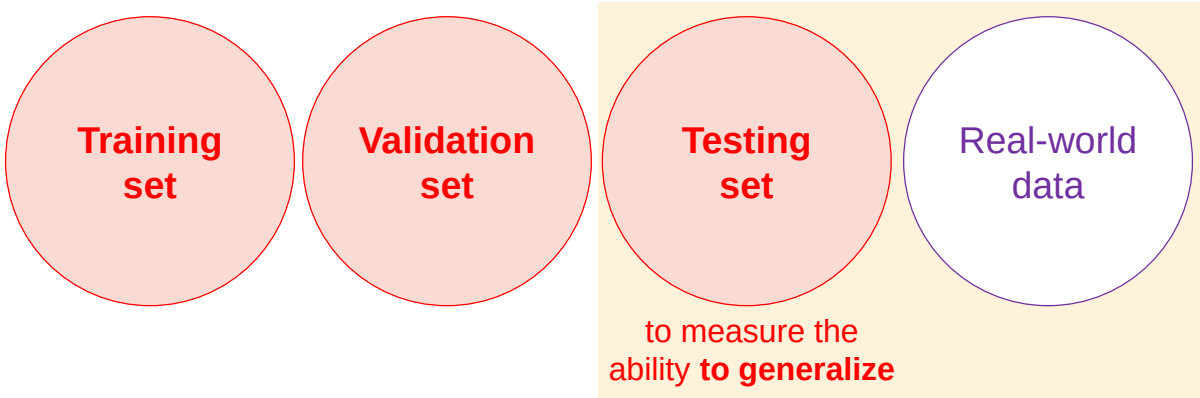
EE-559
Deep Learning

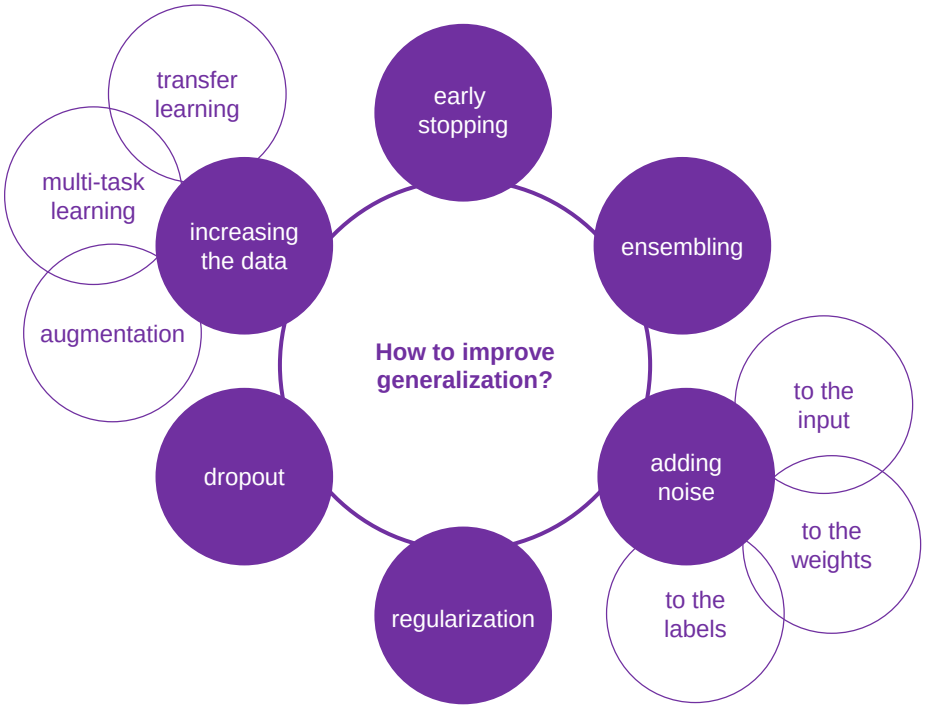
What's on today?

- Generalization: how to improve model performance
- Human-annotated data: on labeling for subjective tasks
- Self-supervised learning: how to train without human labels
- Contamination: on testing for memorized (internet-scale) data
- Reduced models: how to heavily decrease the parameter count
- Exercises: initialization, quantization, and distillation of a network

Generalization

Model performance on unseen data





Regularization

$$\Theta^* = \arg\min_{\Theta} L(\Theta)$$

$$= \arg\min_{\Theta} \left[\sum_{i=1}^{N} l_i(y_i, x_i) \right]$$

$$\Theta^* = \arg\min_{\Theta} \left[\sum_{i=1}^{N} l_i(y_i, x_i) + \lambda\, r(\Theta) \right]$$

positive scalar

regularization term

the more desirable Θ, the smaller $r(\Theta)$

λ controls the relative contributions of the original loss and $r(\Theta)$

Concepts:
Adding a regularization term to the loss function, prior on the weights

Example: Tikhonov regularization

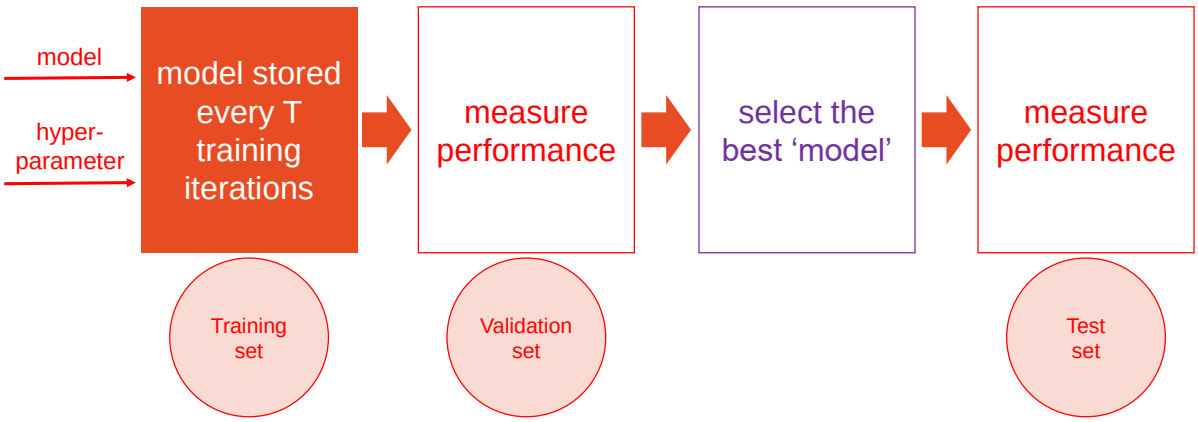
$$\Theta^* = \arg\min_{\Theta} \left[\sum_{i=1}^{N} l_i(y_i, x_i) + \lambda\, r(\Theta) \right]$$

$$= \arg\min_{\Theta} \left[\sum_{i=1}^{N} l_i(y_i, x_i) + \lambda \sum_{j=1}^{P'} \Theta_j^2 \right]$$

penalizes the sum of the squares of the parameter values

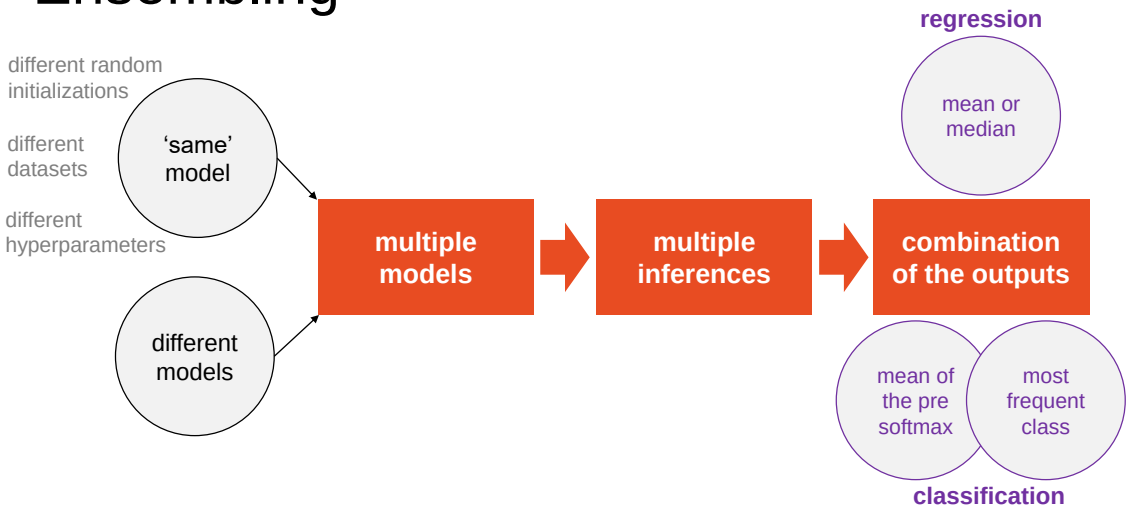
Concept:
Weight decay (*if not applied to biases*)

Early stopping



Concepts: Single-model and single hyperparameter heuristics, reduces model complexity, reduces overfitting, only coarse approximation of the function

Ensembling



Concepts: Building multiple models and averaging the predictions, bootstrap aggregating (bagging) with different datasets, costs: (i) storing models, (ii) training multiple models, (iii) multiple inferences

Dropout

Training

clamp a subsets of hidden units to zero

at each SGD iteration

Testing

network has more hidden units than it had at training

weight scaling inference rule

$(1 - \lambda)$ factor

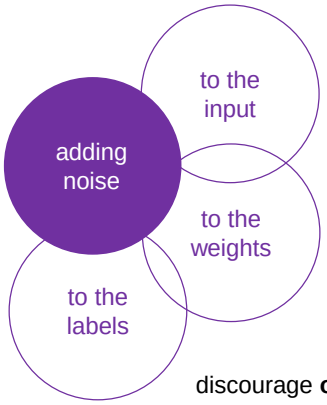
Monte Carlo dropout

run the network multiple times with random subsets of hidden units clamped to 0

combine the results (as in ensembling)

Concepts: Encouraging smaller magnitude weights, applying multiplicative Bernoulli noise to the activations, reduction of dependency on any hidden unit, removing large changes between data points

Adding noise



search for small perturbations that cause **large changes** to the output

adversarial training

convergence to local minima in **wide, flat regions**

discourage **over-confident behavior** (caused by the maximum likelihood criterion)

randomly change a proportion λ of the labels at each training iteration

Concepts: Encouraging sensible predictions for *small perturbations* of the inputs or the weights, adversarial training, label smoothing, maximum likelihood criterion pushes the final network activations (before the softmax) to (i) very large values for the correct values and (ii) very small values for the wrong classes

Adversarial noise

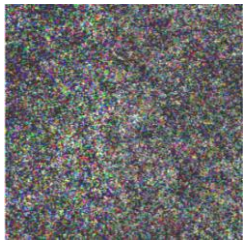
original (clean) input x $y = f(x; \Theta)$

perturbed (adversarial) input $\dot{x}$ $y \neq f(\dot{x}; \Theta)$

$\dot{x} = x + \delta$

adversarial perturbation

x



δ

Concept:
Intentionally perturbed input that misleads a trained model, how to fool a trained model

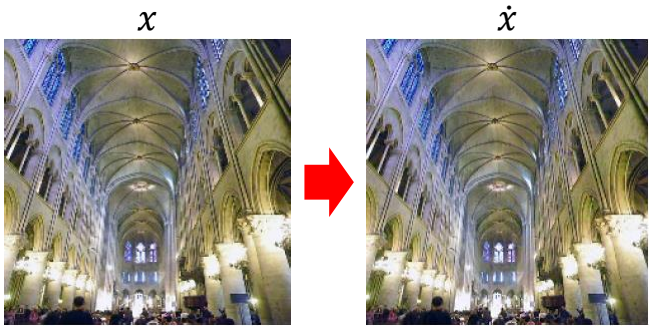
Fast Gradient Sign Method (FGSM)

$$\dot{x} = x + \delta$$

$$= x + \epsilon \, \text{sign}\left(\frac{\partial L(\Theta, x, y)}{\partial x}\right)$$

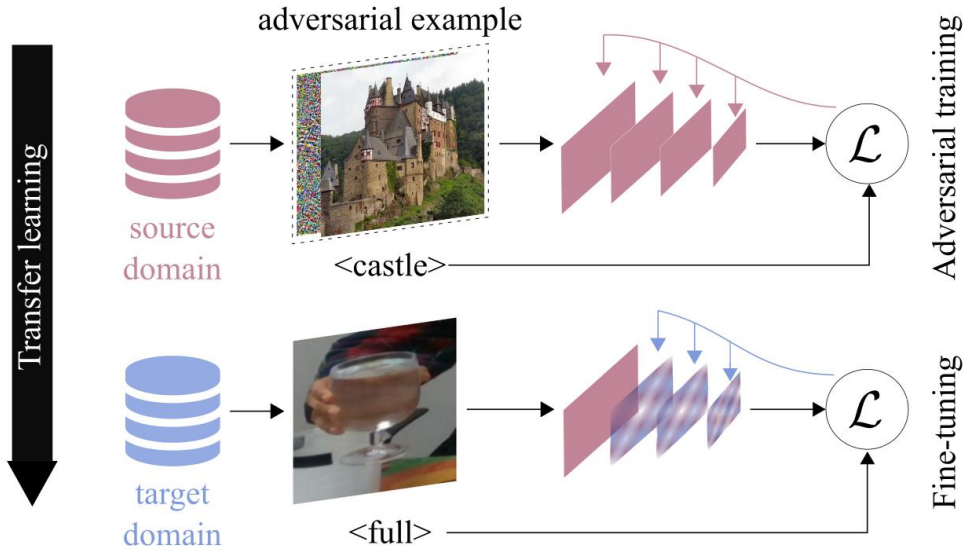
ϵ: controls the magnitude of the perturbation

We will play with FGSM during the lab today!

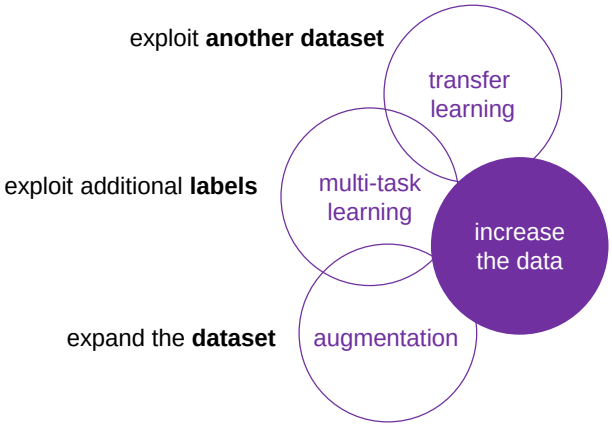


Concepts: Perturbations constrained by ε to maintain visual quality, gradient of the loss with respect to the image, Basic Iterative Method (BIM), clipping, targeted and untargeted approaches

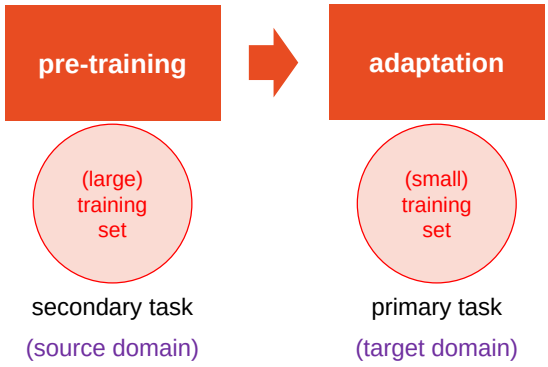
Adversarial training + transfer learning



Increasing the amount of data



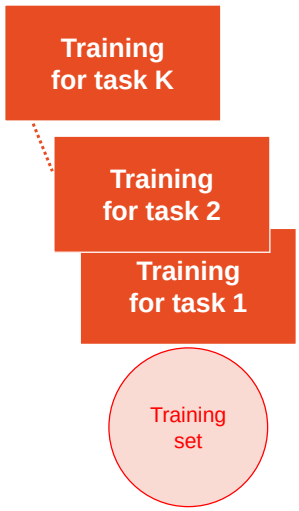
Transfer learning



Concepts: Initializing sensibly (most of) the parameters of the network with the secondary task, internal representation, replacing the last layer, fine-tuning the whole model

Multi-task learning

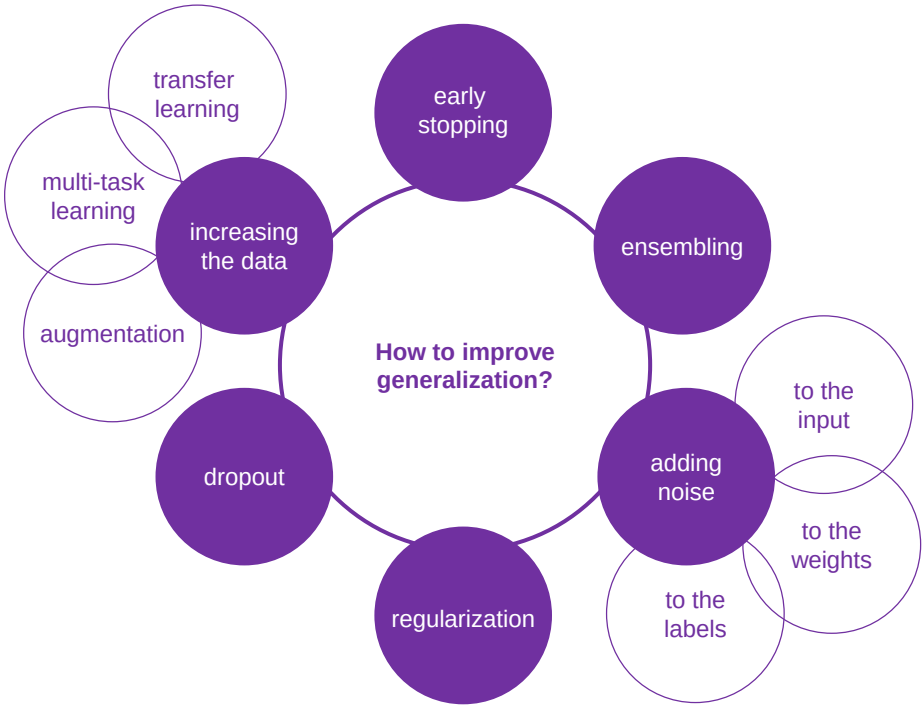
Learning tasks in parallel with a shared representation



Concepts: Train for several tasks concurrently, better representation of the data, improved learning efficiency, improved prediction accuracy, reduced overfitting

Augmentation

Concept: Transform each input sample without changing the label (e.g. manipulate colors, flip, blur, crop, add noise)



Human-annotated data

$$\{x_i, y_i\}_{i=1}^{N}$$

$$\{x_i, y_i\}_{i=1}^{N}$$

Objective vs subjective tasks

Example: instructions

Assume you have taken this photo, and you are about to upload it on your favorite social network or content sharing site.

Tell whether the image is private or public in nature.

Assume that you know the people in the photo.

PrivacyAlert *(simplified instructions)* https://zenodo.org/records/5841576

Example: categories

Clearly private: images that should not be uploaded online at all.

Private: images that should be kept confidential for me and selected trusted people only.

Public: images that anyone in my social network would be OK to see.

Clearly public: images that anyone online would be OK to see.

PrivacyAlert https://zenodo.org/records/5841576

slido

Choose a privacy category for the image.

ⓘ Start presenting to display the poll results on this slide.

slido

Choose a privacy category for the image.

ⓘ Start presenting to display the poll results on this slide.

Socio-cultural factors of annotators

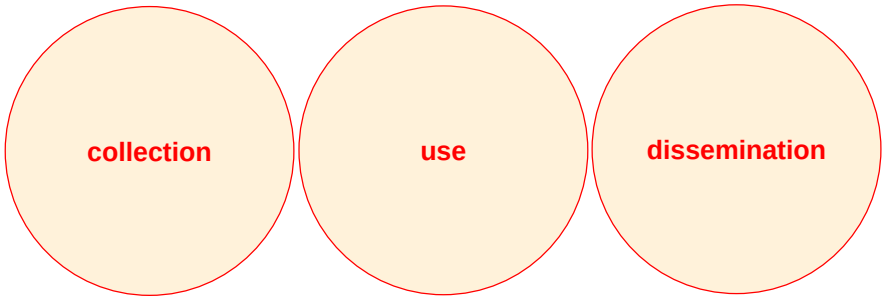
divergent annotations across different communities

skew in annotator demographics

situated domain expertise

Concept:
Annotator's identity and lived experience impact how annotation questions are (interpreted and) responded to

Crowd-sourced annotations



Concepts: Human computation, human intelligence and cognition,
creation of training datasets, human-in-the-loop system, reinforcement learning from human feedback

Datasets capturing subjective phenomena

(e.g. sentiment, hate speech)

Concepts:
How annotator social identity shape their understanding of the world, value-sensitive design

CrowdWorkSheets

task formulation	annotator selection	choice of platform	dataset analysis & evaluation	dataset release & maintenance

arXiv:2206.08931

Task formulation

- At a high level, what are the subjective aspects of your task?
- What assumptions do you make about annotators?
- How did you choose the specific wording of your task instructions? What steps, if any, were taken to verify the clarity of task instructions and wording for annotators?
- What, if any, risks did your task pose for annotators and were they informed of the risks prior to engagement with the task?
- What are the precise instructions that were provided to annotators?

arXiv:2206.08931

Annotator selection

- Are there certain perspectives that should be privileged?
  If so, how did you seek these perspectives out?
- Are there certain perspectives that would be harmful to include?
  If so, how did you screen these perspectives out?
- Were sociodemographic characteristics used to select annotators for your task?
- Do you have reason to believe that sociodemographic characteristics of annotators may have impacted how they annotated the data?
- Consider the intended context of use of the dataset and the individuals and communities that may be impacted by a model trained on this dataset.
  Are these communities represented in your annotator pool?

arXiv:2206.08931

Choice of platform

- What annotation platform did you utilize?
- At a high level, what considerations informed your decision to choose this platform?
- Did the chosen platform sufficiently meet the requirements you outlined for annotator pools? Are any aspects not covered?
- What, if any, communication channels did your chosen platform offer to facilitate communication with annotators? How did this channel of communication influence the annotation process and/or resulting annotations?
- How much were annotators compensated? Did you consider any particular pay standards, when determining their compensation?

arXiv:2206.08931

Dataset analysis & evaluation

- How do you define the annotation quality in your context, and how did you assess quality in your dataset?
- Have you conducted any analysis on disagreement patterns?
  If so, what analyses did you use and what were the major findings?
- Did you analyze potential sources of disagreement?
- How do the individual annotator responses relate to the final labels released in the dataset?

arXiv:2206.08931

Dataset release & maintenance

- Do you have reason to believe the annotations in this dataset may change over time?
  Do you plan to update your dataset?
- Are there any conditions or definitions that, if changed, could impact the utility of your dataset?
- Will you attempt to track, impose limitations on, or otherwise influence how your dataset is used? If so, how?
- Were annotators informed about how the data is externalized?
  If changes to the dataset are made, will they be informed?
- Is there a process by which annotators can later choose to withdraw their data from the dataset?

arXiv:2206.08931

Self-supervised learning

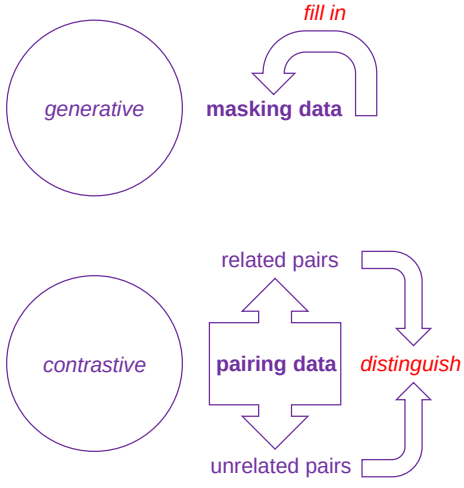
Self supervision

leverage **inherent structures** within the training data

error signals generated with the data themselves

no external labels provided by humans

$$\{x_i\}_{i=1}^{N}$$



Concept: Training using the data themselves, auto-associative self-supervised learning (autoencoders), generative self-supervised learning, contrastive self-supervised learning

Examples: masking data

image inpainting

missing word prediction

This course explores how to design reliable discriminative and neural networks, the ethics of data and model deployment, as well as modern multi-modal models.

Examples: pairing data (related pairs)

transformed versions of one another?

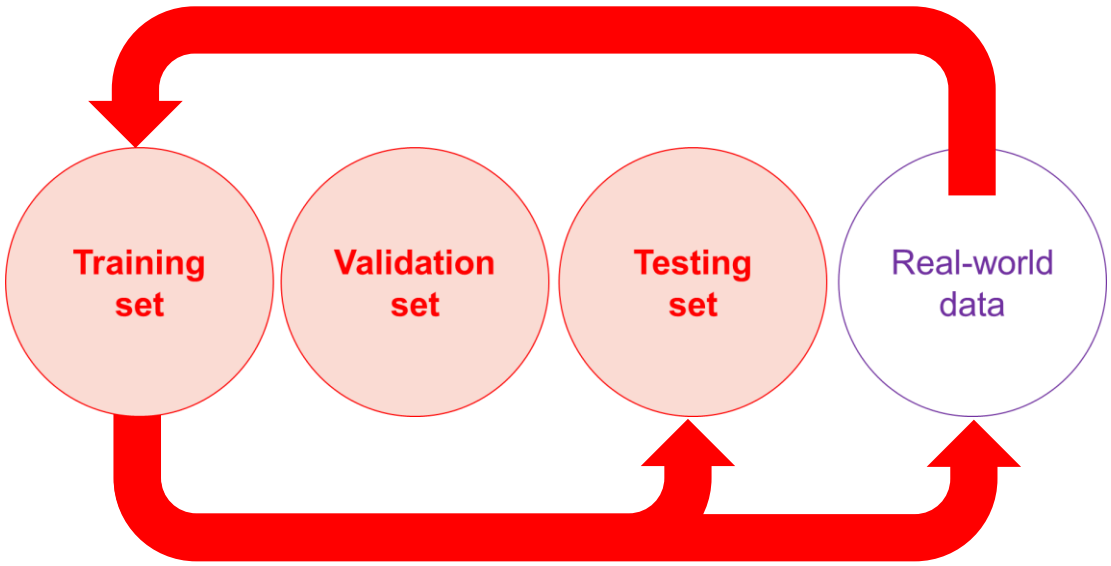
sentences followed one another?

This document was heavily influenced by the US Constitution and the ideals of the French Revolution. The basis for the present day Federal Constitution is the Constitution of 12 September 1848, which established the Swiss federal state.

The basis for the present day Federal Constitution is the Constitution of 12 September 1848, which established the Swiss federal state. This document was heavily influenced by the US Constitution and the ideals of the French Revolution.

Contamination

Memorization



Stored knowledge vs ‘reasoning’

Large Language Models trained on the text on the Internet

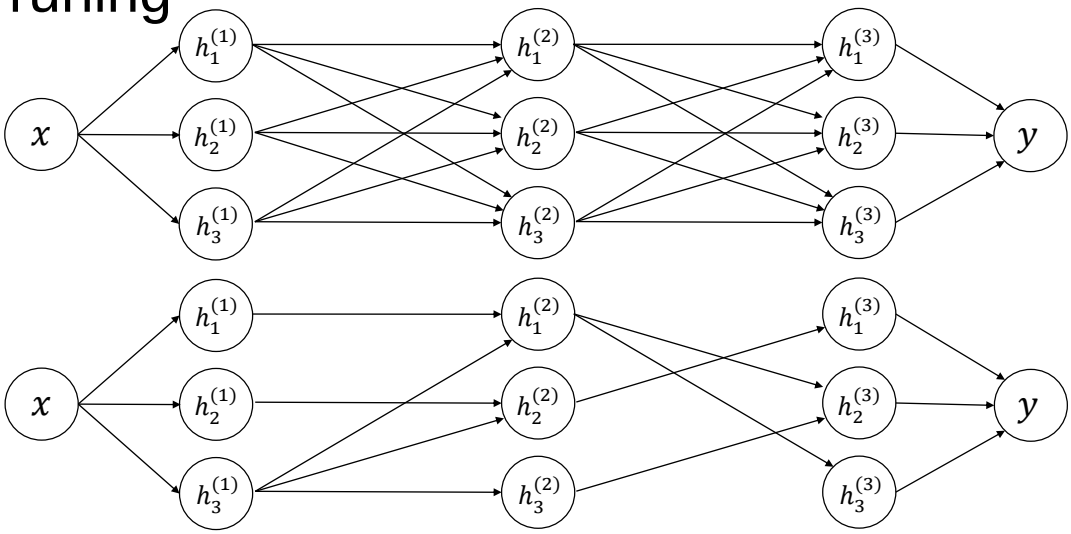
do they pass exams
(e.g. the bar, the medical license)
or did they “just” memorize the answers?

test for contamination!

https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield

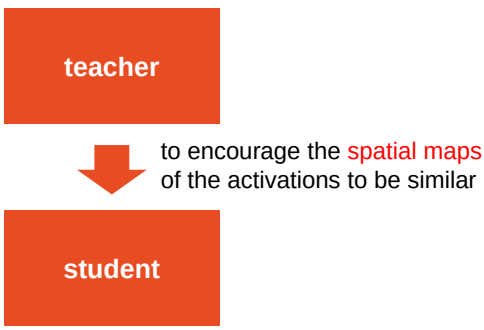
Reduced models

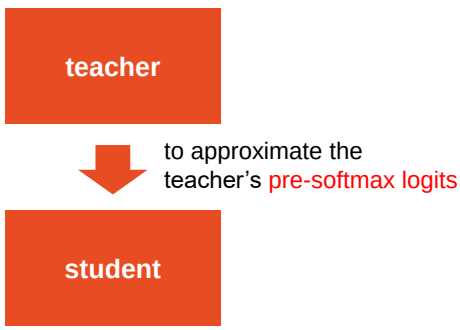
Pruning



Concepts:
Architecture search, removing weights w/o reducing performance, fine-tuning after pruning

Knowledge distillation





Concepts:
Training a smaller model (*student*) to have the same performance as a bigger model (*teacher*), attention transfer

Exercises

Today's exercises

Practice. You will become familiar with:

- data (i.e. text) pre-processing
- performance **metrics**, including the BLEU similarity metric for text
- generating **adversarial examples** with FGSM (re: adversarial training)

Marked. You will experiment with **initialization**, **quantization**, and the **distillation** of a network.

What did we learn today?

- Generalization
- Human-annotated data
- Self-supervised learning
- Contamination
- Reduced models
- Exercises

EE-559
Deep Learning

andrea.cavallaro@epfl.ch