuples used to formulate a classification task for a CNN.
Fig. 2: (a) We sample tuples of frames from high motion windows in a video. We form positive and negative tuples based on whether the three input frames are in the correct temporal order. (b) Our triplet Siamese network architecture has three parallel network stacks with shared weights up to the fc7 layer. Each stack takes a frame as input, and produces a representation at the fc7 layer. The concatenated fc7 representations are used to predict whether the input tuple is in the correct temporal order.

Benefit of unsupervised but in-domain training

Table 2: Mean classification accuracies over the 3 splits of UCF101 and HMDB51 datasets. We compare different initializations and finetune them for action recognition.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Initialization</th>
<th>Mean Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF101</td>
<td>Random</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td>(Ours) Tuple verification</td>
<td><strong>50.2</strong></td>
</tr>
<tr>
<td>HMDB51</td>
<td>Random</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td>UCF Supervised</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>(Ours) Tuple verification</td>
<td><strong>18.1</strong></td>
</tr>
</tbody>
</table>

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton

ICML 2020
https://github.com/google-research/simclr
Domain adaptation
Train a classifier on the training data and directly apply it to the test data
Domain Shift

A classifier trained on one domain may perform poorly on another domain.
Semi-supervised vs Unsupervised

- Semi-supervised: Some labeled target data, but not enough to train from scratch

Source data

Target data

Fully-labeled

A few labels
Semi-supervised vs Unsupervised

- Unsupervised: No labels for the target data

Source data

Target data

Fully-labeled
Single vs Multiple Source Domains

• Moving towards domain generalization
Domain Shift

- The domain shift is defined as a difference in the distribution of the source and target samples.
Metric Learning for Domain Adaptation

- Saenko et al., Adapting Visual Category Models to New Domains, ECCV 2010

(a) Domain shift problem  (b) Pairwise constraints  (c) Invariant space

• Colors = domains, shapes = classes
Transformation Learning

• Learn a mapping to a latent space where the distributions are similar
Domain Adversarial Networks

- Ganin & Lempitsky, ICML 2015; Ajakan et al., 2014
  - With domain-invariant features, classifying from which domain a sample comes should be difficult
Generation
Generative Models

Training data $\sim p_{\text{data}}(x)$

Generated samples $\sim p_{\text{model}}(x)$

Want to learn $p_{\text{model}}(x)$ similar to $p_{\text{data}}(x)$
Generative Models

Addresses density estimation, a core problem in unsupervised learning

Several flavors:
- Explicit density estimation: explicitly define and solve for $p_{model}(x)$
- Implicit density estimation: learn model that can sample from $p_{model}(x)$ w/o explicitly defining it

Want to learn $p_{model}(x)$ similar to $p_{data}(x)$
Problem: Want to sample from complex, high-dimensional training distribution. No direct way to do this!

Solution: Sample from a simple distribution, e.g. random noise. Learn transformation to training distribution.

Q: What can we use to represent this complex transformation?

A: A neural network!

Output: Sample from training distribution

Input: Random noise
Training GANs: Two-player game

**Generator network**: try to fool the discriminator by generating real-looking images

**Discriminator network**: try to distinguish between real and fake images

---

Ian Goodfellow et al., “Generative Adversarial Nets”, NIPS 2014
Adversarial Networks Framework

Generator
\[ x \sim G(z) \]

Discriminator
Real vs. Fake

[Goodfellow et al. 2014]
Training GANs: Two-player game

Generator network: try to fool the discriminator by generating real-looking images
Discriminator network: try to distinguish between real and fake images

Train jointly in minimax game

Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[ \mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

Ian Goodfellow et al., “Generative Adversarial Nets”, NIPS 2014

Serena Young
Training GANs: Two-player game

Generator network: try to fool the discriminator by generating real-looking images

Discriminator network: try to distinguish between real and fake images

Train jointly in minimax game

Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[ \mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log (1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

Discriminator output for real data x

Discriminator output for generated fake data G(z)
Training GANs: Two-player game

Ian Goodfellow et al., “Generative Adversarial Nets”, NIPS 2014

Generator network: try to fool the discriminator by generating real-looking images

Discriminator network: try to distinguish between real and fake images

Train jointly in minimax game

Minimax objective function:

\[
\min_{\theta_g} \max_{\theta_d} \left[ \mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log (1 - D_{\theta_d}(G_{\theta_g}(z))) \right]
\]

- Discriminator (\(\theta_d\)) wants to maximize objective such that \(D(x)\) is close to 1 (real) and \(D(G(z))\) is close to 0 (fake)
- Generator (\(\theta_g\)) wants to minimize objective such that \(D(G(z))\) is close to 1 (discriminator is fooled into thinking generated \(G(z)\) is real)
Training GANs: Two-player game

Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[ \mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log (1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

Alternate between:

1. **Gradient ascent** on discriminator
   $$\max_{\theta_d} \left[ \mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log (1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

2. **Gradient descent** on generator
   $$\min_{\theta_g} \mathbb{E}_{z \sim p(z)} \log (1 - D_{\theta_d}(G_{\theta_g}(z)))$$
Training GANs: Two-player game

Generator network: try to fool the discriminator by generating real-looking images

Discriminator network: try to distinguish between real and fake images

After training, use generator network to generate new images
Interpretable Vector Math

Smiling woman  Neutral woman  Neutral man

Samples from the model

Average features, do arithmetic

Radford et al, ICLR 2016

Adapted from Serena Young
Celebrities Who Never Existed

Karras et al., “Progressive Growing of GANs for Improved Quality, Stability, and Variation”, ICLR 2018
StarGAN

Choi et al., “StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation”, CVPR 2018
Edges → Images

Edges from [Xie & Tu, 2015]
Trained on Edges → Images

Data from [Eitz, Hays, Alexa, 2012]
#edges2cats  [Christopher Hesse]

@gods_tail

Ivy Tasi @ivymyt

Vitaly Vidmirov @vvid

https://affinelayerv.com/pixsrv/
Changing artistic style

Input  |  Monet  |  Van Gogh  |  Cezanne  |  Ukiyo-e

Pix2pix / CycleGAN
Changing seasons

Pix2pix / CycleGAN
Denoising diffusion models consist of two processes:

- Forward diffusion process that gradually adds noise to input
- Reverse denoising process that learns to generate data by denoising

Sohl-Dickstein et al., Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML 2015
Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020
Song et al., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021

Slide from: https://cvpr2022-tutorial-diffusion-models.github.io/
Conditioning Generation on Preferences

“GLIGEN: Open-Set Grounded Text-to-Image Generation” Yuheng Li et al., CVPR 2023.
3D Vision