

Basic Concept

AI, NI, Cognitive Function

- AI (artificial intelligence)
 - the intelligence exhibited by machines
- NI (natural intelligence)
 - the intelligence exhibited by humans and animals
- Cognitive Function
 - mental capabilities that enable individuals to process information, reason, learn, and solve problems

Elements of NN

- Neurons
 - the basic unit of a neural network
- Activation Function
 - a function that determines the output of a neuron

4 branches of ML

- Supervised Learning

Train a model on **labeled** or annotated examples.

- Unsupervised Learning:

Find patterns in a dataset without **label**

- Self-supervised Learning:

Does not use human-labeled training data but generating labels.

- Reinforcement Learning:

choose action to maximize some reward function based on the env

Division of dataset

- 100 ~ 10k data:
 - 70/30 (70% training, 30% testing)
 - 60/20/20 (60% training, 20% dev, 20% testing)
- 100K data: 90/5/5 (90% training, 5% dev, 5% testing)
- 1M labeled data: 98/1/1 (98% training, 1% dev, 1% testing)
- n millions labeled data, can use 99.5/0.4/0.1

Overfitting vs. Underfitting

- Bias refers to the error that is due to **overly simplistic assumptions** in the learning algorithm.
- **High bias** can lead to **underfitting**

bias = TrainingSet Error

- Variance refers to the error that is due to **excessive complexity** in the learning algorithm.
- **High variance** can lead to **overfitting**

variance = DevSet Error - TrainingSet Error

Vanishing / Exploding gradients

This occurs in training a deep neural network especially when dealing with **very deep layers**.

- Vanishing gradients can lead to a **slow convergence**

- Exploding gradients can result in very large and **unpredictable updates to the weights**

Weights Regularization

L_1 Regularization: $\frac{\lambda}{m} \|w\|_1$ where $\|w\|_1 = \sum(|w_i|)$

L_2 Regularization: $\frac{\lambda}{2m} \|w\|_2^2$ where $\|w\|_2^2 = \sum(w_i^2)$

The purpose of this term is to **constrain the model's complexity** by forcing the weights to take on smaller values, which prevent the model from fitting the noise in the training data.

Dropout Regularization

- Randomly set some neurons to zero in each iterations of **training**
- prevent overfitting by **discouraging reliance** on any single feature
- The dropout rate refers to the **fraction of output features that are randomly set to zero** during the training of a neural network

Data Augmentation

- images: flipping, rotating, randomly cropping
- text: cropping, back translation

Data Normalization

- form a **standard normal distribution** (mean = 0, variance = 1) using $x \leftarrow \frac{x-\mu}{\sigma}$

Algorithms for finding the minimum of the cost function

- Gradient Descent
- Gradient Descent with Momentum

$$V_{dw} \leftarrow \beta V_{dw} + (1 - \beta) dW$$

$$V_{db} \leftarrow \beta V_{db} + (1 - \beta) db$$

$$W \leftarrow W - \alpha V_{dw}$$

$$b \leftarrow b - \alpha V_{db}$$

- Root mean square propagation
- Stochastic gradient descent

Small updates for **large** oscillations, and **large** updates for **small** oscillations.

Learning rate decay

- Same rate for all iterations \rightarrow wander around the minimum
- LR decay \rightarrow allow taking smaller steps as it approaches the minimum
- Taking big steps at the beginning and small steps at the end

Optima

In most optima, we have **saddle points** rather than min or max points, because the training data are in high dimensions, you may **descending in one dimensions but ascending in another dimensions**.

Problems near saddle points:

- slope is small, causing the optimization algorithm to take small steps
- solution: **stochastic gradient descent**

Explain the difference between parameters and hyperparameters. Give examples of each.

When do we need to retune hyperparameters?

- When applying a Model to a different application
- When new data is introduced or the model performance degraded

Explain the “panda” versus the “caviar” approach in tuning hyperparameters. In which situation would you use panda? Use caviar?

- “Panda” approach: Train only one model, adjust hyperparameter each day
 - You have lots of data but not much computational resources
- “Caviar” approach: Train many models in **parallel** with different hyperparameters
 - You have lots of computational resources

Briefly explain the main idea in batch normalization. How is batch norm similar to normalizing inputs (C2M1L09)? How are they different?

How does batch norm improve the calculations? Under what circumstance would you use batch norm?

How:

- Speeds up learning (training)
- Slight regularization effect (prevent overfitting)

When:

- new data with different distribution is introduced

Briefly explain softmax. What is it used for?

- Softmax is a mathematical function that transforms a vector of raw scores (logits) into probabilities that sum to one.
- Softmax is typically used in **classification problems**, where it enables models to predict the likelihood of each class given an input.

How do we calculate softmax?

$$\text{softmax}(x) = \frac{e^{x_i}}{\sum(e^{x_i})}$$

Name some of the deep learning frameworks presented in this class. Which two are used the most today?

- Pytorch, developed by Facebook
- Tensorflow / Keras, developed by Google

C3M1

What is the meaning of perfect precision? What is the meaning of perfect recall?

- Perfect precision means **no false positives**
- Perfect recall means **no false negatives**

When you combine precision and recall, what is that metric called? Give the formula.

The metric that combines precision and recall is called the **F1 Score**.

$$\text{F1 Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

What is the meaning for “dev set is like setting the target”? What is the consequence of this statement?

- It means the dev set is served as the benchmark to evaluate the performance of model
- Biased dev set → biased model performance (overfitting in dev set, poor real-world performance)

How should we divide our labeled data into training / dev / testing if we have a) 1K labeled data b) 100K labeled data c) 1 million labeled data?

Amount labeled data	Train/Dev/Test %
100 to 10K	60/20/20
100K	90/5/5
1M	98/1/1

What is Bayes optimal error?

Bayes optimal error is the **lowest possible error rate** for any classifier on a given classification problem with known class distributions.

How do we compute avoidable bias? How do we compute the variance?

$$\text{Avoidable Bias} = \text{Bayes Error} - \text{Training Error}$$

$$\text{Variance} = \text{Dev Error} - \text{Training Error}$$

C3M2

What is the purpose of error analysis? Briefly describe how you would use it? Draw a table to illustrate.

Purpose: understand the underlying causes of the errors by examining the misclassified examples

- Manually examine 100 mistakes (takes 2 hours)
- Categorize errors into meaningful groups
- Count frequency of each error type
- Focus on categories with highest potential impact

Supposed we have 2 different sets of labeled data from different distributions. One large set is downloaded from the Internet, and a smaller set is specifically made for the app we want to build. How should we divide the data for training, dev and testing?

Total Data:

- 200,000 internet images
- 10,000 app images

Split:

- Training: 200,000 internet + 5,000 app images
- Dev: 2,500 app images
- Test: 2,500 app images

Key Principle:

- Dev/test must reflect future real-world data you want to perform well on
- Don't randomly mix distributions, even if tempting

C4M1

Know how to calculate convolutions, compute output sizes and the number of parameters in each layer.

image ($W \times H$), kernel ($K_w \times K_h$), padding (P_w and P_h), stride (S_w and S_h)

$$W_{\text{out}} = \left\lfloor \frac{W - K_w + 2P_w}{S_w} + 1 \right\rfloor \quad H_{\text{out}} = \left\lfloor \frac{H - K_h + 2P_h}{S_h} + 1 \right\rfloor$$

Parameters calculation:

- Conv layer: $K_w \times K_h \times C_{\text{in}} \times C_{\text{out}} + C_{\text{out}}$ (bias for each filter)
- Fully connected layer: $N_{\text{in}} \times N_{\text{out}} + N_{\text{out}}$ (bias for each neuron)

Why do CONV layers has so few parameters compared to densely connected layers?

- Shared parameters: the same filter is applied to every position in the input
- Sparse connections: each output value depends only on a small number of input values
- Reduce the risk of overfitting: fewer parameters \rightarrow less likely to overfit

Is convolution linear or nonlinear? Is maxpool linear or nonlinear?

- convolution: linear
- max pool: nonlinear

Why do we place a maxpool layer between conv layers?

\Rightarrow add max pool layer between conv layers to introduce **non-linearity**

In the architecture for VGG, ResNet, etc., we always place a maxpool layer between conv layers. Why?

- Make image smaller, so the computation is faster and more robust
- Capture the most prominent features instead of noise
- Introduce non-linearity

C4M2

What are the classical networks presented in the lecture? Briefly explain the main idea in each network.

- LeNet: first successful CNN, used for handwritten digit recognition
- AlexNet: bigger and deeper than LeNet (deep nn is effective).
- VGG-16:
 - 3×3 conv block with 2×2 max pool, very deep network (16 layers with weights)
 - half image size but double channel size in each layer
- ResNet:
 - deep NN is hard to train \Rightarrow vanishing gradient problem
 - **skip connection**: prevent the lost of information in the deep layers

Briefly describe how a residual block works in ResNet. You can use a diagram or equations.

$$y_{k+1} = f(x_k) + x_k$$

What is a 1×1 convolution? When would you use it?

1×1 conv consider every pixel in the image, but across all the channels

- use to reduce the number of channels (e.g. VGG-16 have too many channels in deep layer, we can use 1×1 conv to reduce the channel size)
- making a full connection to another layer

How should we reduce the image size? How should we reduce the channel size?

- Reduce image size: max pool layer
- Reduce channel size: 1×1 conv layer

Describe the main ideas in transfer learning. How should you apply deep learning when you have (a) only a little of your own data for your app, (b) moderate amount of your own data, (c) and lots and lots of your own data?

What are some of the common augmentation methods?

C4M3

What is the difference between image classification, classification with localization, and object detection?

- Image classification: classify the image into a single category
- Classification with localization: classify the image and locate the object
- Object detection: locate multiple objects in the image and classify them

In an image, suppose we want to detect pedestrian, car, bicycle and background.

1. What are the components in the training label?
2. What is the loss function for this object detection?

Components in training label:

- Bounding boxes for each object
 - b_x, b_y : center of the box
 - b_h, b_w : height and width of the box
- Class labels (pedestrian/car/bicycle/background)

Loss Function:

Assuming output vector is $[p_c, b_x, b_y, b_w, b_h, c_1, c_2, \dots, c_n]$, then the loss function is:

$$\mathcal{L} = \begin{cases} \sum (y_i - \hat{y}_i)^2 & \text{if } y_1 = 1 \\ (y_1 - \hat{y}_1)^2 & \text{if } y_1 = 0 \end{cases}$$

What are landmark points? How do we use it in deep learning?

landmark points: important points in the image

application in DL:

- face emotion recognition
- pose position detection

What is a sliding window used for? How does it work? What is its drawback?

- It can be used to detect objects in the image
- How it works:
 - select window size
 - place window on the upper left corner
 - feed that part into ConvNet (output labels)
 - increase the window size and repeat the process
- Drawbacks
 - computation speed is slow, since a lot of windows need to be computed
 - resulting boundingbox is not good

What does YOLO stand for in deep learning? What does it do? What are its main idea?

YOLO stands for **You Only Look Once**.

What does IoU stand for? What is it use for? Describe its main idea.

- IoU stand for **Intersection over Union**.
- measure the overlap between 2 bounding box \Rightarrow evaluate the perf of object detection algorithms.
- main idea:
 - if IoU close to 1, the bounding box is good
 - if IoU close to 0, the bounding box is bad

What is object segmentation? What is class segmentation?

- Object segmentation: identifies and outlines individual instances of objects in an image
- Class segmentation labels each instance according to its category

C4M4

What is face recognition? Briefly describe the issues involved.

- Database has k people's face images, given a new image, determine if it is one of the k people
- Problem:
 - convnet does not work well with just few training example
 - need to retrain the network if new data added

In one-shot learning, how do we determine if two facial images match?

We check the difference between the two images. $d(\text{img}_1, \text{img}_2)$

- if $d \leq t$ then the two images match
- if $d > t$ then the two images do not match

Briefly describe Siamese network in facial recognition.

Convert the two images into a feature vector, then calculate the distance between the two vectors.

$$v_1 = f(x_1), v_2 = f(x_2), d = \|v_1 - v_2\|^2$$

What is the meaning of triple loss? What is the meaning of A, P and N?

Main idea: minimize the distance between the anchor image and the positive image, while maximizing the distance between the anchor image and the negative image.

$$\frac{\|f(A) - f(P)\|^2}{d(A, P)} \leq \frac{\|f(A) - f(N)\|^2}{d(A, N)} + \alpha(\text{margin})$$

A: anchor image P: positive image N: negative image

$$\mathcal{L}(A, P, N) = \max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0)$$

Briefly describe neural style transfer.

- combine the content of one image with the style of another image

In Neural Style Transfer, we use S , C and G to denote the images. What do S , C and G stand for? Briefly explain their roles in Neural Style Transfer. Please give an example.

- S : style image
- C : content image
- G : generated image

In neural style transfer (NST), what does the content involve? Which layer(s) the neural network is the content located?

- Content = higher level macrostructure (overall shape) of the C image
- the information is located in the deep layer of the neural network

In NST, what does the style involve? Where in the neural network is style located?

- Style = textures, colors, visual patterns of the S image
- Style is determined by how correlated the activations are across different channels in a layer.

C5M1

What is named entity recognition? What are the input and output?

- NER extract entities from the text, such as person names, locations, organizations, dates, etc.
- Input: a sentence
- Output: a list binary value, 1 if the word is an entity, 0 otherwise

Briefly describe RNN. What does it stand for? What are the input and output? You can draw a diagram.

What does a language model do? Please give an example.

A language model (LM) is designed to predict the likelihood of a sequence of words, essentially predicting the next word in a sentence based on the context provided by the preceding words.

How does a language model predict a missing word?

It use conditional probability to predict the next word given the previous words.

What is the drawback or weakness of (plain) RNN?

Vanishing gradient problem: the gradient becomes very small as it backpropagates through time, which makes it hard to learn long-range dependencies.

What do you use to overcome this weakness?

use GRU or LSTM

What is GRU? What does it stand for? What does it do?

- GRU stands for **Gated Recurrent Unit**
- It use two types of gate:
 - update gate: decide how much of the past information to keep
 - reset gate: decide how much of the past information to forget

What is LSTM? What does it stand for? What does it do?

- LSTM stands for **Long Short-Term Memory**
- use one more gate called **output gate** to decide how much of the current information to output

C5M2

What is one-hot vector? What are the elements in a one-hot vector?

- one-hot vector is a vector that has only one element as 1 and the rest are 0.
- used to represent word in NLP
- elements: binary values (0 or 1), all 0 except for one 1

What is the drawback of one-hot vector?

they have no relationship with each other, even if the words are related linguistically.

Briefly describe word embedding. What does the word embedding matrix show? What do columns show? What do the rows show? Please give example.

What is the advantage of using word embedding?

- Dense (not many zeros)
- Lower dimensionality (300 dim)
- Capture relationships between words

Compare one hot vector with word embedding vector.

Give an example of using transfer learning in word embedding. E.g. in named entity recognition.

- Pretrain a word embedding model on a large corpus (e.g. Wikipedia)
- Specific Task: NER
 - fine-tune the word embedding model on a smaller dataset

How do we use word embedding in analogy, e.g. Man is to Woman as King is to ???

In word embedding, analogies are computed using vector arithmetic.

Briefly explain the vector space projections from Gensim. (See the Gensim pdf file.) What do the plots show?

What is the cosine similarity function? What is it used for?

In sentiment classification, what is the input and output?

C5M3

In machine translation (MT) or neural machine translation (NMT), what is the input? What is the output?

- Input: a sentence in one language
- Output: a sentence in another language

How is MT and language model (LM) related? You can draw a diagram.

- Language Model: always start with a vector with all zeros $a^{<0>}$
- MT: start with a vector $a^{<0>}$ that is the output of the encoder

Under what basis in MT do we choose a sentence in translation, say from French to English?

In MT, compare greedy search with beam search.

- Greedy search: choose the word with the highest probability at each time step
- Beam search: keep track of the top k candidates at each time step, then choose the best one at the end (in probability)

Beam search can produce better results than greedy search, and also quicker than exhaustive search.

Describe the main ideas (up to 3) about the attention model. What is its intuitive meaning? What does the attention model consists of? What is the relation between context and attention weights? Drawing a diagram may help.