TRANSFER LEARNING (RESEARCH)

- A method to use pre-existing ML models as a starting point for our own usecase and dataset. Instead of training models from scratch we use a well- trained model like Resnet18 and ImageNet to adapt to our task.
- Transfer Learning helps to preserve time as we do not need to build a model from scratch as it requires huge amount of resources.
- Steps in transfer learning:
  1) Choose a pre-trained model
  2) Freeze early layers (retain basic knowledge) and fine tune top layers.
  3) Replace top layers that matches our concern
  4) Train data

Some applications of Transfer learning are image classification, object classification, NLP, etc

PYTORCH

- Open-source deep learning framework
- Heavily used in neural networks, CNN's, RNN's
- Convolutional neural networks- specializes in image and video processing
- Recurrent neural networks- for text processing

Everything In pytorch is based on tensor operations. A tensor is a multi-dimensional matrix consisting of same type of data.

```
import torch

# torch.empty(size): uninitiallized
x = torch.empty(1) # scalar
print("empty(1):", x)
x = torch.empty(3) # vector
print("empty(3):",x)
x = torch.empty(2, 3) #matrix
print("empty(2,3):",x)
```

Torch.empty helps us create a tensor, depends on our use case how many dimension and size we need.

- Requires_grad arg: default set to false, need to set to true for calculation of gradients later on. (requires_grade= True).
- Numpy is a fundamental library that supports large multi-dimensional arrays and matrices, heavily used in deep learning.

TRANSFER LEARNING (RESEARCH)

How Pytorch actually works and the math behind it?

- Linear regression: the most basic algorithm in machine learning. Used to predict a value using linear graphs.
- Eg: wanting to predict weight of a person using linear approach **: y=mx+c**
- The goal of linear regression is to minimise mean squared error between predicted and actual values.

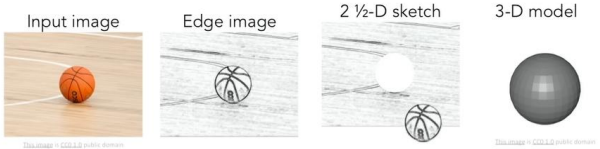
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{pred}_i} - y_{\text{actual}_i})^2$$

  Lower the mse, better will be the model.
- Gradient descent: a method used to minimise loss function, ie finding the best value of m and c in the linear regression equation.
- Linear regression is the foundation for understanding how the models fit the data
- Gradient descent is almost how all the ML models learn.

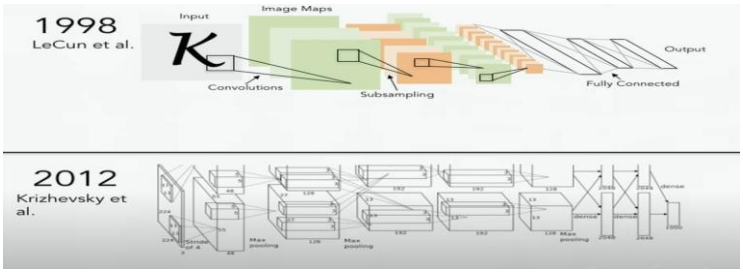
COMPUTER VISION

- From course cs231 we understand how computer vision evolved from history and how we tend to use it right now.
- Primal sketch

  Input image->edge image->2D->3D



- Problems of object recognition:
  1. Image classification
  2. Object detection
  3. Segmentation
  4. Facial recognition, scene understanding
- Convolutional neural networks found a breakthrough in the imageNet challenge helping it perform better than any other algorithm out there.
- Other name for CNN is deep learning using neural networks.



TRANSFER LEARNING (RESEARCH)

Why use **Google Colab**

1. Free Access to GPUs and TPUs

- You get free cloud-based access to powerful NVIDIA GPUs and TPUs.
- Ideal for training deep learning models like CNNs, transformers, etc.
- No need to buy an expensive GPU laptop/PC.

2. Runs in the Cloud

- You don't need to install Python, Jupyter, PyTorch, TensorFlow, etc., locally.
- Just log into your browser, and everything works.
- Your code runs on Google's servers, not your laptop.

3. Jupyter Notebook Interface

- Google Colab is built on Jupyter, so it's very beginner-friendly.
- You can:
    - Write & run Python code
    - Add markdown cells for notes
    - Visualize charts, graphs, and images inline

4. Great for Collaboration

- You can share your notebook just like Google Docs.
- Teammates can comment, suggest, or even run code with you.

5. Seamless with Google Drive

- Your notebooks are saved in your Google Drive.
- Easy to organize, access from anywhere, and back up

1. Linear Regression

What is it?

Linear regression is a supervised learning algorithm used for predicting a continuous value. It finds the best-fit straight line (or hyperplane in higher dimensions) through the data.

Equation:

$y=wx+b$

- x: input features
- w: weight/parameter
- b: bias/intercept
- y: predicted value

Goal:

Find the values of w and b that minimize the error between predictions and actual values.

TRANSFER LEARNING (RESEARCH)

---

2. Loss Function: Mean Squared Error (MSE)
The error is the difference between predicted and actual values.
MSE is the most commonly used loss for regression tasks:

- $$\mathcal{L}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
- - y_i: actual output
  - $\hat{y}_i$: predicted output
  - n: number of data points

---

3. Gradient Descent
Why Gradient Descent?
To minimize the loss function and find the optimal weights and bias.
Idea:
Take small steps in the opposite direction of the gradient of the loss function to reduce it.
Update Rule:

$$w = w - \alpha \cdot \frac{\partial \mathcal{L}}{\partial w}$$

$$b = b - \alpha \cdot \frac{\partial \mathcal{L}}{\partial b}$$

- α = learning rate (step size)
- $\frac{\partial \mathcal{L}}{\partial w}$ = gradient w.r.t. weight

---

How it Works Step-by-Step:

1. Start with random w and b.
2. Compute predictions using current w, b.
3. Calculate the loss using MSE.
4. Compute gradients of loss w.r.t. w and b.
5. Update w and b using gradient descent.
6. Repeat for multiple epochs until loss is minimized.

2. Activation Functions
Why do we need them?

- Without activation functions, a neural network would just be a linear model, no matter how many layers it has.
- Activation functions **introduce non-linearity**, allowing the network to learn more complex patterns.

---

1. Sigmoid Function
Equation:

Equation:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Output Range:

$$(0, 1)$$

TRANSFER LEARNING (RESEARCH)

Pros:

- Smooth gradient
- Good for binary classification (as output layer)

Cons:

- Vanishing gradient problem
- Outputs not centered around zero

2. Tanh (Hyperbolic Tangent)

Equation:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Output Range:

$$(-1, 1)$$

Pros:

- Zero-centered output
- Better than sigmoid in hidden layers

Cons:

- Still suffers from vanishing gradients

3. ReLU (Rectified Linear Unit)

Equation:

Equation:

$$f(x) = \max(0, x)$$

Output Range:

$$[0, \infty)$$

Pros:

- Simple and efficient
- Solves vanishing gradient issue (partially)

Cons:

- Can "die" during training if neurons stop updating (output = 0)

Function	Use Case
Sigmoid	Output layer in binary classification
Tanh	Hidden layers when data is zero-centered
ReLU	Default for hidden layers (fast and effective)

3. Loss Functions

What is a Loss Function?

A loss function measures how far off the model's predictions are from the actual values. It provides a **quantitative measure of error**, which is then minimized using **gradient descent**.

---

1. Mean Squared Error (MSE)

TRANSFER LEARNING (RESEARCH)

Use:

Used in **regression problems**.

Formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

y_i: actual value

$\hat{y}_i$: predicted value

n: number of samples

Pros:

- Simple and widely used
- Penalizes large errors

Cons:

- Sensitive to outliers

2. Cross-Entropy Loss

Use:

Used in **classification problems**.

Formula:

$$\mathcal{L} = -\sum_i y_i \log(\hat{y}_i)$$

y_i: true label (one-hot encoded)

$\hat{y}_i$: predicted probability

Pros:

- Ideal for probabilistic outputs
- Works well with Softmax activation

Cons:

- Output must be log-probabilities or raw logits

3. SVM Loss (Hinge Loss)

SVM LOSS:

Use:

Used in **Support Vector Machines (SVMs)**.

Formula:

$$\mathcal{L}(y, f(x)) = \max(0, 1 - y \cdot f(x))$$

y: actual label (should be +1 or -1)

$f(x)$: predicted score

Pros:

- Focuses on margin maximization

TRANSFER LEARNING (RESEARCH)

- Works well for clear decision boundaries

Cons:

- Doesn't output probabilities

Loss Function	Task Type	Use With
MSE	Regression	Linear regression, DNN
Cross-Entropy	Classification	Softmax output
Hinge (SVM)	Classification	SVM models

1. Forward Pass

Goal: Calculate the output (prediction) of the model given an input.

How It Works:

- Data flows **forward** through each layer of the network.
- Each layer applies:
    - A linear transformation: $z=wx+b$
    - An activation function (e.g., ReLU, Sigmoid)

2. Loss Calculation

TRANSFER LEARNING (RESEARCH)

After the forward pass, we compare the prediction with the true value using a **loss function** like MSE or CrossEntropy.

3. Backward Pass (Backpropagation)

Goal: Compute gradients of the loss w.r.t. weights and biases using **chain rule** of calculus.

How It Works:

- The loss is propagated **backward** from the output layer to each layer.
- Gradients are computed for all parameters (weights & biases).

4. Parameter Update (Gradient Descent)

Once gradients are computed, we update parameters to reduce the loss:

```
optimizer.step()      # updates weights using gradients

optimizer.zero_grad() # clears old gradients before next backward pass
```

FULL EXAMPLE:

```
for epoch in range(n_epochs):

    output = model(inputs)            # forward pass

    loss = loss_fn(output, targets)   # compute loss

    loss.backward()                   # compute gradients

    optimizer.step()                  # update weights

optimizer.zero_grad()                 # reset gradients
```

What Backpropagation Gives Us

- It gives the **gradient** (slope) of the loss w.r.t. each parameter.
- This allows us to **know the direction** in which we need to change weights to reduce the error.

Input ➜ [Linear + Activation] ➜ Output ➜ Loss

▲ |

| ▼

Gradient ← Backward ← Loss Function

TRANSFER LEARNING (RESEARCH)

5. Batch Normalization

Why Do We Need It?

When training deep neural networks:

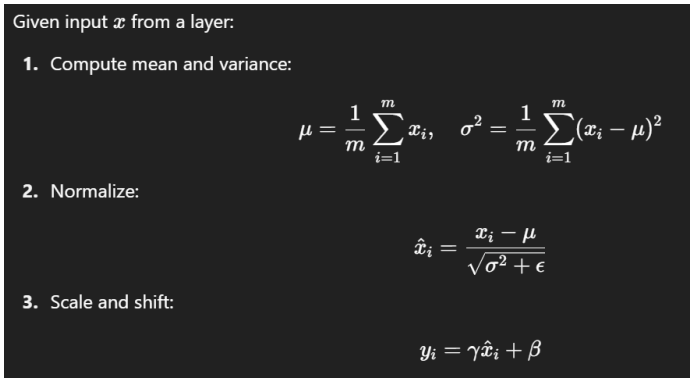
- The distribution of inputs to each layer changes during training.
- This slows down training and makes convergence harder (called **Internal Covariate Shift**).

Batch Normalization solves this by **normalizing** the inputs of each layer.

What It Does:

For each mini-batch during training, it:

1. **Normalizes** the inputs to have mean = 0 and variance = 1
2. **Scales and shifts** using learnable parameters γ\gammaγ (scale) and β\betaβ (shift)



Benefit	Description
Faster Training	Allows higher learning rates
Stability	Reduces exploding/vanishing gradients
Regularization	Acts like a mild regularizer, reduces need for dropout

BatchNorm helps make each layer learn **independently** from changes in previous layers' distributions — **making deep models easier to train**.

6. Transfer Learning

TRANSFER LEARNING (RESEARCH)

What is Transfer Learning?

Transfer Learning is the technique of **using a pre-trained model on a new but related task**.

Instead of training a model from scratch (which needs tons of data and time), we **leverage the knowledge** learned from a model trained on a large dataset (like ImageNet) and **fine-tune it** for our own task.

Common Use Case:

- Use a model trained on **ImageNet** (over 1 million images) for a smaller **image classification** task.

Part	**Role**
Base layers (Frozen)	Extract general features (edges, textures, shapes)
Final layer (Trainable)	Adapt to new task (like classifying cats vs dogs)
You Have...	Recommended Approach
Very small dataset	Freeze base model, train final layer
Medium dataset	Fine-tune top few layers
Large dataset, new task type	Train model from scratch

Key Insight:

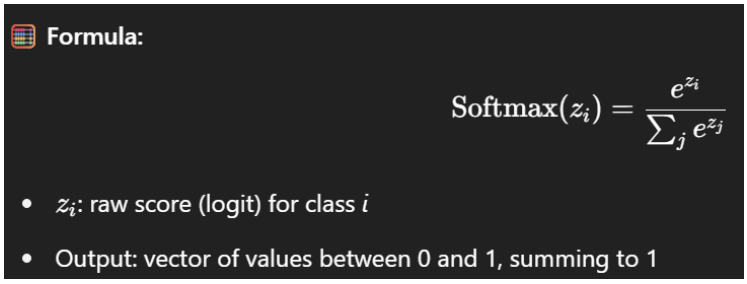
Transfer learning allows models to **generalize from one task to another**, just like how humans use prior knowledge to learn faster in new situations.

7. Softmax vs. SVM (Support Vector Machine)

🔸 What is Softmax?

Softmax is an **activation function** typically used in the final layer of a classification model to produce **probabilities** over classes.

TRANSFER LEARNING (RESEARCH)



Used With:

- **CrossEntropy Loss**

Pros:

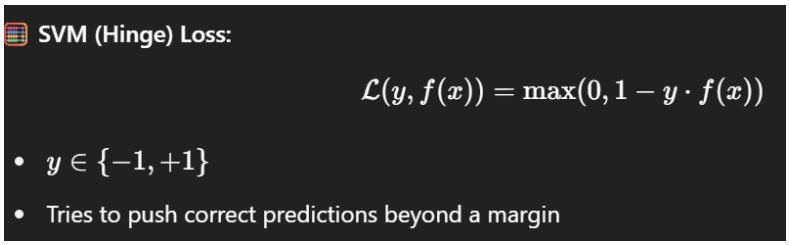
- Provides probabilities (confidence scores)
- Works well for multiclass classification

Cons:

- Can be overconfident
- Can struggle with small datasets unless regularized

🔸 **What is SVM?**

Support Vector Machines are **margin-based classifiers** that aim to **maximize the margin** between data points and the decision boundary.



Pros:

- Works well for small/medium datasets
- Focuses on support vectors → robust to outliers

Cons:

- Doesn't give probabilities
- Slower to train with large datasets

TRANSFER LEARNING (RESEARCH)

Feature	Softmax	SVM
Output	Probabilities	Raw scores (margins)
Loss Function	CrossEntropy Loss	Hinge Loss
Use Case	Neural Networks (deep learning)	Classic ML or shallow networks
Multiclass Support	Built-in	Requires one-vs-rest/one-vs-one
Optimization Goal	Maximize likelihood	Maximize margin
Confidence Output	Yes	No

Key Insight:

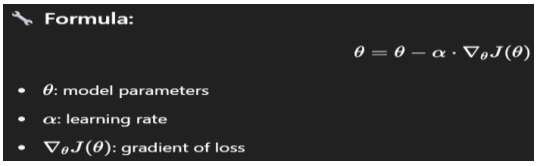
- Use **Softmax + CrossEntropy** in deep learning when you need **probabilities** and multiclass support.
- Use **SVM** when you're focusing on **maximizing margins**, especially in small datasets or classical ML setups.

What Is Optimization in ML?

Optimization refers to the method of **adjusting the model's parameters** (like weights and biases) to **minimize the loss function** — that is, improve predictions.

We use **gradients** (from backpropagation) to decide how to change weights. We apply this after gradients are computed after backprop.

A. **SGD (Stochastic Gradient Descent)**



Characteristics:

- Updates weights after each mini-batch
- Can be noisy → but helps escape local minima

Pros:

- Simple and memory efficient
- Good for large datasets

Cons:

- Needs careful tuning of learning rate
- Can get stuck or oscillate

B. **Momentum (Improved SGD)**

Formula:

$$v_t = \beta v_{t-1} + \alpha \nabla_\theta J(\theta) \quad \theta = \theta - v_t$$

- β: momentum coefficient (e.g., 0.9)

It adds velocity — so updates carry forward some momentum from the previous updates, smoothing out oscillations.

C. **RMSprop (Root Mean Square Propagation)**

Formula:

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma) g_t^2 \quad \theta = \theta - \frac{\alpha}{\sqrt{E[g^2]_t + \epsilon}} \cdot g_t$$

- Keeps a moving average of squared gradients
- Scales learning rate for each parameter

Good for:

- non-stationary objectives

D. **Adam (Adaptive Moment Estimation)**

Combines Momentum + RMSprop

- Maintains both:
  - **Exponential moving average of gradients** (1st moment)
  - **Exponential moving average of squared gradients** (2nd moment)

TRANSFER LEARNING (RESEARCH)

Update Rule:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad \theta = \theta - \alpha \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}$$

Pros:

- Requires less tuning
- Fast convergence
- Works well with sparse gradients

Optimizer	Best Use Case	Key Feature
SGD	Large datasets	Simplicity
Momentum	Faster convergence	Smooths update direction
RMSprop	RNNs, non-stationary objectives	Normalizes gradients
Adam	General-purpose optimizer	Combines RMSprop + Momentum

COMPUTER VISION

1. Image Classification

What is it?

Image classification is the most basic and widely used computer vision task. It involves assigning a **single label** to an **entire image**. For example, if you input an image of a dog, the model classifies it as **"dog"**.

You give it:

- An image (e.g., of a cat)

It gives you:

- A label: "cat"

Real-life Applications:

- Classifying X-ray images as normal or pneumonia.
- Detecting spam in memes (offensive, adult content).
- Animal or plant species classification.

Common Models

1. **ResNet (Residual Network)**
   - Solves the vanishing gradient problem.
   - Allows very deep networks by using **skip connections** (residual blocks).
   - Example: ResNet-50, ResNet-101.
2. **VGG (Visual Geometry Group)**
   - Uses **very deep layers** with small (3×3) filters.
   - Easy to implement, but heavy on computation.
3. **DenseNet**
   - Every layer is connected to all subsequent layers.
   - Helps reuse features and gradients more efficiently.
4. **EfficientNet**
   - Scales depth, width, and resolution systematically.
   - Very efficient and powerful for mobile and cloud.
5. **Vision Transformer (ViT)**
   - Uses self-attention mechanisms instead of convolution.
   - Performs great on large datasets.

TRANSFER LEARNING (RESEARCH)

Evaluation Metrics

1. **Accuracy**
   - Percentage of correct predictions over total predictions.
   - Good for balanced datasets.
2. **Top-k Accuracy**
   - Top-1 accuracy: correct label is the first prediction.
   - Top-5 accuracy: correct label is in the top 5 predicted labels.
3. **Precision / Recall / F1-Score**
   - Especially useful in **imbalanced datasets** (e.g., 90% cats, 10% dogs).
   - **Precision**: How many predicted cats are actually cats?
   - **Recall**: How many actual cats were detected?
   - **F1**: Harmonic mean of precision and recall.
4. **Confusion Matrix**
   - A grid showing where the model confused classes.
   - Helps diagnose specific mistakes (e.g., cat misclassified as dog).

Metric	Formula
Accuracy	$\frac{\sum TP}{\text{Total Samples}}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1 Score	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
Confusion Matrix	Raw counts of TP, FP, FN, etc.

Object Detection

What It Does

Object detection not only tells you **what** is in the image, but also **where** it is.

It gives:

- **Class label** (e.g., "dog")
- **Bounding box** (e.g., x=110, y=65, width=150, height=200)
- **Confidence score** (e.g., 0.89)

Architecture Types

There are **2 main types** of object detectors:

TRANSFER LEARNING (RESEARCH)

---

1. Two-Stage Detectors

Example: Faster R-CNN

How it works:

1. **Stage 1**: Region Proposal Network (RPN) suggests candidate object areas (called Regions of Interest - ROIs).
2. **Stage 2**: Each ROI is classified and refined.

Pros:

- High accuracy
- **Cons**:
- Slower (not ideal for real-time)

---

2. Single-Stage Detectors

Examples: YOLO, SSD

How it works:

- Directly predicts bounding boxes and class probabilities from the image in one go.

Pros:

- Extremely fast (real-time)
  Cons:
- Slightly lower accuracy than two-stage (but newer versions are closing that gap)

YOLO Architecture (Simplified)

1. Divides image into an **S × S grid**
2. Each cell:
   - Predicts **bounding boxes**
   - Predicts **objectness score**
   - Predicts **class probabilities**
3. Combines predictions → outputs final list of objects

In YOLOv5 and v8:

- Everything is learned end-to-end.

TRANSFER LEARNING (RESEARCH)

- Uses anchor boxes and confidence thresholds.
- Non-Max Suppression (NMS) removes overlapping boxes.

	Description
Metric	
IoU	Measures overlap between predicted box & ground truth (0 to 1)
mAP	mean Average Precision — averaged over all classes and IoU thresholds
AP@[IoU]	AP at specific IoU threshold (e.g., 0.5, 0.75)
Precision/Recall	Used to analyze false positives and false negatives
FPS	Speed: How many frames can the model process per second
Term	Meaning
Precision	Of all boxes predicted, how many were correct? (/All Preds)
Recall	Of all actual objects, how many did we detect? (/All GT)
Confidence score	Model's certainty (0 to 1) that this prediction is correct

Eg:

Detected 3 objects:

- person (92%) at [x1, y1, x2, y2]

- dog (88%) at [x1, y1, x2, y2]

- bicycle (79%) at [x1, y1, x2, y2]

Semantic Segmentation

TRANSFER LEARNING (RESEARCH)

What is it?

Semantic segmentation is the process of **classifying every pixel** in an image into a **predefined category**.

Unlike object detection (which gives bounding boxes), semantic segmentation gives a **pixel-level mask**.

Example:

Input image:

A street scene with cars, people, buildings, road

Output:

- Pixels labeled as:
  - road = gray
  - car = blue
  - pedestrian = red
  - building = orange
  - sky = cyan

Every pixel is assigned a **semantic class**, but not a unique object identity. That's what **instance segmentation** does .

Model	Highlights
U-Net	Very popular in medical imaging; uses encoder-decoder with skip connections
DeepLabV3(+):	Uses atrous (dilated) convolutions + pyramid pooling
FCN (Fully Convolutional Network)	First deep learning approach to semantic segmentation
SegFormer	Transformer-based, efficient and accurate

Evaluation Metrics

Metric	Meaning
Pixel Accuracy	% of correctly classified pixels

TRANSFER LEARNING (RESEARCH)

Metric	Meaning
IoU (Jaccard Index)	Intersection over Union (per class)
mIoU (mean IoU)	Average IoU over all classes
Dice Coefficient	Like F1-score, used for imbalanced segmentation (especially medical)

Common Datasets

- **PASCAL VOC**: 20 categories (person, car, dog, etc.)
- **Cityscapes**: Urban street scenes (great for ADAS projects)
- **ADE20K**: 150 categories (objects + stuff)
- **COCO-Stuff**: Adds “stuff” categories to COCO
- **Medical**: BraTS, ISIC, etc.

Feature	Semantic Segmentation
Granularity	Pixel-level
Output	Mask (same size as input image)
Good for	Spatial understanding
Limitation	Can't separate multiple instances of the same class (e.g., 2 people overlap)

4. Instance Segmentation

What Is It?

Instance Segmentation = Object Detection + Semantic Segmentation

It not only **classifies every pixel** but also:

- **Separates each object instance**, even if they belong to the **same class**.

Semantic vs Instance Segmentation

TRANSFER LEARNING (RESEARCH)

Task	Detects "what"	Detects "where"	Separates objects
Semantic Segmentation	✅	✅ (pixel-level)	❌ (groups same-class pixels together)
Instance Segmentation	✅	✅ (pixel-level)	✅ (individual objects)

Example:

- Two dogs in an image
  - Semantic segmentation: All dog pixels labeled "dog"
  - Instance segmentation: **Dog #1** mask, **Dog #2** mask

Popular Models

Model	Key Idea
Mask R-CNN	Extends Faster R-CNN by adding a third branch to predict segmentation masks
YOLACT / YOLACT++	Real-time, combines YOLO speed with masks
SOLO / SOLOv2	Segments objects directly without bounding boxes
Detectron2	Facebook's full framework for detection/segmentation tasks

---

How Mask R-CNN Works (Simplified)

1. **Backbone (e.g., ResNet)** extracts features from input image
2. **RPN (Region Proposal Network)** finds likely object regions
3. **RoIAlign** crops these regions and sends to:
   - Classifier
   - Bounding box regressor
   - **Mask branch** (predicts a binary mask for each instance)

Evaluation Metrics

Metric	Meaning
IoU (for masks)	Intersection over Union of predicted mask vs true mask
mAP@[IoU]	Average Precision for masks over various IoU thresholds

TRANSFER LEARNING (RESEARCH)

Metric	Meaning
Dice Score / F1 for masks	Especially in medical or binary segmentation
Per-instance Precision/Recall	Measures individual object mask accuracy

---

Datasets

- **COCO** (Common Objects in Context) — popular for object detection + instance masks
- **LVIS** — large vocabulary + instance segmentation
- **Cityscapes** — includes instance labels
- **Kitti-MOTS** — autonomous driving, multi-object tracking + segmentation

---

Applications

- Medical: segment tumors, cells, or organs per patient
- Autonomous driving: identify individual pedestrians, cars
- Retail: product instance detection
- Robotics: pick up specific objects
- Image editing: object cutout, manipulation

Summary

Feature	Instance Segmentation
Granularity	Pixel-level
Separates Instances?	Yes
Use Case	Multi-object, spatially detailed analysis
Models	Mask R-CNN, YOLACT, SOLOv2

5. Pose Estimation

What is Pose Estimation?

Pose estimation refers to the task of **locating key points (joints)** of a person (or object) in an image or video.
In humans, these keypoints could be:

TRANSFER LEARNING (RESEARCH)

- Eyes
- Shoulders
- Knees
- Ankles
- Hands, wrists, elbows

It’s not just about finding a “person” — it’s about **understanding how they're positioned**.

Real-world Example:

Take an image of a person walking:

- Output: Coordinates for all 17 joints (like COCO format)
- You can reconstruct their **skeleton** and even determine their **action** (walking, jumping, etc.)

MOST POPULAR MODELS FOR POSE ESTIMATION

Model	Description
OpenPose	First multi-person real-time pose estimator
HRNet	Maintains high-resolution features throughout the network
MediaPipe Pose	Google’s ultra-fast real-time pose estimation (good for mobile)
DeepPose	First to use deep learning for pose (by Google)
PoseNet	TensorFlow.js-compatible, lightweight pose estimation

How It Works

1. **Input Image**
2. CNN extracts feature maps
3. For each keypoint (e.g., elbow), predict a **heatmap** where it's most likely to be
4. Output final coordinates of all keypoints (max value in heatmap)

Evaluation Metrics

Metric	Description
PCK (Percentage of Correct Keypoints)	A keypoint is correct if it's within a radius from the ground truth point
OKS (Object Keypoint Similarity)	Like IoU, but for keypoints — accounts for scale and location

TRANSFER LEARNING (RESEARCH)

Metric	Description
mAP for keypoints	Average precision across all joints and thresholds

6. Face Recognition / Verification

What is it?

This task involves using a person's face to:

- ✅ **Recognize** who they are (Face **Identification**)
- ✅ **Verify** if two faces belong to the same person (Face **Verification**)

It's different from face **detection**, which only finds the location of faces in an image.

Examples:

- Face ID on iPhones → **Face verification**
- Facebook auto-tagging → **Face recognition**
- Security systems → both

Two Modes:

Mode	Description
1-to-1 (Verification)	Is this person A? → Yes/No (e.g., unlocking phone)
1-to-N (Recognition)	Who is this person among many? (e.g., attendance)

Popular Models

Model	Notes
FaceNet	Learns embeddings; distance-based verification (Google)
Dlib	Lightweight C++ library with Python bindings
ArcFace	Uses angular margin loss; highly accurate
DeepFace	High-level Python API using multiple models underneath
VGGFace2	Model trained on large-scale celebrity dataset
InsightFace	State-of-the-art, optimized for production

How It Works:

TRANSFER LEARNING (RESEARCH)

1. Face is **detected**
2. Extract **embedding vector** (e.g., 128D or 512D)
3. For comparison:
    - If distance < threshold → **same person**
    - Else → different

Common distance metrics:

- **Cosine similarity**
- **Euclidean distance**

Evaluation Metrics

Metric	Description
Accuracy	Correct matches / total comparisons
ROC Curve / AUC	Visualizes true positive vs false positive rates
FAR (False Acceptance Rate)	% of wrong matches accepted
FRR (False Rejection Rate)	% of correct matches rejected
EER (Equal Error Rate)	Point where FAR = FRR (used in benchmarks)

Datasets

- **LFW** (Labelled Faces in the Wild)
- **MS-Celeb-1M**
- **VGGFace2**
- **CASIA-WebFace**
- **FaceScrub**

7. OCR – Optical Character Recognition

What is OCR?

OCR stands for **Optical Character Recognition** — it's the task of extracting **text from images**.

This includes:

- Scanned documents
- Handwritten notes
- Street signs in photos

TRANSFER LEARNING (RESEARCH)

- License plates
- Screenshots of code or articles

OCR Pipeline (Simplified)

1. **Text Detection**
    - Locate where text is in the image (bounding boxes)
2. **Text Recognition**
    - Convert the image inside each box into readable characters

---

Popular Models & Frameworks

Tool/Model	Purpose
Tesseract	Most common open-source OCR engine (by Google)
EAST	Efficient and Accurate Scene Text Detector
CRAFT	Character-Region Awareness for text detection
CRNN	Combines CNN + RNN + CTC for robust recognition
TrOCR	Transformer-based OCR (Microsoft)
EasyOCR	High-level Python wrapper over deep OCR stack
PaddleOCR	Very accurate; supports 80+ languages

Evaluation Metrics

Metric	Description
CER (Character Error Rate)	% of characters incorrectly recognized
WER (Word Error Rate)	% of words with mistakes
BLEU score (if comparing to reference text)	How close the output matches ground truth
Precision/Recall (for detection stage)	How accurately it finds text locations

8. Image Captioning

What is Image Captioning?

Image captioning is the task of generating a **natural language description** of an image. It blends **computer vision** (to understand the image) and **natural language processing** (to generate sentences).

TRANSFER LEARNING (RESEARCH)

Example:

Input:
A photo of a man riding a horse on a beach.

Output (Caption):

“A man is riding a horse along the shoreline.”

How it Works (Conceptually)

1. **CNN encoder** extracts features from the image (e.g., ResNet, EfficientNet)
2. **RNN / Transformer decoder** generates words one by one based on those features
3. The process is trained on pairs of (image, caption)

Popular Models

Model	Description
Show and Tell	CNN + LSTM model by Google (first deep learning-based image captioning model)
Show, Attend and Tell	Adds **attention** to focus on parts of image while generating each word
NIC (Neural Image Captioner)	Early encoder-decoder framework using InceptionNet + LSTM
BLIP / BLIP-2	Vision-language model with transformer decoder
ViT + GPT combos	Transformers on both vision and text sides (zero-shot capable)
CLIP + Decoder	Uses CLIP image embeddings and language decoders like GPT-2

Evaluation Metrics

Metric	What it Measures
BLEU	N-gram overlap with reference captions (precision-like)
ROUGE	Measures recall of overlapping units (more NLP focused)
CIDEr	Measures consensus with multiple reference captions (best for captions)
METEOR	Accounts for synonyms, stemming — better linguistic match
SPICE	Evaluates scene-graph level meaning

TRANSFER LEARNING (RESEARCH)

Example BLEU Calculation:

GT Caption: "A dog is running"
Predicted: "A dog is playing"
→ 2 out of 3 words match → BLEU score ≈ 0.66

Datasets

- **MS-COCO** (standard dataset with 5 captions per image)
- **Flickr8k / Flickr30k**
- **Visual Genome**
- **Conceptual Captions** (web-scaled dataset)

Real-World Use Cases

- Image accessibility for the visually impaired (screen readers)
- Automatic alt-text generation for the web
- AI-assisted photo tagging
- Visual question answering (VQA) building blocks
- News/media caption automation

🔎 **9. Image Super-Resolution**

What is Image Super-Resolution?

Image Super-Resolution (SR) is the task of **enhancing the resolution of a low-quality image** — essentially turning a blurry or pixelated image into a sharper, clearer version.

You input:

- A **low-resolution (LR)** image
  It outputs:
- A **high-resolution (HR)** version of the same image with improved detail

---

Real-world Example:

Input:
32×32 pixel face

Output:
128×128 or even 512×512 face with enhanced details

TRANSFER LEARNING (RESEARCH)

Types of SR

Type	Description
Single Image SR (SISR)	Enhance one image at a time
Video SR	Enhance resolution of frames in a video
Multi-image SR	Fuse multiple low-res views into one better image

---

Popular Models

Model	Highlights
SRCNN	First deep learning model for SR (simple and elegant)
SRGAN	Introduced perceptual + adversarial loss for realism
ESRGAN	Enhanced SRGAN with better detail recovery
Real-ESRGAN	Trained on real-world image degradation — great for photos
SwinIR	Transformer-based, state-of-the-art SR quality
EDSR	Very deep CNN without batch normalization for performance

---

How It Works (Typical Flow):

1. **Input:** LR image (e.g., 64x64)
2. **Upsampling Layer:** Bicubic or learned
3. **Deep CNN / GAN layers:** Restore lost features (e.g., textures)
4. **Output:** HR image (e.g., 256x256)

Evaluation Metrics

Metric	What it measures
PSNR (Peak Signal-to-Noise Ratio)	Higher = better pixel-level accuracy
SSIM (Structural Similarity Index)	Measures perceptual similarity (0–1)
LPIPS (Learned Perceptual Image Patch Similarity)	Learned metric aligned with human judgment (lower = better)
FID (Fréchet Inception Distance)**	If using a GAN-based model, FID helps measure realism

---

Use Cases

- **Upscaling old photos** (AI photo enhancers)

TRANSFER LEARNING (RESEARCH)

- **Satellite imagery** (sharpen terrain, roads, etc.)
- **Video streaming** (improve video quality at low bandwidth)
- **Forensics** (enhance blurry CCTV frames)
- **Medical imaging** (CT, MRI clarity improvement)

10. Image Generation

What is Image Generation?

Image generation is the task of **creating completely new images** using AI models — either:

- From **random noise** (like GANs)
- From **text prompts** (like DALL·E or Stable Diffusion)
- From **other images** (like style transfer or image-to-image translation)

Examples

- Generate fake faces: "a photo of a person who doesn't exist"
- Text-to-image: "a futuristic car driving on Mars"
- Image editing: remove background or colorize black-and-white photos

---

Major Techniques in Image Generation

Method	Description
GANs (Generative Adversarial Networks)	Learn to create realistic images by pitting two networks (Generator vs Discriminator)
Diffusion Models	Start with random noise → gradually "denoise" into a high-quality image
VQ-VAE (Vector Quantized VAE)	Discrete latent representation learning
Autoregressive Models	Predict next pixel/patch (e.g., PixelCNN)

TRANSFER LEARNING (RESEARCH)

Method	Description
Text-to-Image	Uses both vision and language models (CLIP + UNet)

Popular Models

Model	Description
StyleGAN2 / StyleGAN3	State-of-the-art GANs for photorealistic face generation
BigGAN	High-res class-conditional generation
CycleGAN	Translates images across domains (horse ↔ zebra)
Stable Diffusion	Text-to-image generation with stunning detail
DALL·E 2	OpenAI's model that generates images from text prompts
Midjourney	Proprietary, highly stylized text-to-image generation
DreamBooth	Fine-tunes a model on *you* (personalized generation)

How GANs Work (Simplified)

1. **Generator (G)** tries to make fake images
2. **Discriminator (D)** tries to tell if they're fake or real
3. They train together until the fake images are **indistinguishable** from real

---

Evaluation Metrics

Metric	Measures
FID (Fréchet Inception Distance)	Closeness of generated to real data (lower is better)
IS (Inception Score)	How diverse and high-quality the images are
LPIPS	Measures perceptual similarity (used for image-to-image tasks)
Human Evaluation	Sometimes the best option for creativity tasks

---

Applications

TRANSFER LEARNING (RESEARCH)

AI art and design (Midjourney, DALL·E)

Avatar & face generation

Photo enhancement & editing

- **Fashion try-ons or virtual product mockups**
- **Data augmentation** for training CV models
- **AI-generated video frames** (future of animation)

Summary

Feature	**Image Generation**
Input	Noise / Text / Image
Output	Fully generated image
Top Models	StyleGAN, Stable Diffusion, DALL·E
Metrics	FID, IS, LPIPS
Creativity	💯 Unmatched — truly generative AI

11. 3D Reconstruction / Depth Estimation

What is it?

This task focuses on understanding the **3D structure** of a scene or object from **2D images**.

- **Depth Estimation**: Predicts how far each pixel is from the camera
- **3D Reconstruction**: Builds a full 3D model (point cloud, mesh, or volume) from one or more 2D images

---

Example:

Input:
A photo of a road

Output:
A grayscale **depth map** where brighter pixels = closer

TRANSFER LEARNING (RESEARCH)

Or:
A **3D mesh** of the object or environment

Types of Depth Estimation

Type	Description
Monocular	Predict depth from a **single image**
Stereo	Use two camera views (like human vision)
Multi-view	Use multiple images from different angles
RGB-D	Combine color + depth sensor (like Kinect)

Popular Models

Model	Notes
MonoDepth / MonoDepth2	Monocular depth estimation from a single RGB image
MiDaS	Multi-scale, trained on many datasets, generalizes well
DPT (Dense Prediction Transformer)	Transformer-based for high-quality depth maps
NeRF (Neural Radiance Fields)	Volumetric 3D rendering from multiple images
COLMAP	Traditional multi-view 3D reconstruction (SfM/SLAM)

Evaluation Metrics

Metric	Description
RMSE (Root Mean Square Error)	Distance between predicted & true depth (lower is better)
MAE (Mean Absolute Error)	Average difference in depth values
Abs Rel Error	Average error relative to true depth
Threshold Accuracy (δ)	% of pixels where prediction is within factor (e.g., $\delta < 1.25$)

Real-World Applications

- **Autonomous vehicles**: depth sensing for driving & collision avoidance
- **AR/VR**: creating immersive environments

TRANSFER LEARNING (RESEARCH)

- **Medical imaging**: reconstructing 3D scans from 2D slices
- **Architecture & mapping**: building 3D models of spaces
- **Robotics**: environment perception for grasping or navigation

12. Video Action Recognition

What is it?

Video action recognition is the task of **classifying the action** taking place in a **video clip or sequence of frames**.

You're not just identifying *what* is in the scene — you're understanding *what is happening over time*.

Example:

- A 3-second video of a person jumping → Model predicts: "jumping"
- A sports video → Predicts: "throwing a basketball", "kicking", "swimming"

Key Difference from Image Classification:

You're working with **space + time**, not just pixels in a static image.

That means temporal patterns matter — like **motion**, **velocity**, and **frame changes**.

Popular Models

Model	Description
C3D (3D ConvNet)	Applies 3D convolutions over space & time
I3D (Inflated 3D ConvNet)	Inflates 2D kernels into 3D, using pretrained image models
SlowFast Networks	One stream captures slow features (semantics), the other fast motion
TimeSformer	Pure transformer model for video action recognition
VideoMAE	Masked autoencoder for self-supervised video learning
MoViNet	Optimized for mobile & real-time performance

How It Works

TRANSFER LEARNING (RESEARCH)

1. Sample video frames (e.g., 16 frames)
2. Extract **spatiotemporal features** using CNNs or Transformers
3. Use temporal pooling or attention
4. Predict an action label (e.g., “running”)

Evaluation Metrics

Metric	Description
Top-1 Accuracy	% of videos where the top prediction is correct
Top-5 Accuracy	% where correct label is in top 5 predictions
mAP (mean Average Precision)	Useful in multi-label settings
Precision / Recall	For per-action performance breakdown
Confusion Matrix	Shows what actions get confused with others

Datasets

Dataset	Description
UCF-101	101 action classes (sports, human activities)
Kinetics-400/600/700	Large-scale YouTube clips labeled with actions
HMDB-51	51 actions from movies and public video clips
Something-Something	Actions that require temporal understanding (e.g., “putting object on table”)
AVA	Annotated Video Actions with spatial + temporal localization

TRANSFER LEARNING (RESEARCH)