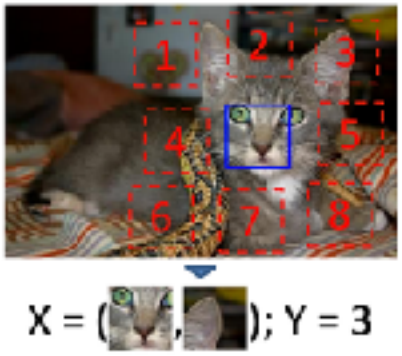
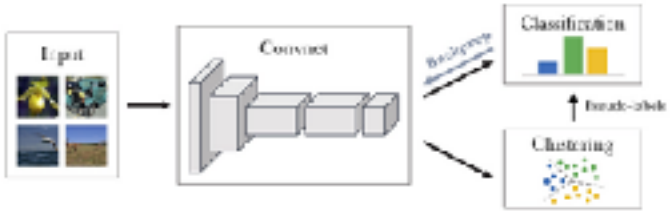
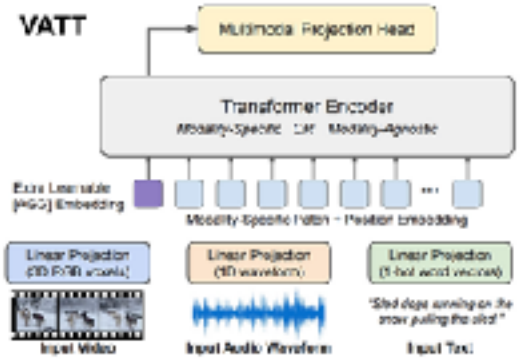
COMP 590/790 Self-Supervised Visual Representation Learning







https://www.gedasbertasius.com/comp590790-22f

Course Introduction

Gedas Bertasius

About Me

- Originally from Lithuania.
- Came to the US to play basketball.
- Got a PhD from UPenn.
- Spent 2 years at Facebook AI Research.
- Joined UNC last summer.

Penn

FB

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Introductions in Canvas

- Name?
- BA / BS / MS / PhD?
- Year?
- What are you excited about in computer vision, self-supervised learning, or AI in general?
- Why are you taking this course?

Plan for Today

- Motivation for the course
- Overview of self-supervised learning
- Course logistics overview

Plan for Today

- Motivation for the course
- Overview of self-supervised learning
- Course logistics overview

Why Self-Supervised Learning?

- Manually labeling visual data is costly and not scalable.

14M Images
21K Object Categories

3,670 hours of daily life videos
text descriptions for all 3,670 hours

Why Self-Supervised Learning?

- Manually labeling visual data is costly and not scalable.

14M Images

21K Object Categories

Took ~22 human years to label

3,670 hours of daily life videos

text descriptions for all 3,670 hours

Why Self-Supervised Learning?

- Manually labeling visual data is costly and not scalable.

14M Images

21K Object Categories

Took ~22 human years to label

3,670 hours of daily life videos

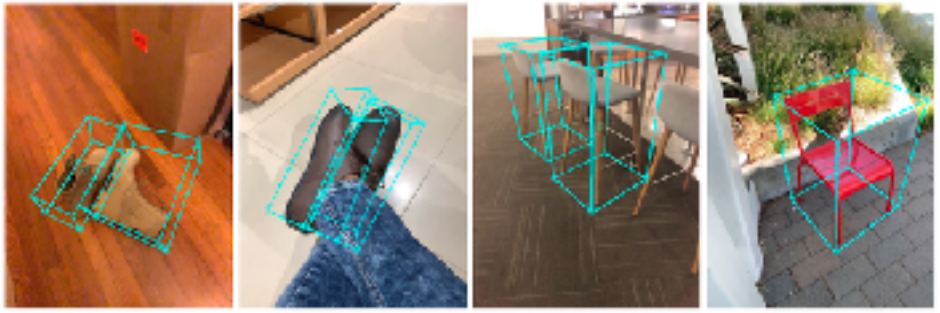
text descriptions for all 3,670 hours

Cost > $1M to label

Why Self-Supervised Learning?

- Manually labeling visual data is costly and not scalable.

Semantic Segmentation



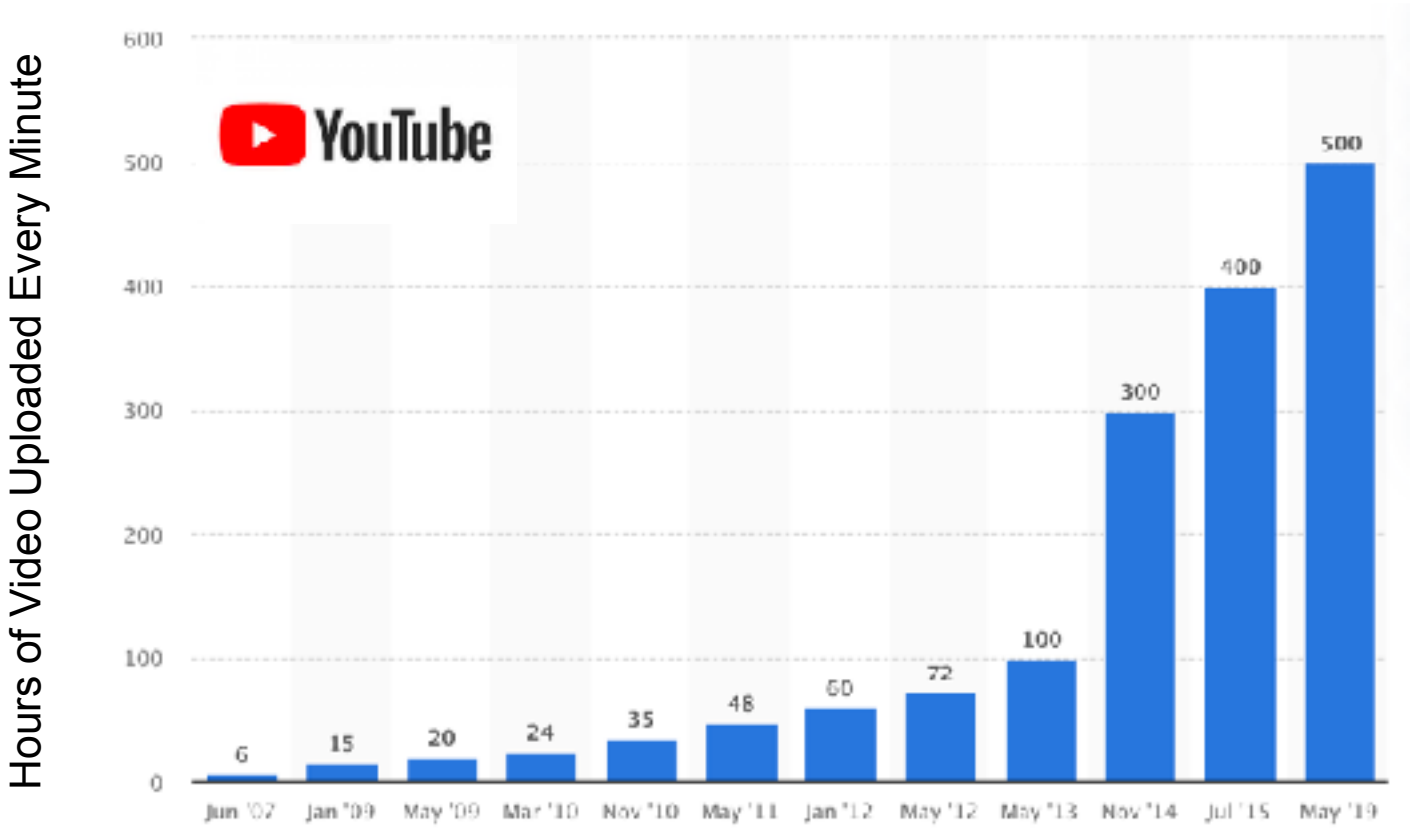
3D Object Detection

3D Point Cloud Semantic Segmentation

Medical Imaging

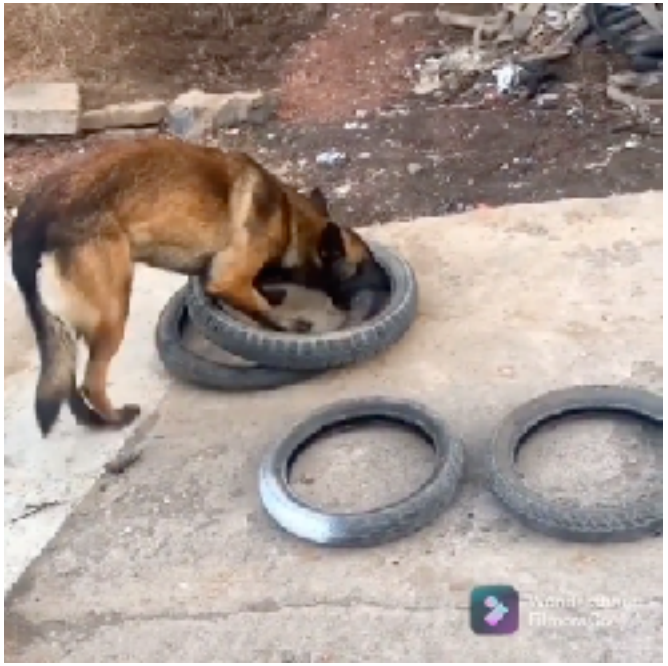
Why Self-Supervised Learning?

- Availability of vast amount of unlabelled image/video data.



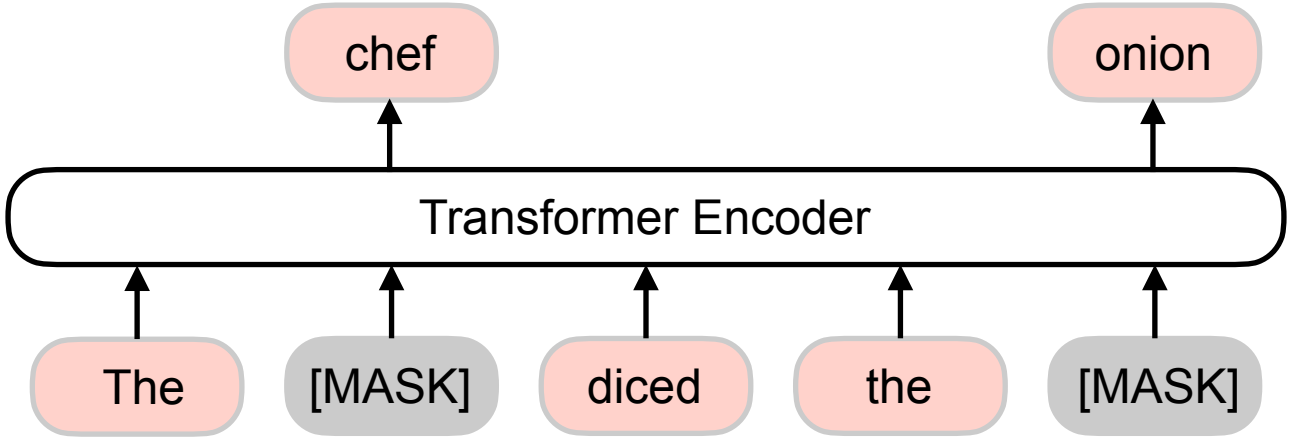
Why Self-Supervised Learning?

- Humans and animals don’t learn as most of our current AI models.



BERT Moment in Computer Vision?

- Self-supervised BERT pretraining has revolutionized the field of natural language processing.



Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, NAACL 2019

Yann Lecun's Cake

How Much Information is the Machine Given during Learning?

Y. LeCun

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

Summary of Motivation

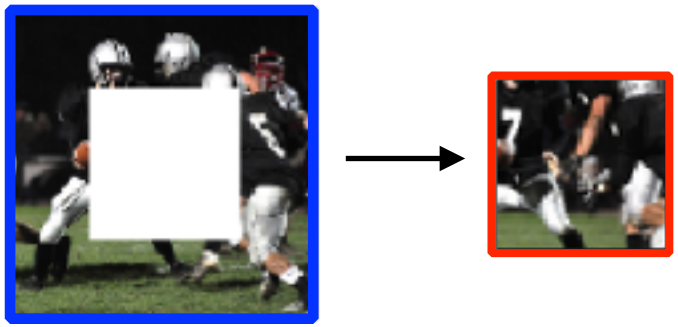
- Manual labeling is costly and unscalable.
- Massive amounts of unlabeled image/video data that we could exploit using self-supervised learning.
- Humans don’t learn by sifting through tons of manually annotated images/videos.
- The hope to discover the BERT moment in CV.

Plan for Today

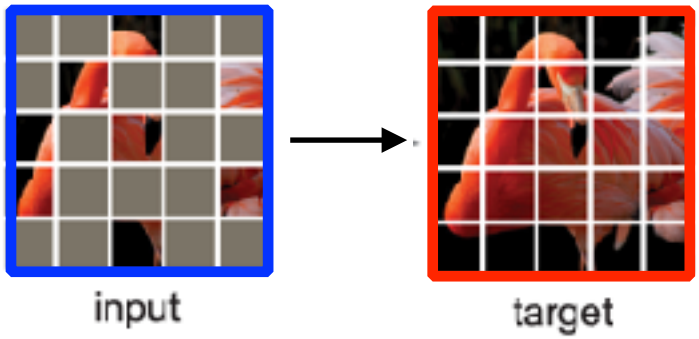
- Motivation for the course
- Overview of self-supervised learning
- Course logistics overview

What is Self-Supervised Learning?

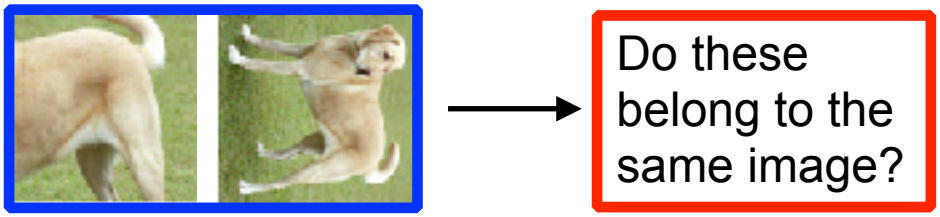
- Learning to predict **unobserved or hidden** part of the input from any **observed or unhidden part** of the input.



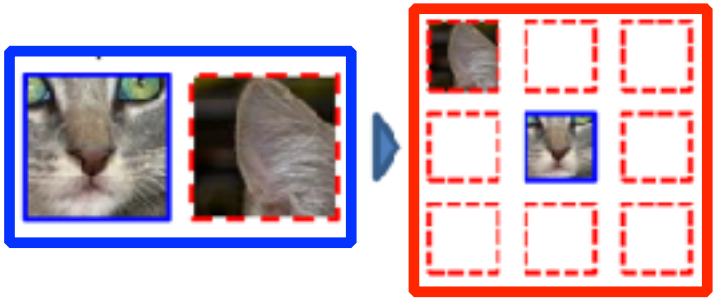
a) Inpainting



b) Reconstruction



c) Similarity-based Matching



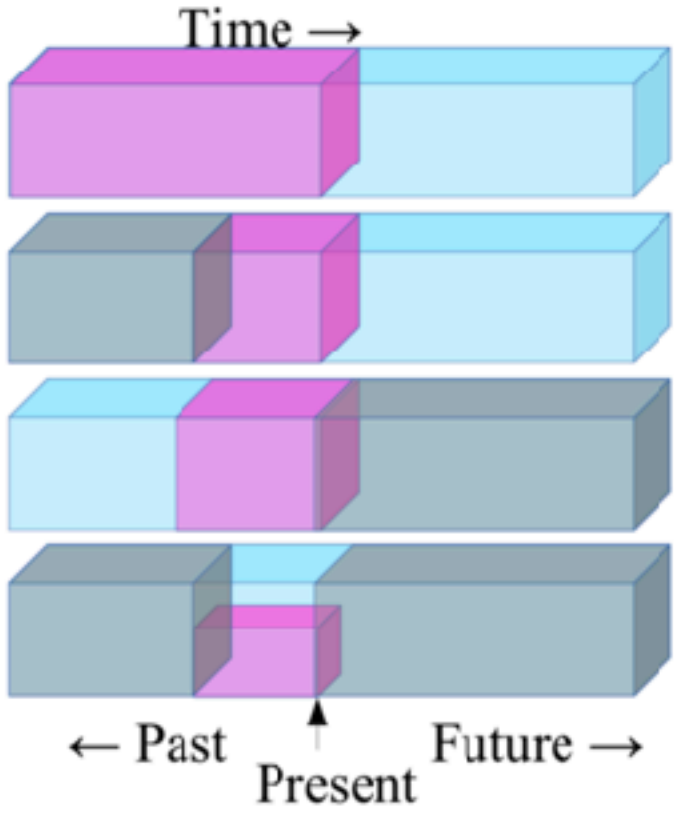
d) Relative Positioning

What is Self-Supervised Learning?

Self-Supervised Learning

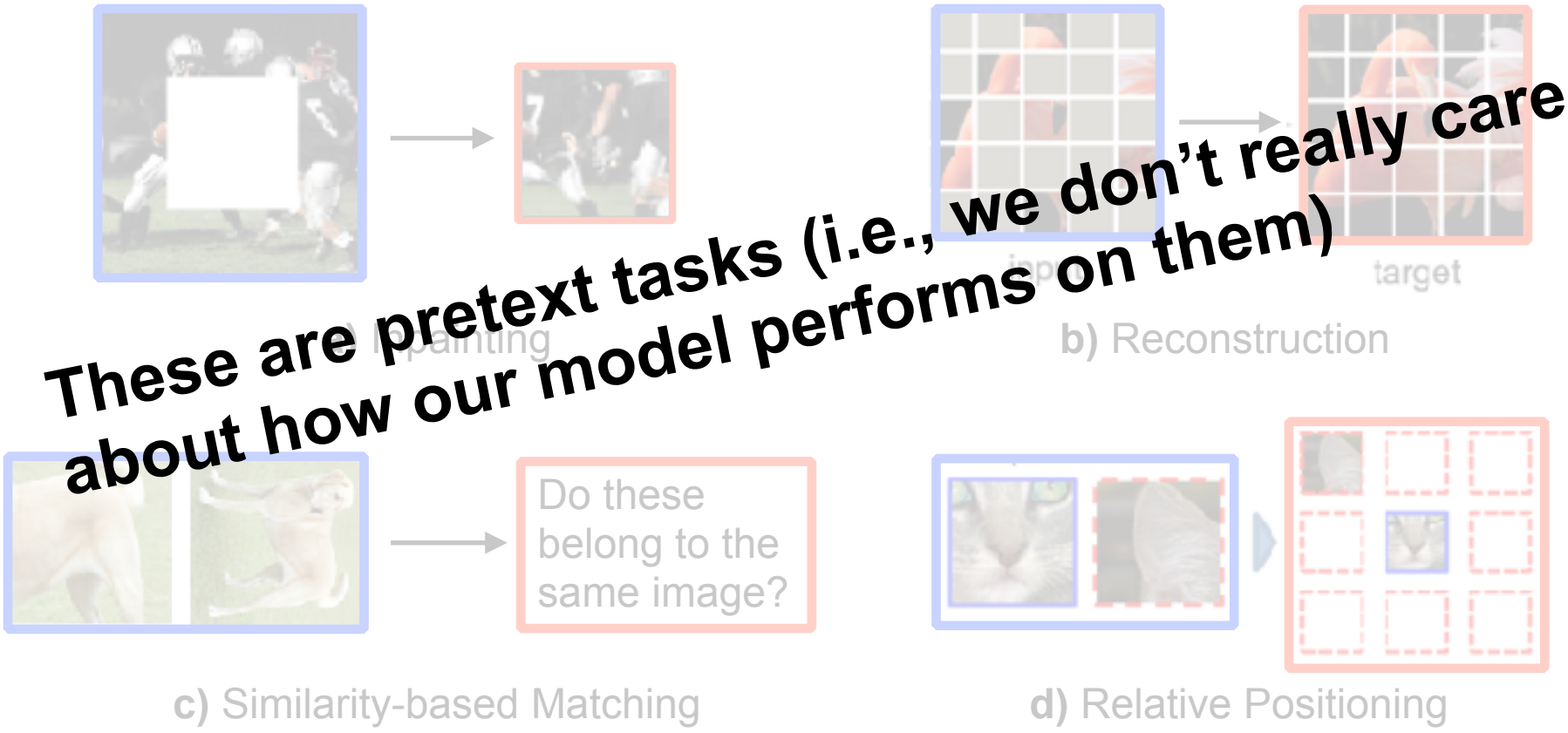
Y. LeCun

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



What is Self-Supervised Learning?

- Learning to predict **unobserved or hidden** part of the input from any **observed or unhidden part** of the input.

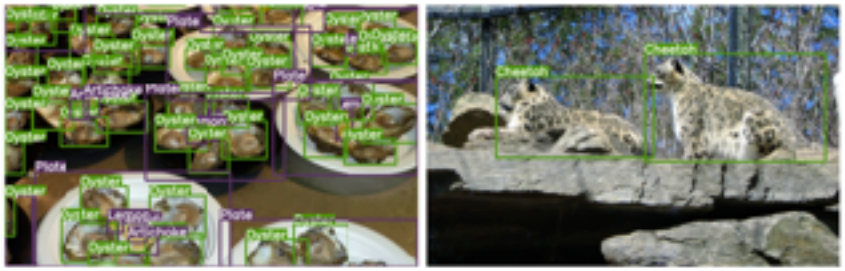


These are pretext tasks (i.e., we don't really care about how our model performs on them)

Generalization

- We want our learned representation to transfer to the tasks that we care about.

a) Image Classification



b) Object Detection

c) Semantic Segmentation

d) Spatiotemporal Action Recognition

Learning Visual Semantics

- Our learned representation should capture high-level semantics (e.g., objects, object parts, scenes, actions, etc.).

Patch A

Patch B

Learning Visual Semantics

- Our learned representation should capture high-level semantics (e.g., objects, object parts, scenes, actions, etc.).

Which location is the correct one ????

Learning Visual Semantics

- Our learned representation should capture high-level semantics (e.g., objects, object parts, scenes, actions, etc.).

Learning Visual Semantics

- Our learned representation should capture high-level semantics (e.g., objects, object parts, scenes, actions, etc.).

Learning Visual Semantics

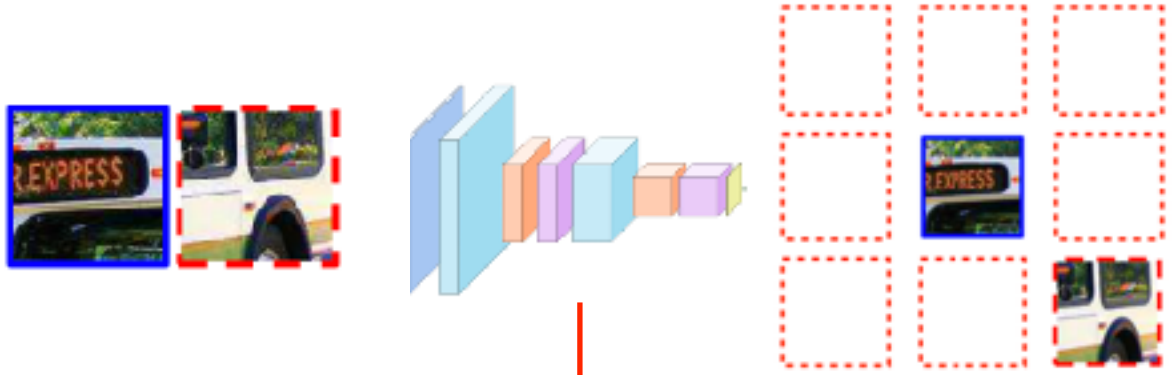
- Our learned representation should capture high-level semantics (e.g., objects, object parts, scenes, actions, etc.).

Self-Supervised Pretraining

Learning Visual Semantics

- Our learned representation should capture high-level semantics (e.g., objects, object parts, scenes, actions, etc.).

Self-Supervised Pretraining

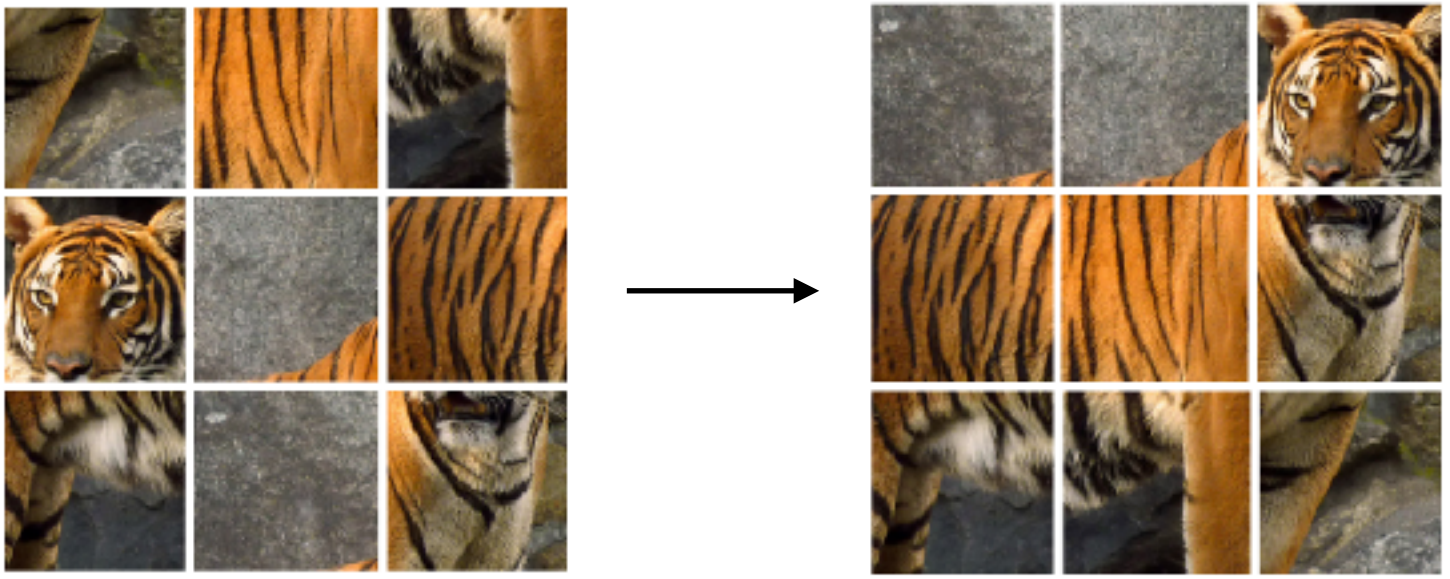


Representation Transfer

a Bus

Learning Visual Semantics

- Our learned representation should capture high-level semantics (e.g., objects, object parts, scenes, actions, etc.).



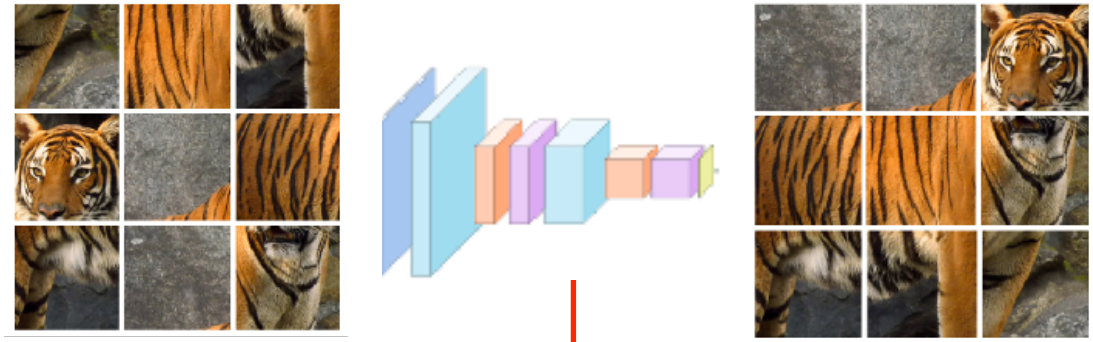
Scrambled Patches of an Image

Unscrambled Image

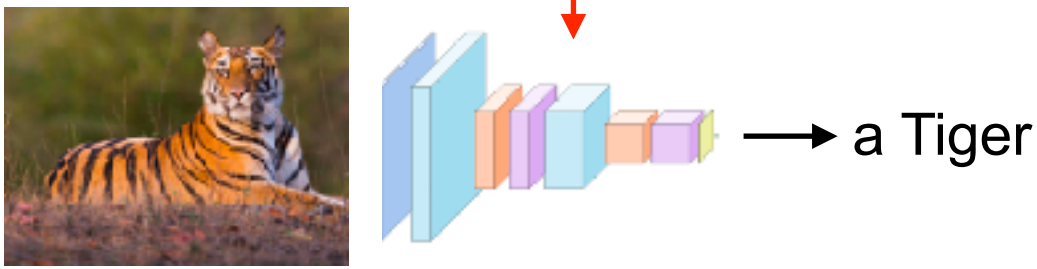
Learning Visual Semantics

- Our learned representation should capture high-level semantics (e.g., objects, object parts, scenes, actions, etc.).

Self-Supervised Pretraining



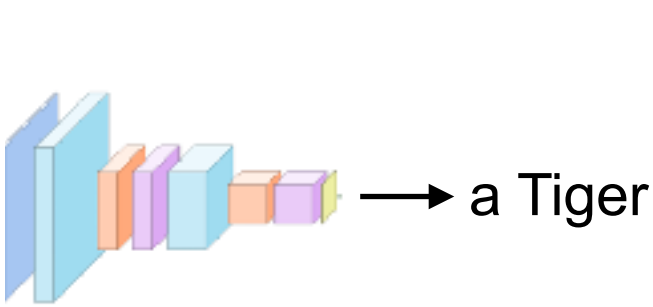
Representation Transfer



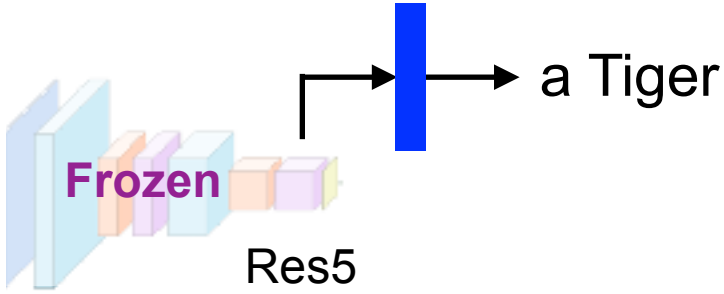
Evaluating Learned Representation

1. Self-Supervised Pretraining

2. Evaluating Learned Representation



a) Full finetuning

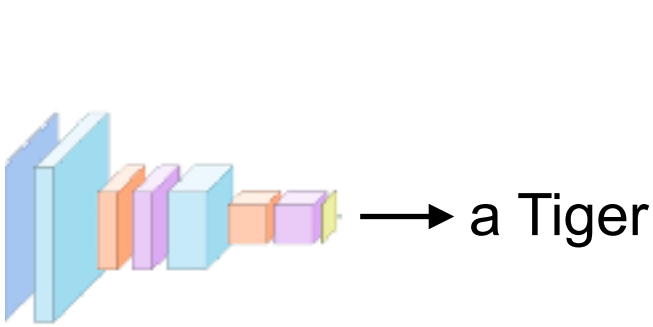


b) Linear Probing

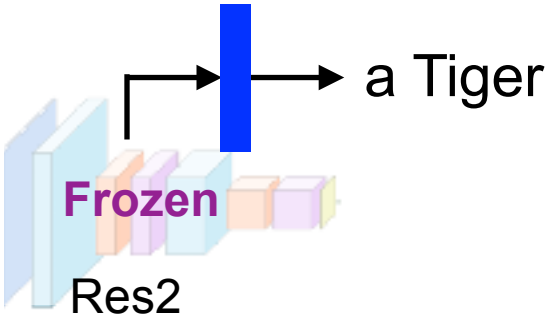
Evaluating Learned Representation

1. Self-Supervised Pretraining

2. Evaluating Learned Representation



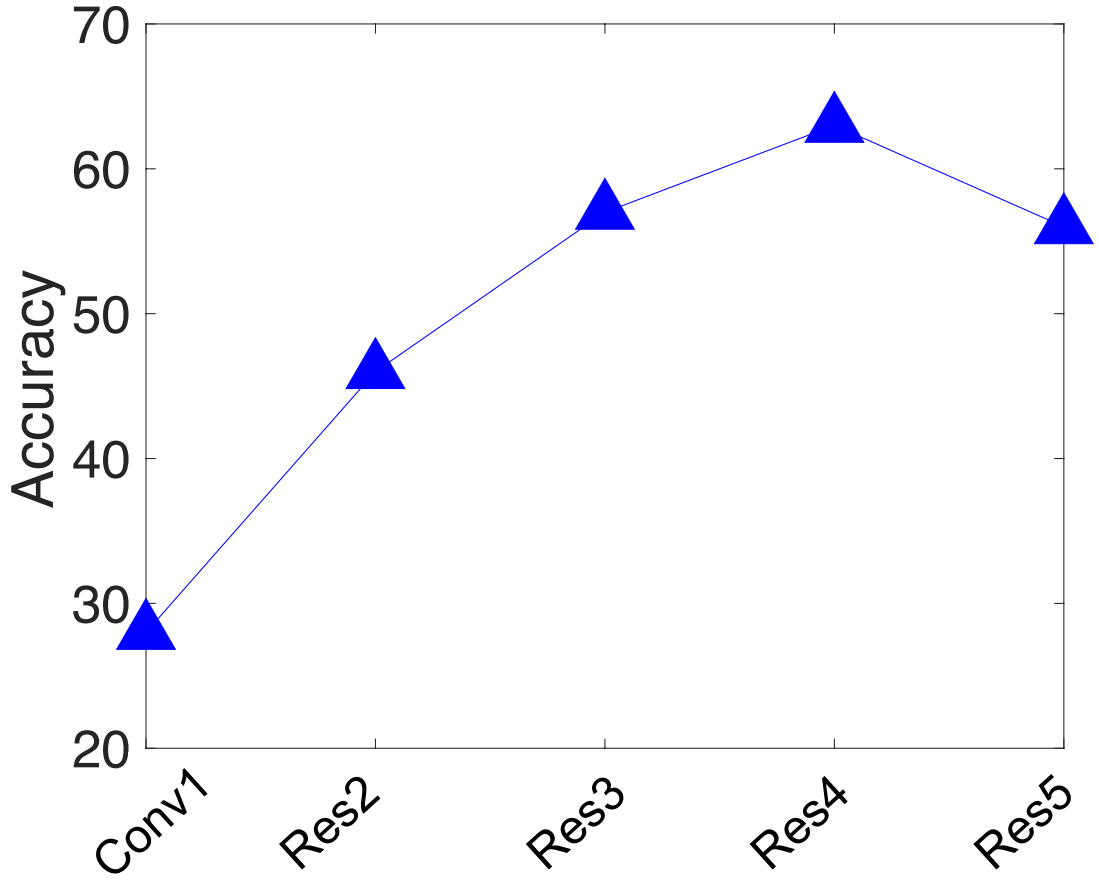
a) Full finetuning



b) Linear Probing

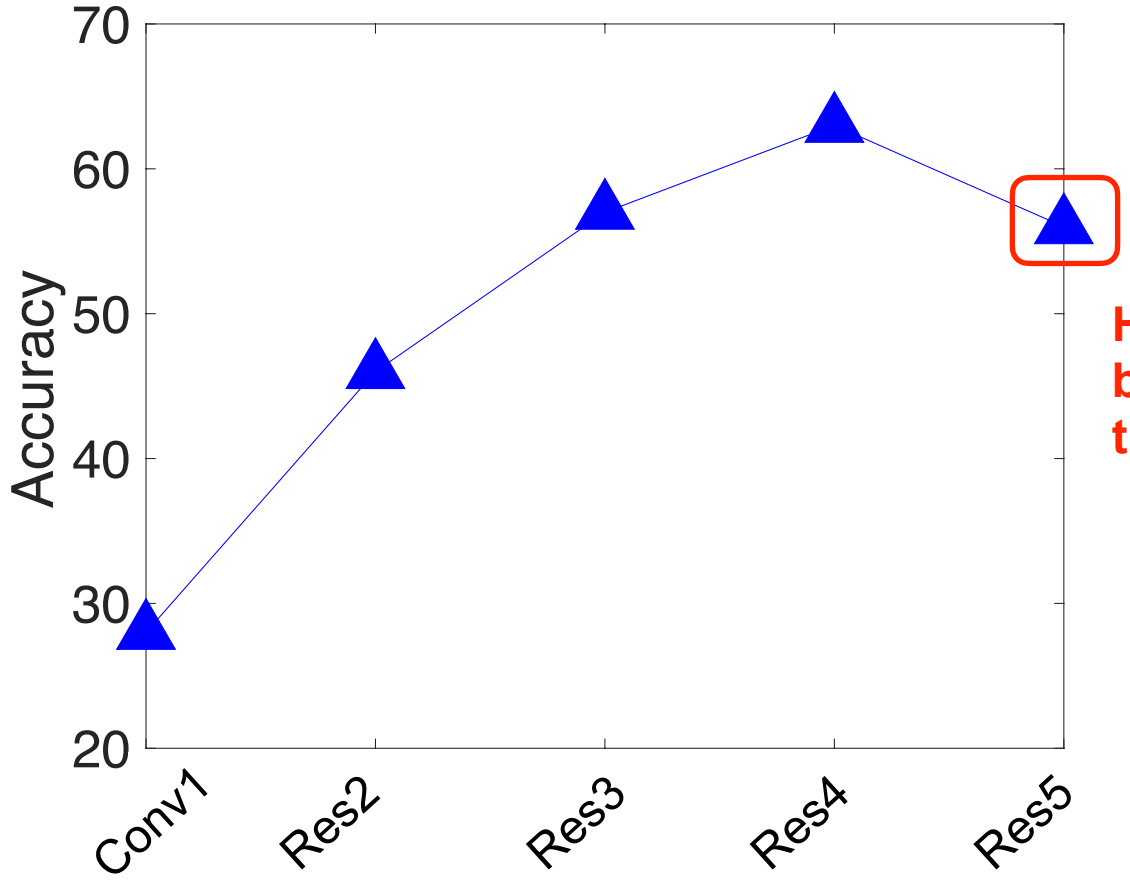
Evaluating Learned Representation

- Image classification accuracy after training a linear classifier using features from different layers of a pretrained network.



Evaluating Learned Representation

- Image classification accuracy after training a linear classifier using features from different layers of a pretrained network.



Higher layers might be too specialized for the pretext task.

Categorization of SSL Methods

1. Pretext Task-based
   - Context Encoders (2016), Colorization (2016), Jigsaw (2017), etc.
2. Generative (GAN-based)
   - BigBiGAN (2019)
3. Clustering
   - DeepCluster (2018), SeLA (2019), SwaV (2020)
4. Contrastive
   - PIRL (2020), MoCo (2020), SimCLR (2020)
5. Distillation
   - BYOL (2020), SimSiam (2020).
6. Generative (transformer-based)
   - iGPT (2020), BEIT (2021), MAE (2021).

Areas in SSL that We Will Cover

- Images
- Videos
- 3D
- Multimodal
- Robotics

Plan for Today

- Motivation for the course
- Overview of self-supervised learning
- Course logistics overview

Course Objectives

- Provide a thorough overview of state-of-the-art in this area of research.
- Learn how to analyze and present research papers.
- Discuss cutting edge research and speculate about future research directions.
- Carry out a semester-long project culminating in a paper of nearly publishable quality.

Prerequisites

- Understanding of fundamental machine learning concepts.
- Experience with deep learning.
- The ability to analyze research papers published in major machine learning and computer vision conferences.
- If you are not sure about your background, check the papers listed on the course website, and see if you would be comfortable presenting them (also come talk to me).

Grading

- Class Participation: 10%
- Written Paper Critiques: 20%
- Paper Presentations: 30%
- Course Project: 40%

Paper Presentations

- Everybody will give two types of presentations:
  1. One 30min or one 45min paper presentation (presented solo or in pairs).
  2. One 20min paper + discussion for a paper battle (presented in groups of 4-5).

- Rehearse your talk to make sure it fits within the time limit listed in the Schedule next to each paper.

- If you need help understanding the paper, send me an email and we'll set up a meeting.

Paper Presentations

- Focus on the most important high-level concepts. No need to present every single technical detail / experiment.

- Your audience should understand:
  1. The research problem.
  2. The motivation of the proposed research.
  3. Any necessary background info.
  4. The main technical details.
  5. The key experimental results.

- Spend time on description of the experiments.

Written Paper Critiques

- The goal is to provide a critical analysis of the paper (positive or negative).
- You will need to submit paper critiques for 10 of my selected papers (each critique worth 2% of the total course grade).
- Each critique graded as Pass or Fail (no detailed feedback).
- Use the template here (also provided in Canvas).
- Please write your critiques independently.
- Upload the critiques in a PDF format on Canvas by 1:00 PM on the day of the class.

QA Prompts for a Paper Discussion

- With each paper critique, you will also include one paper discussion question and your answer to that question.
- We will use your submitted discussion questions for detailed ~30min paper discussions.
- Check out some general info here on how to come up with good questions for a discussion.
- I will read every single one of these so you should come up with meaningful questions.

Detailed Paper Discussions

- For 10 of the selected papers, we will have detailed ~30min paper discussions.
- I will use your submitted QA discussion prompts (from your paper critiques) to compile a set of 6-8 discussion questions.
- We will then break out into small groups where each group will discuss one of the questions among themselves.
- Afterward, we will reconvene to discuss all of the questions together.

Paper Battles

- For 5 paper pairs, we will have paper battles, i.e., detailed head-to-head paper comparisons between two groups of students.
- Assume that the two given papers represent two conference paper submissions.
- However, only one of those papers can be accepted.
- Two groups of students will try to convince the audience that their presented paper should be “accepted”.

Paper Battles (Continued)

- Each group will give a brief 20min overview of their assigned paper (time limit will be strictly enforced).
- Following both presentations, each group will present slides with 3 main reasons why their presented paper is better.
- Afterward, we will have a brief discussion allowing each group to rebut another group's points.
- Lastly, the students in the class who were not presenting will vote on which paper is better (and provide a justification for their vote).

Course Project

- You can propose any project involving self-supervised learning for a CV task of interest to you.
- Review the paper list for inspiration, or come talk to me for topic related suggestions.
- Projects should be completed in groups of 3. If you want to pursue a project individually (e.g., for your dissertation, etc.), please talk to me before doing so.
- If you do not have access to GPUs, send me an email or talk to me after class.
- Start thinking early!

Project Submissions

- Proposal (10% of the total grade):
  - Presentations on **09/19/22** & **09/21/22**.
  - Write-up due **09/25/22.**
  - Overview of your project plan.

- Milestone (10% of the total grade):
  - Presentations on **10/24/22** & **10/26/22**.
  - Write-up due **10/30/22.**
  - A checkpoint to make sure you are making progress.

- Final (20% of the total grade):
  - Presentations on **11/28/22** & **11/30/22**.
  - Write-up due **12/02/22.**
  - Final findings of the project.

- Use the template here (also available on Canvas) for your project write-up.
- Upload the PDF of your slides & write-ups on Canvas in the Assignment section.

Paper Presentation Feedback

- Giving individual paper presentation feedback takes a lot of time.
- Last year, many students ignored my feedback.
- Thus, if you would like to get feedback from me please sign up for a meeting at https://calendly.com/gedasb (either in-person or via zoom)
- You have to sign up before your presentation so that I could take detailed notes on your talk.

Office Hours

- Office hours are available by appointment.

- Please sign up for a meeting at https://calendly.com/gedasb

Canvas

- We will use Canvas for many course-related activities.
- All the announcements will be made on Canvas so please check it regularly.
- You will need to upload your assignments on Canvas.
- The discussion and collaboration pages are enabled on Canvas. Please share any interesting papers, blog posts, or general ideas in the discussion page.
- You can find collaborators for project on Canvas as well.

First Assignment

- The reading list is posted here.

- Select the following:
  1. Seven 30min or 45min papers for standard paper presentations (marked **red** and **purple** in the schedule). Any combo of the papers suffice (e.g., five 30min & two 45min papers, all 30min papers, etc.)
  2. Three 20min papers for paper battles (marked **green** in the schedule).

- Make sure that the papers that you selected will **NOT** be presented by me.

- Rank the papers in each of these lists in descending order of preference (from highest to lowest) and upload them to Canvas **by Sunday, Aug 21st, 11:59 PM** (please include paper IDs in your lists!!).

- I will then update the website with the paper assignments.

Second Assignment

- Complete the paper critique for paper [3] Deep Clustering for Unsupervised Learning of Visual Features.

- Upload it to Canvas by **1 PM on Monday, August 22th**.