SUSTech
Southern University of Science and Technology

神经计算与控制实验室
Neural Computing & Control Lab

Brain Intelligence and Artificial Intelligence

人脑智能与机器智能

Lecture 6 – Data for Deep learning and Pretraining

Quanying Liu （刘泉影）

SUSTech, BME department

Email: liuqy@sustech.edu.cn

Recall Lecture 5 – Intro to AI & hands on

- **A general introduction to AI**
  - 3 key components
  - The network **Architecture**
- **AI learning: Gradient Descent (GD)** to minimize a *loss function*
  - What is Gradient Descent?
  - Gradient Descent to train deep NNs → Error Backpropagation
- **Hands-on** (pytorch), thanks to three TAs
- **Error Back-propagation (BP)** in fully-connected NN
  - Backpropagation – forward pass
  - Backpropagation – backward pass
  - BP in the brain

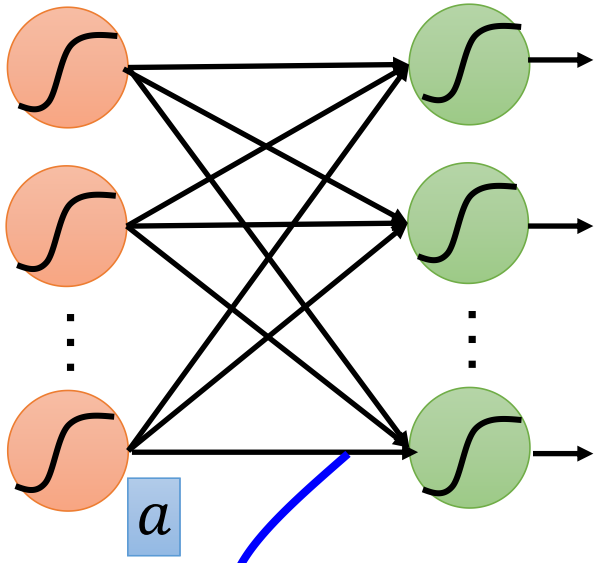
Recommendations of AI courses

Self-learning is all you need.

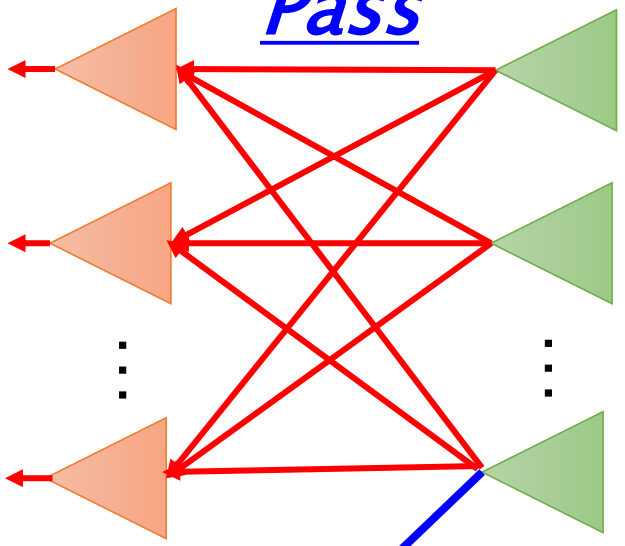
- Machine learning courses (Caltech CS156, Stanford CS229)
- B站 – 机器学习白板推导系列
- Deep Learning for Computer Vision (Stanford CS231N)
  - 2021年，https://www.bilibili.com/video/BV1TQ4y1B7yx/
- Natural Language Processing with Deep Learning (Stanford CS224N)
  - 2023年，https://www.bilibili.com/video/BV1d6421f7oW
- Deep Generative Models (Stanford CS236, Prof. Stefano Ermon)
  - 2023年，https://www.bilibili.com/video/BV1Wf421m7xP/
- B站 - 跟李沐学AI
- B站 - bryanyzhu

Backpropagation – Overview

Forward Pass



Backward Pass



$$\frac{\partial l}{\partial z_{i,j}} = \sigma'(z_{i,j})\left[\sum w_{i,j+1}\frac{\partial l}{\partial z_{i,j+1}}\right]$$

$$\frac{\partial l}{\partial z_{1,n}} = \frac{\partial l}{\partial y_{1,n}}\frac{\partial y_{1,n}}{\partial z_{1,n}}$$

- Initiate parameters all w^0
- Go forward pass to get all a^0
- $\frac{\partial z}{\partial w_{i,j}^0} = a_{i,j}^0$

$$\frac{\partial z}{\partial w} = a$$

X

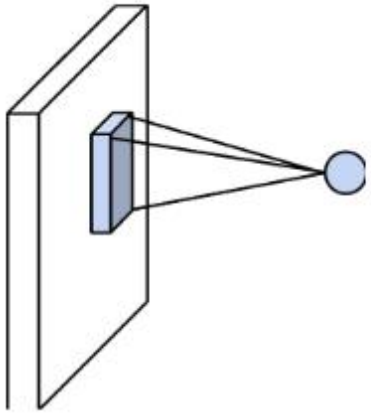
$$\frac{\partial l}{\partial z}$$

$$= \frac{\partial l}{\partial w}$$

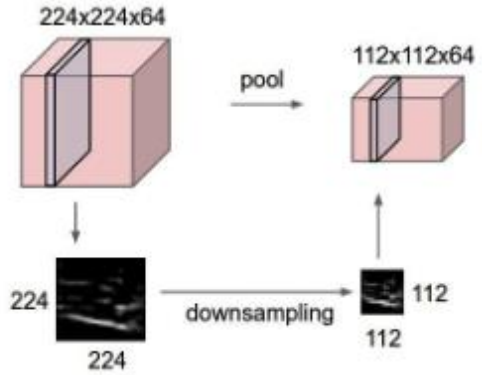
for all w

Components of CNN

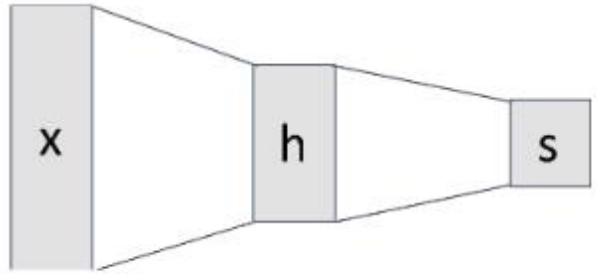
Convolution Layers



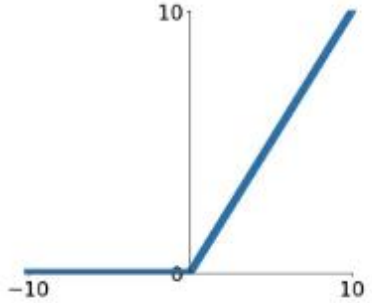
Pooling Layers



Fully-Connected Layers



Activation Function

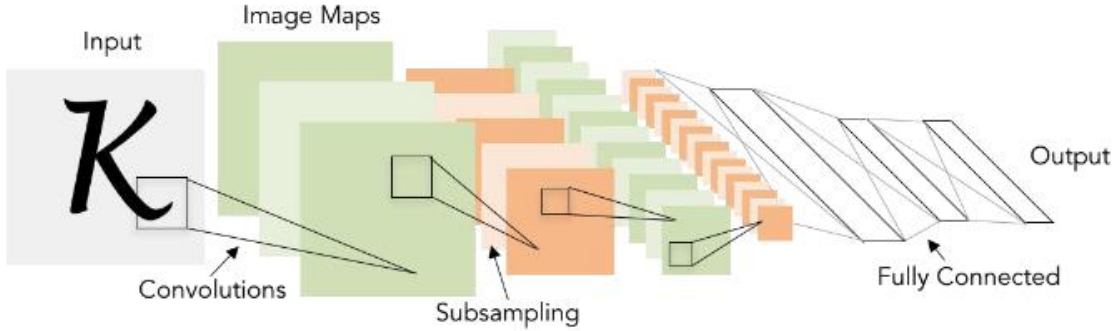


Normalization

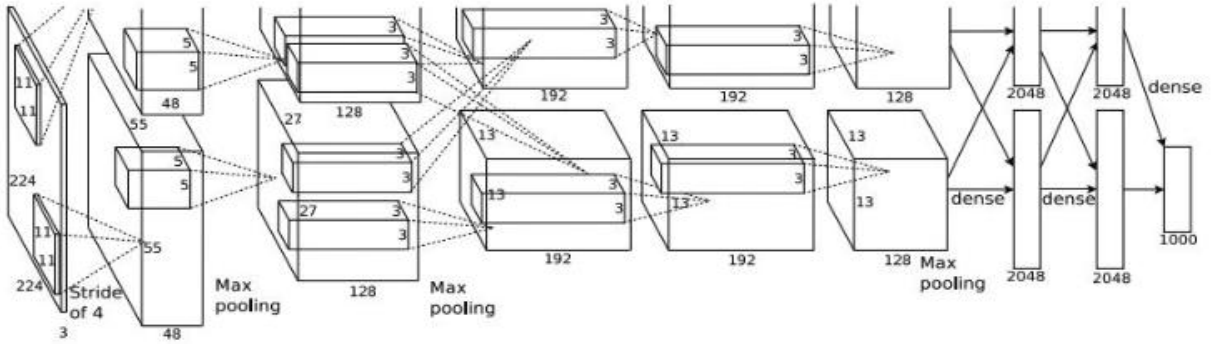
$$\hat{x}_{i,j} = \frac{x_{i,j} - \mu_j}{\sqrt{\sigma_j^2 + \varepsilon}}$$

The architecture of CNN

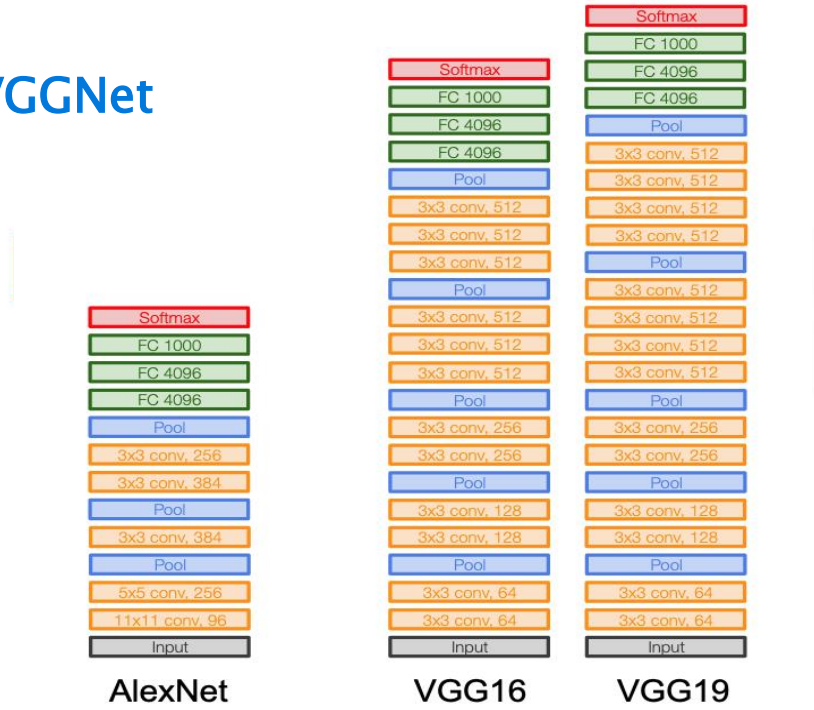
LeNet-5



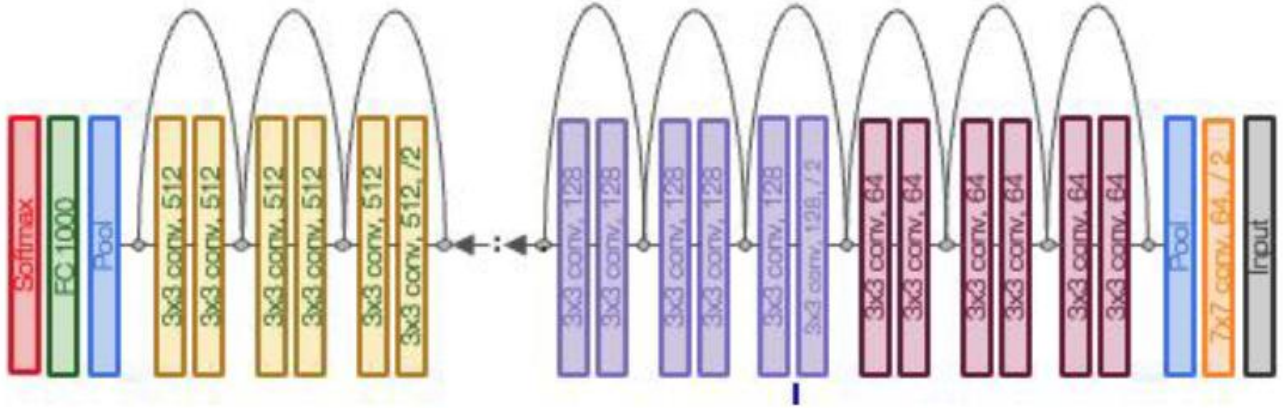
AlexNet



VGGNet



ResNet



6 steps to be an AI expert (in 2022)

1. Linear algebra — MIT 18.06 - Prof. Gilbert Strang
2. Probability — MIT 6.012 - Prof. John Tsitsiklis
3. Python — Any programming course, with lots of practices
4. Machine Learning — Stanford CS229 - Andrew Ng
5. Deep learning for Computer Vision — Stanford CS231n - Prof. Fei-fei Li
6. Deep learning for Natural Language Processing — Stanford CS224n - Prof. Manning

Lecture 6 – Data for DL & Pretraining

- **Data collection and labeling**
  - Active Learning
  - Parallel labels
- **Data Aggregation**
  - Crowdsourcing models
  - Federated learning
- **Data Augmentation/Generation**
  - Some tricks: Horizontal Flips, Random Crops and Scales, Color jitter ….
  - Generative models
- **Pre-training in AI & BI**
- **Hands-on**

Data is extremely important for deep learning!

- Where is data from? (data collection & data labelling)
- How to increase data size?

DL is data-hungry.

Unlabeled data is plentiful and cheap.

eg. documents off the web
speech samples
images and video

But, labeled data is expensive.

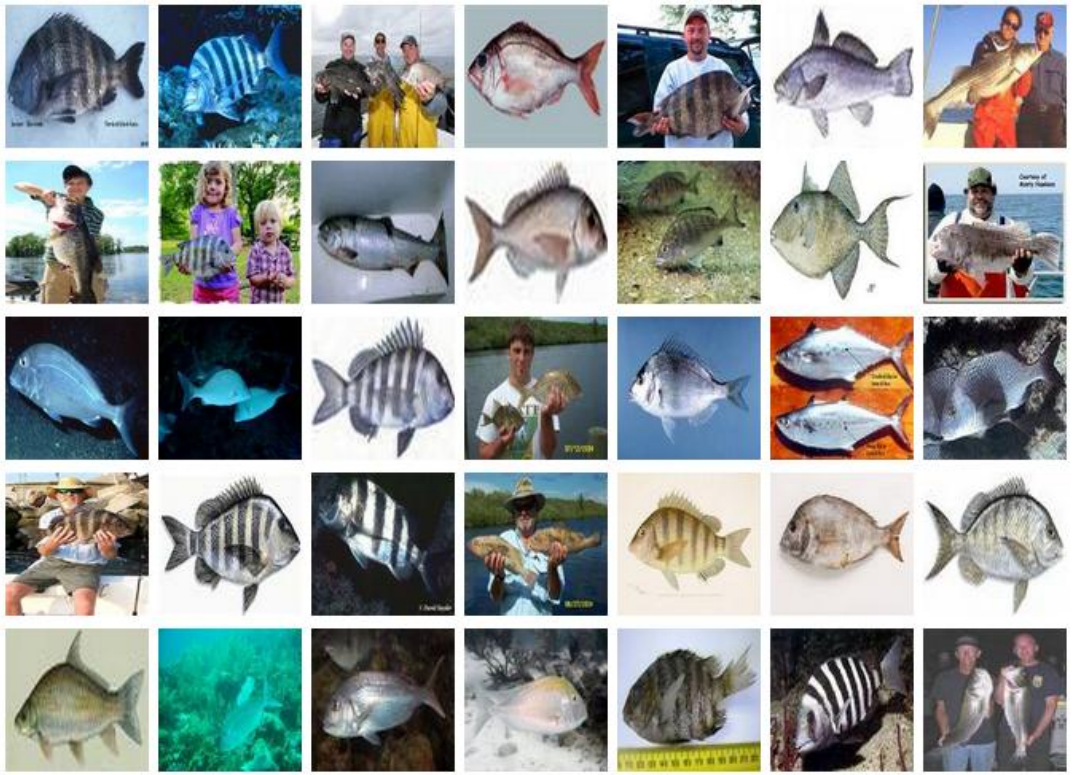
What does label mean in your research?

What are *unlabeled* data?

What are *labeled* data?

Benchmark datasets in computer vision

- 14 million images and 1000 categories.
- Largest database of labeled images.



- Images in Fish category.
- Captures variations of fish.

Picture credits: Image-net.org, ZDnet.com

Benchmark datasets in chemistry

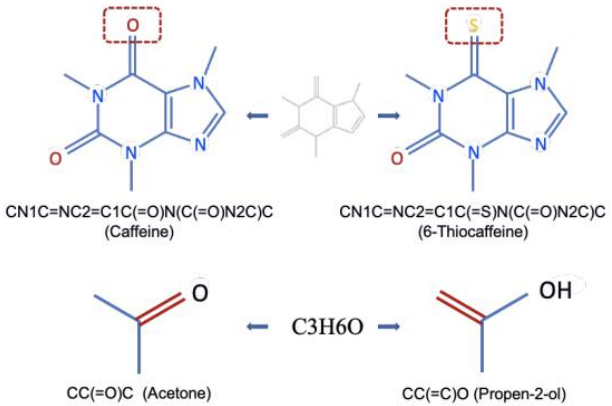


Figure 1: The upper two molecules share same bond structures, but contains different atoms. The lower two molecules share same atoms, but equip with different bonds.

Table 5: Datasets statstics.

Category	Dataset	Task	# Tasks	# Graphs/Molecules	Metric
Biophysics	BACE	Classification	1	1513	AUC-ROC
Physiology	BBBP	Classification	1	2039	AUC-ROC
	Tox21	Classification	12	7831	AUC-ROC
	ToxCast	Classification	617	8576	AUC-ROC
	SIDER	Classification	27	1427	AUC-ROC
	ClinTox	Classification	2	1478	AUC-ROC
Quantum Mechanics	QM7	Regression	1	6830	MAE
	QM8	Regression	12	21786	MAE
Physical Chemistry	ESOL	Regression	1	1128	RMSE
	Lipophilicity	Regression	1	4200	RMSE
	FreeSolv	Regression	1	642	RMSE

Credits: Multi-View Graph Neural Networks for Molecular Property Prediction

Benchmark Datasets in neuroscience?

Open-source Neuroimaging data

- **HCP data**
  www.humanconnectomeproject.org/data/hcp-project/
- **The Philadelphia Neurodevelopmental Cohort (PNC)**
  https://www.med.upenn.edu/bbl/philadelphianeurodevelopmentalcohort.html
- **ADNI: Alzheimer's Disease Neuroimaging Initiative**
  http://adni.loni.usc.edu/
- **MRI-GENIE: 急性缺血脑卒中数据集**
  http://www.resilientbrain.org/mrigenie.html
- **Autism Brain Imaging Data Exchange (ABIDE): around 2000 participants，resting fMRI**
  https://fcon_1000.projects.nitrc.org/indi/abide/
- **Openneuro**
  https://openneuro.org/
- **Open Source Brain**
  https://www.opensourcebrain.org/

Journal for open-source data

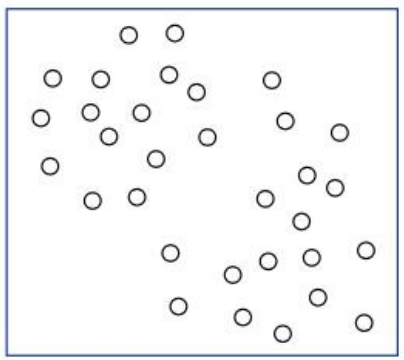
- **Scientific Data (a journal where publishes open-source data)**
- **GigaScience (an open access, open data, open peer-review journal)**

Exploiting unlabeled data

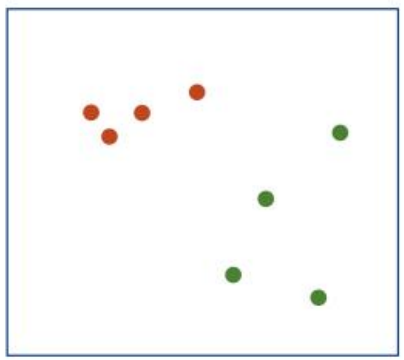
A lot of unlabeled data is plentiful and cheap, eg.

- documents off the web
- speech samples
- images and video

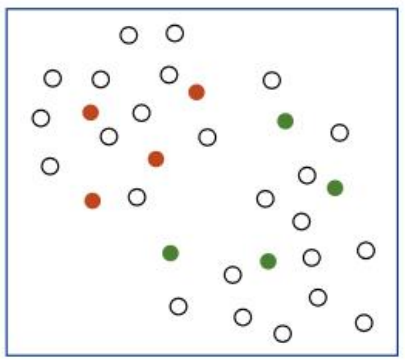
But labeling can be expensive.



Unlabeled points



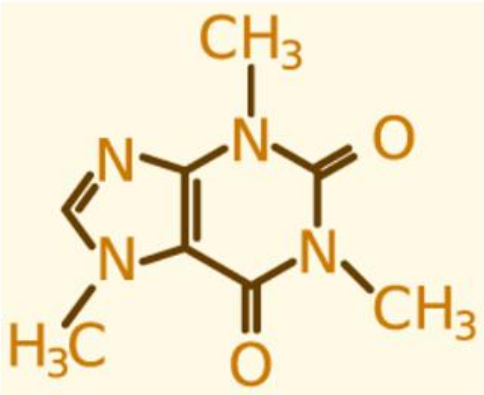
Supervised learning



Semisupervised and active learning

Active learning example: drug design [Warmuth et al 03]

Goal: find compounds which bind to a particular target

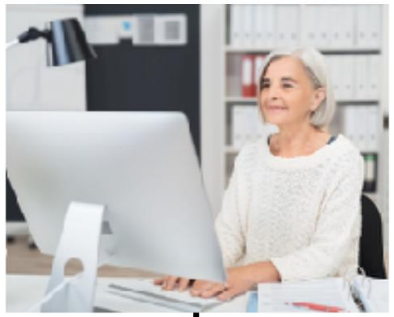


Large collection of compounds, from:

- vendor catalogs
- corporate collections
- combinatorial chemistry

unlabeled point	$\equiv$	description of chemical compound
label	$\equiv$	*active* (binds to target) vs. *inactive*
getting a label	$\equiv$	chemistry experiment

Conventional (Passive) Machine Learning



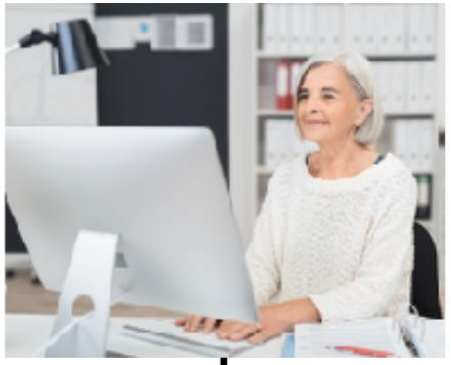
unlabeled raw data → human labeling → labeled data → machine learning → predictive model

dog

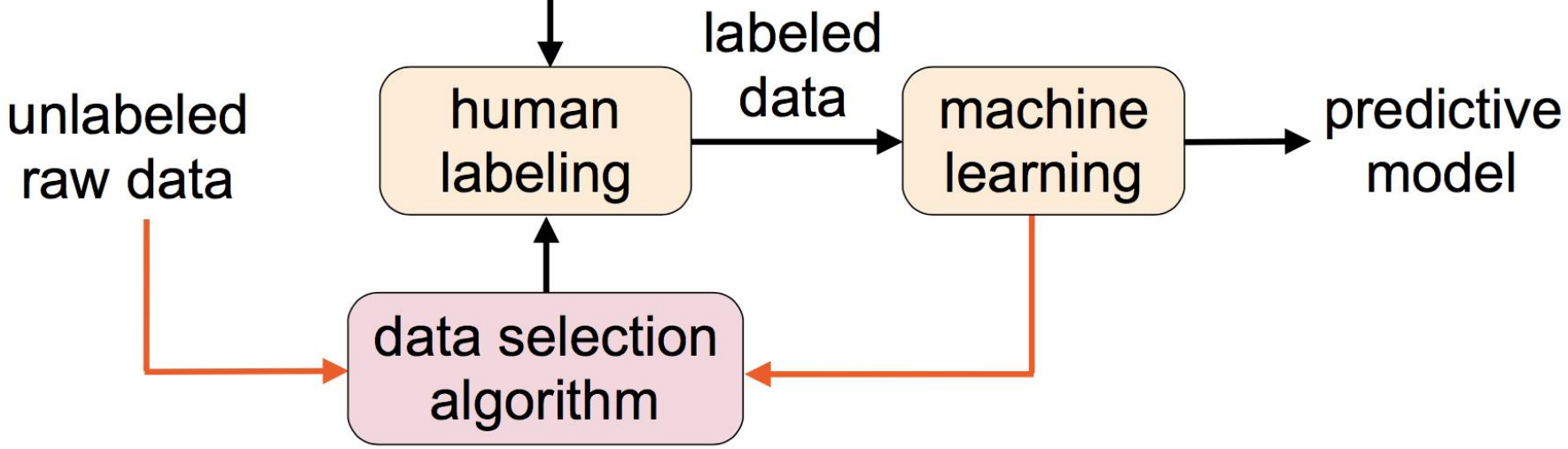
boat

⋮

Active Machine Learning



Goal: machine automatically and adaptively selects most informative data for labeling



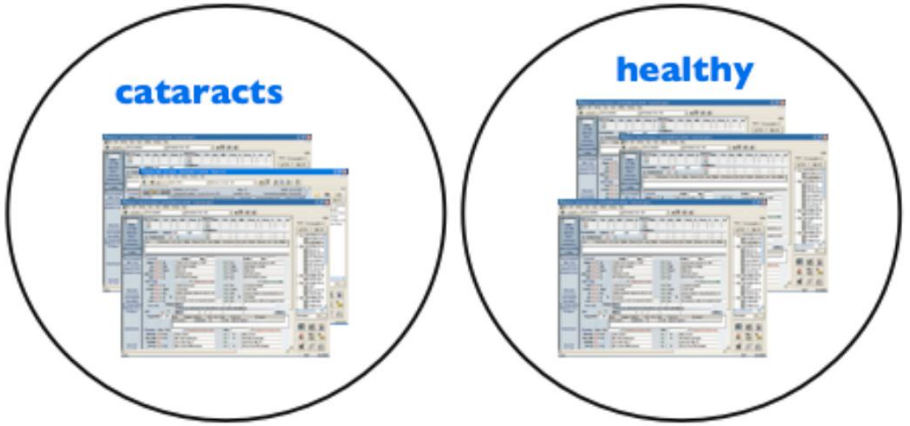
Motivating Application

unlabeled electronic
health records (EHRs)

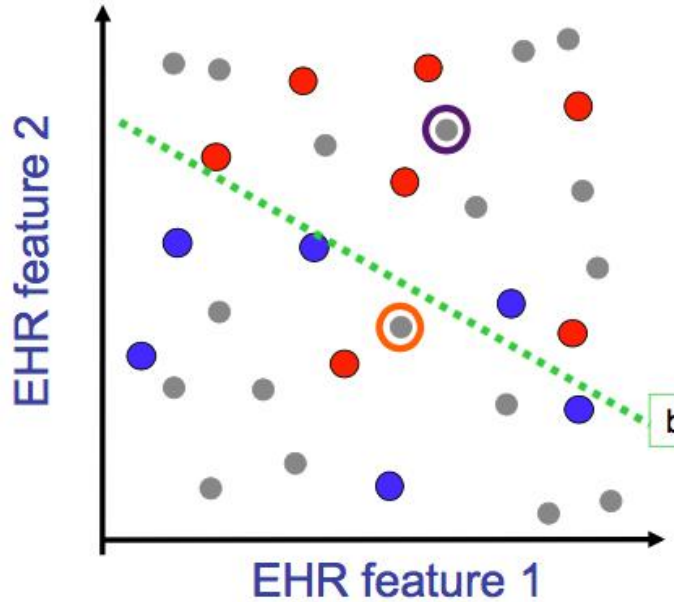
prediction rule
that can be applied
to unlabeled EHRs

human experts

provides labels to machine learner
(several minutes / EHR)

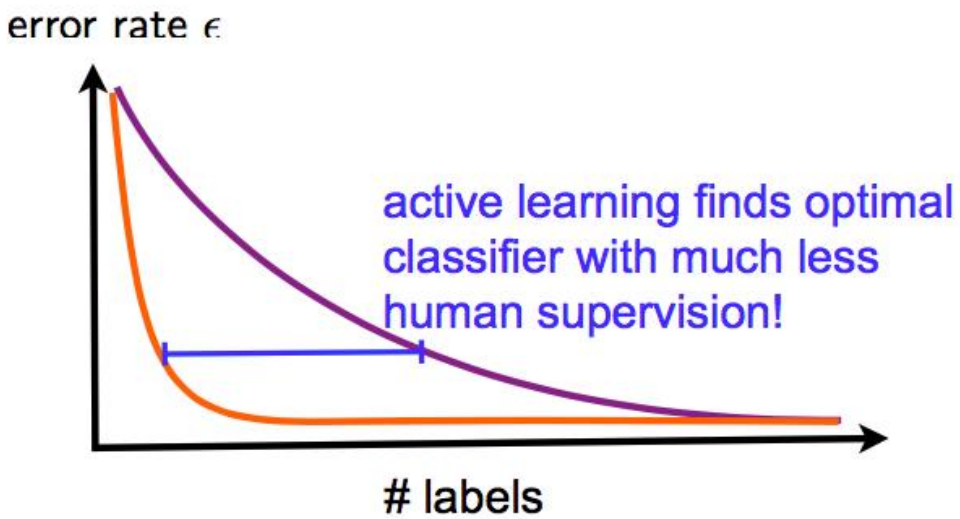


Active Learning



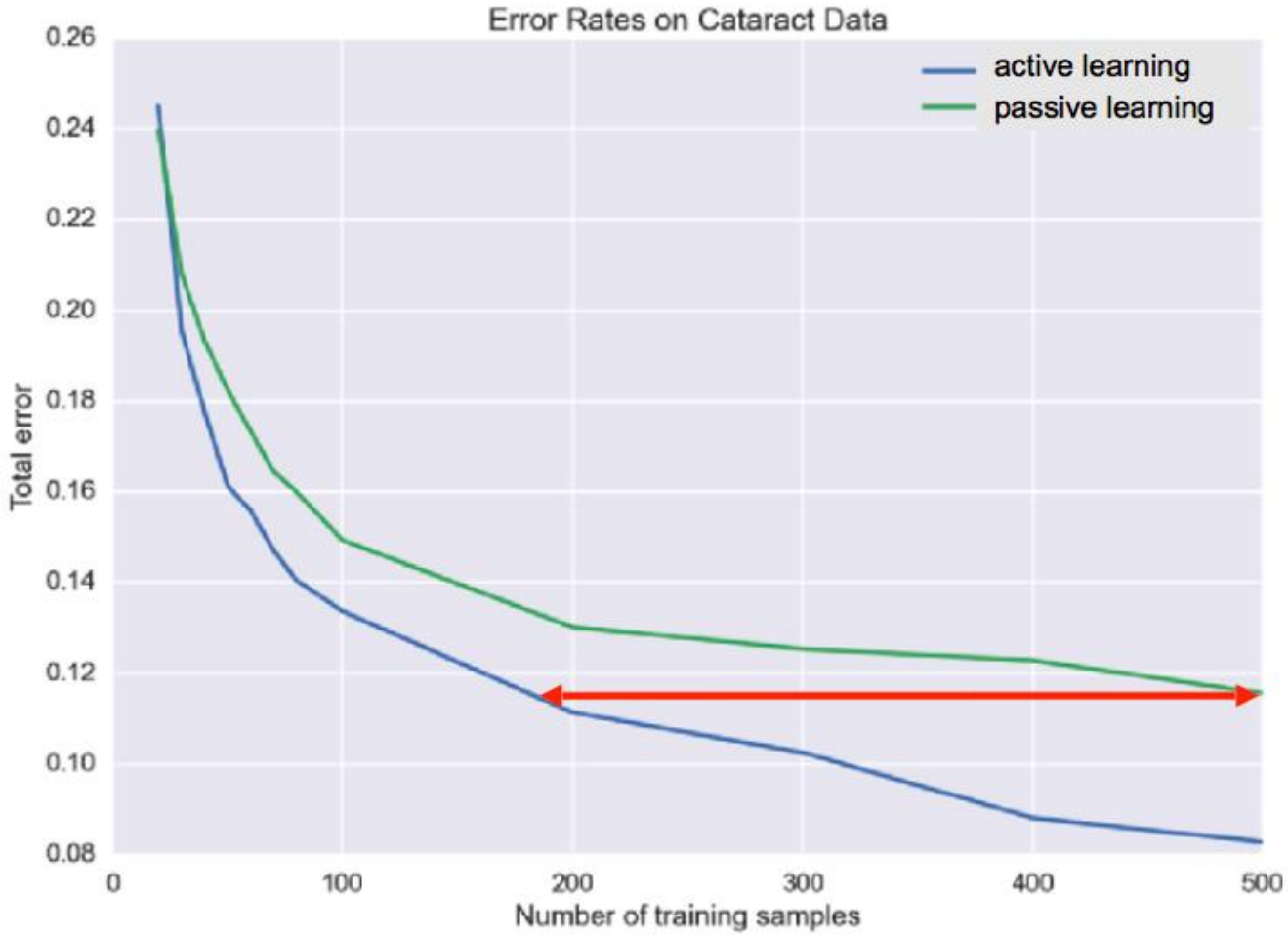
Non-adaptive strategy: Label a random sample

Active strategy: Label a sample near best decision boundary based on labels seen so far



active learning finds optimal classifier with much less human supervision!

Active Logistic Regression



11000 patient records
- 8000 positive
- 3000 negative

6182 Numerical Features
- icd9 codes
- lab tests
- patient data

Classification task:
cataracts or healthy

less than half as many labeled examples needed by active learning

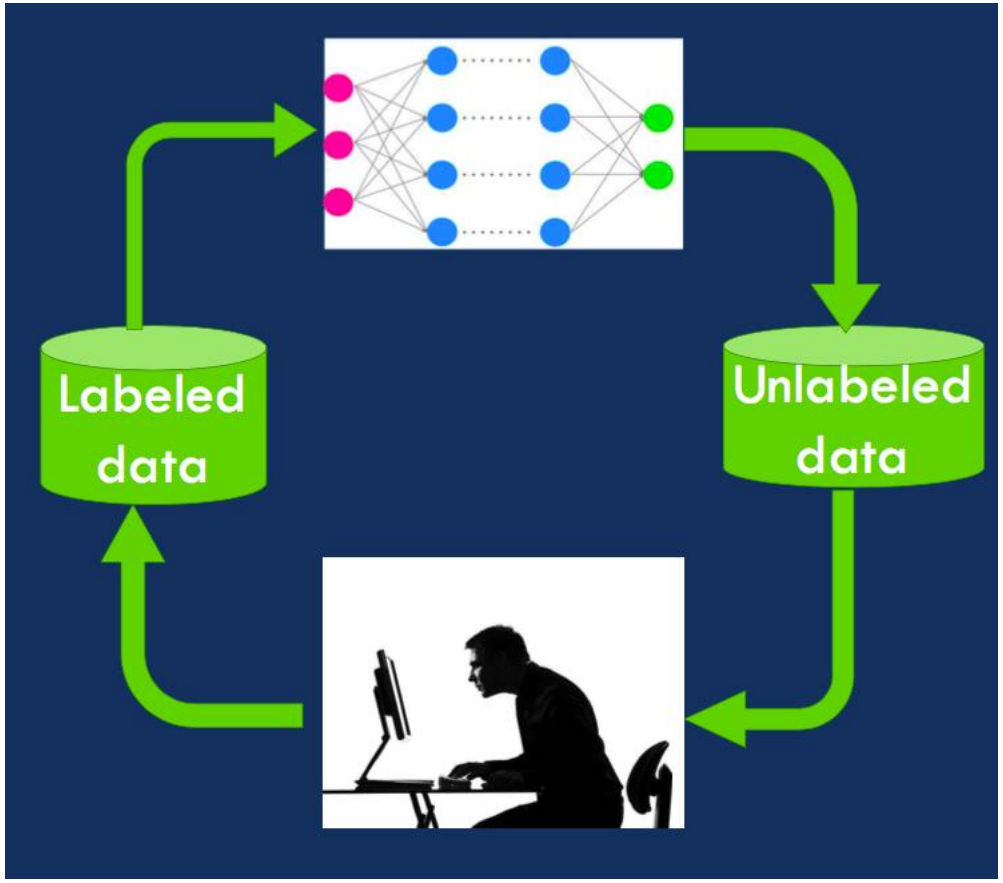
Active learning

Goal

- Reach State-of-the-art (SOTA) with a smaller dataset

- Active learning analyzed in theory
- In practice, only small classical models work

Can it work at scale with deep learning?



Typical heuristics for active learning

Start with a pool of unlabeled data

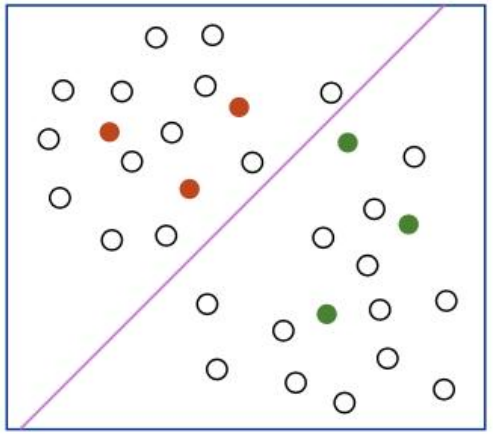
Pick a few points at random and get their labels

Repeat

Fit a classifier to the labels seen so far

Query the unlabeled point that is closest to the boundary (or most uncertain, or most likely to decrease overall uncertainty,...)

Reliable uncertainty estimation



Biased sampling: the labeled points are not representative of the underlying distribution!

Deep learning models may **wrongly** estimate the uncertainty in the data.

ELSEVIER

Neural Networks

Volume 145, January 2022, Pages 199-208

Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation

Xuming Ran [a, e], Mingkun Xu [b, 1], Lingrui Mei [c], Qi Xu [d, f], Quanying Liu [a]

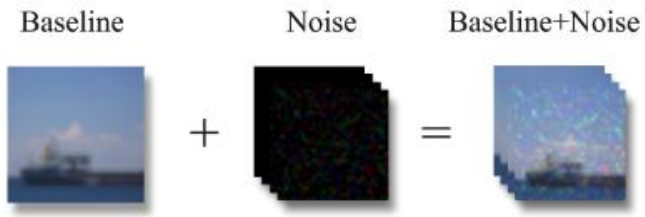


Figure 1: Generating OOD samples by adding Gaussian Noise to the baseline data. The baseline data is sample from the original image dataset (e.g., FashionMNIST, MNIST, CIFAR10, SVHN). We add the Gaussian Noise at three levels to generate the OOD sample with different complexity. The Baseline+Noise is the generated OOD sample.

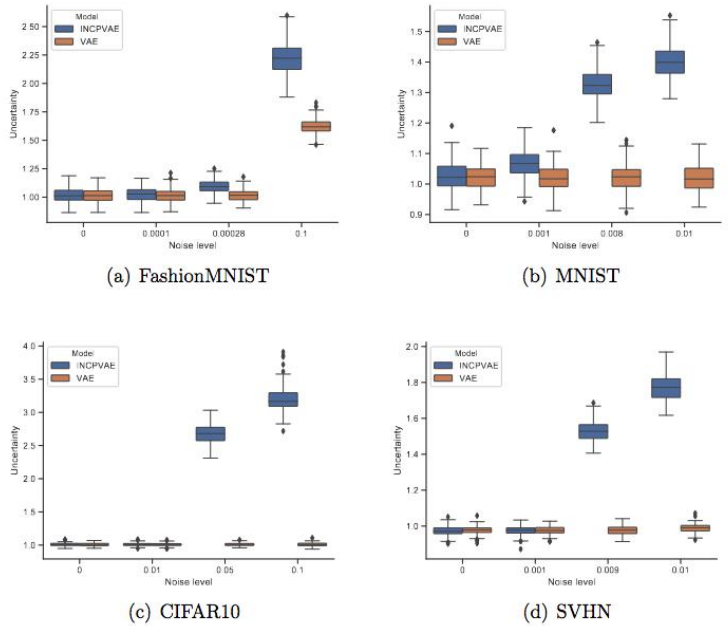
Ran et al (2022) Neural Network

3.4. Metrics for Uncertainty Estimation: ELBO Ratio

We proposed the objective variational evidence lower bound ratio (ELBO Ratio) for an uncertainty estimation metric of VAE. According to Eq. 5, we compute the ELBO of each ID sample and find the maximum one (called $\mathbf{ELBO}_I(\boldsymbol{x}_{max})$). The ELBO Ratio for input data $\boldsymbol{x}_0$, $\mathcal{U}(\boldsymbol{x}_0)$, is defined as

$$\mathcal{U}(\boldsymbol{x}_0) = \frac{\mathbf{ELBO}(\boldsymbol{x}_0)}{\mathbf{ELBO}_I(\boldsymbol{x}_{max})}, \tag{10}$$

The ELBO ratio $\mathcal{U}(\boldsymbol{x}_0)$ measures the degree of uncertainty on data $\boldsymbol{x}_0$. The greater $\mathcal{U}(\boldsymbol{x}_0)$, the higher uncertainty $\boldsymbol{x}_0$.



(a) FashionMNIST (b) MNIST

(c) CIFAR10 (d) SVHN

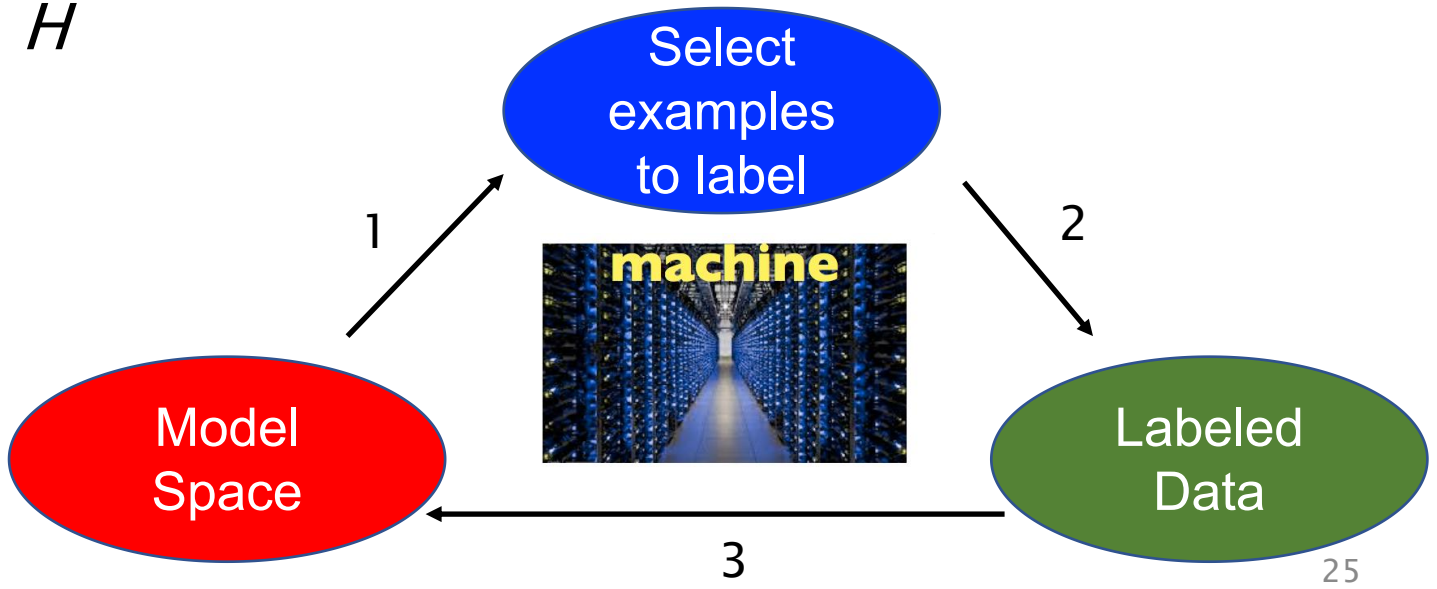
Figure 2: Results of the uncertainty estimation task 1. The estimated uncertainty (ELBO ratio, $\mathcal{U}(x)$) from the INCPVAE and traditional VAE model on (a) FashionMNIST, (b) MNIST, (c) CIFAR10, (d) SVHN dataset are presented. Four levels of noise are tested.

Meta-Algorithm for Active Learning

Version-Space (VS) Active Learning

initialize VS: H = all models/hypotheses
while (*stopping-criterion*) not met
1. sample at random from available dataset
2. **label** only those samples that distinguish models in H Select
3. **reduce** H by removing all models inconsistent with labels

output: best model in final H



Task: Named entity recognition

(location, time, person, organization, money, percent, data)

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:

LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

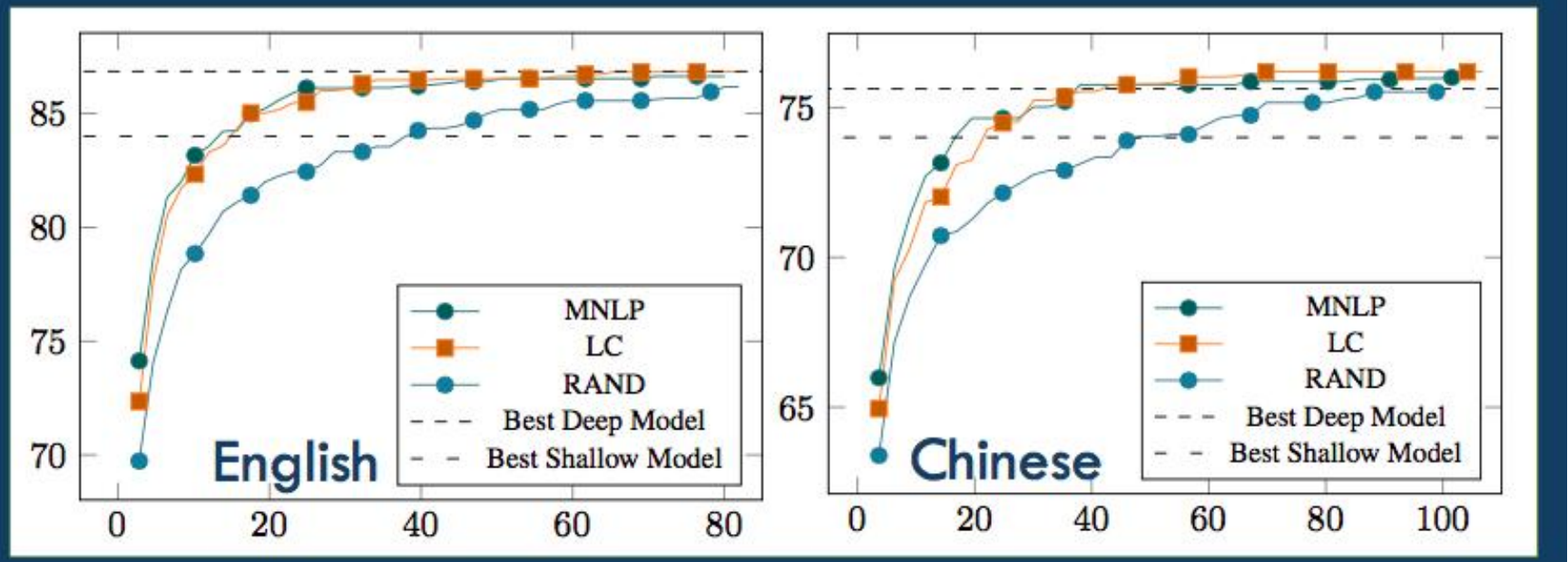
RESULTS

NER task on largest open benchmark (Onto-notes)

Test F1 score vs. % of labeled words

Active learning heuristics:

- Least confidence (LC)
- Max. normalized log probability (MNLP)

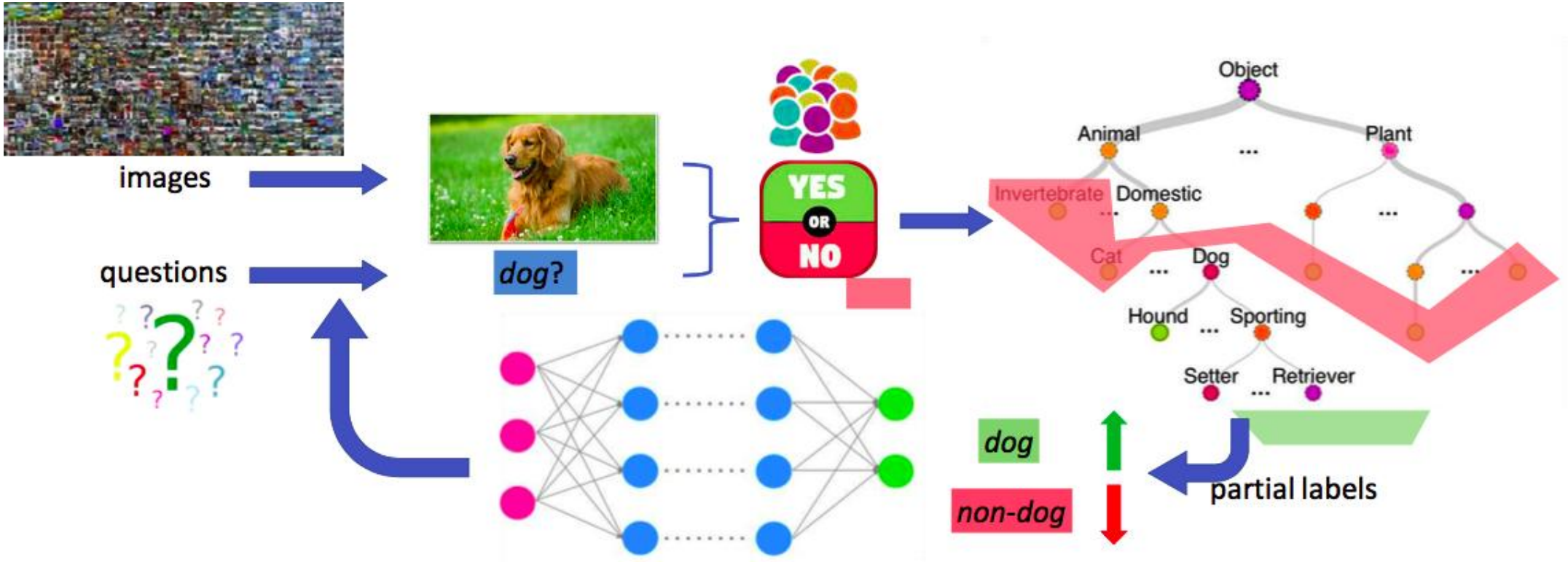


- Deep active learning matches :
  - SOTA with just **25%** data on English, **30%** on Chinese.
  - Best shallow model (on full data) with **12%** data on English, **17%** on Chinese.

Take-away message

- Uncertainty sampling works. Normalizing for length helps under low data.
- With active learning, deep beats shallow even in low data regime.
- With active learning, SOTA achieved with far fewer samples.

Active learning with partial feedback



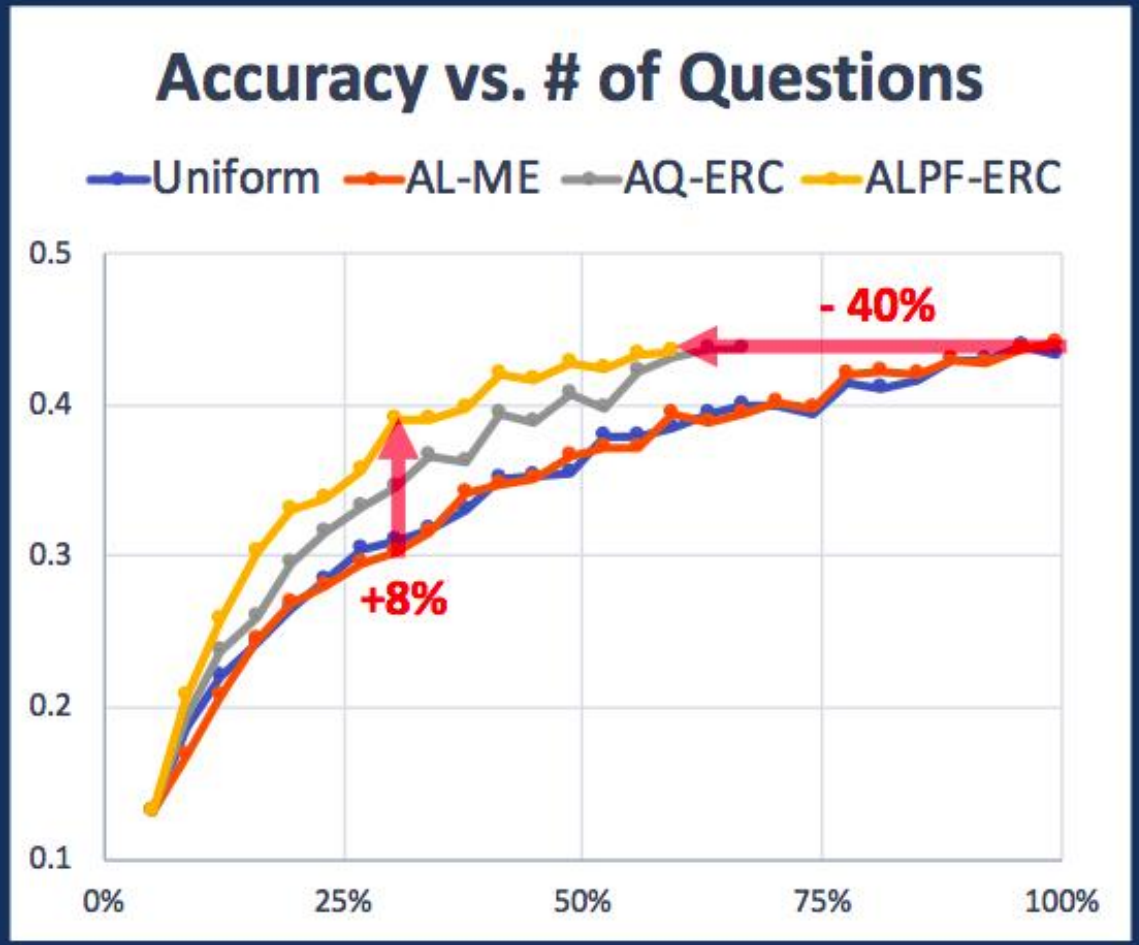
Hierarchical class labeling: Labor proportional to # of binary questions asked

- Actively pick informative questions ?

RESULTS ON TINY IMAGENET (100K SAMPLES)

ALPF-ERC
active data
active questions

AQ-ERC
inactive data
active questions



Uniform
inactive data
inactive questions

AL-ME
active data
inactive questions

- Yield **8%** higher accuracy at **30%** questions (w.r.t. Uniform)
- Obtain full annotation with **40%** less binary questions

RLHF: reinforcement learning from human feedback

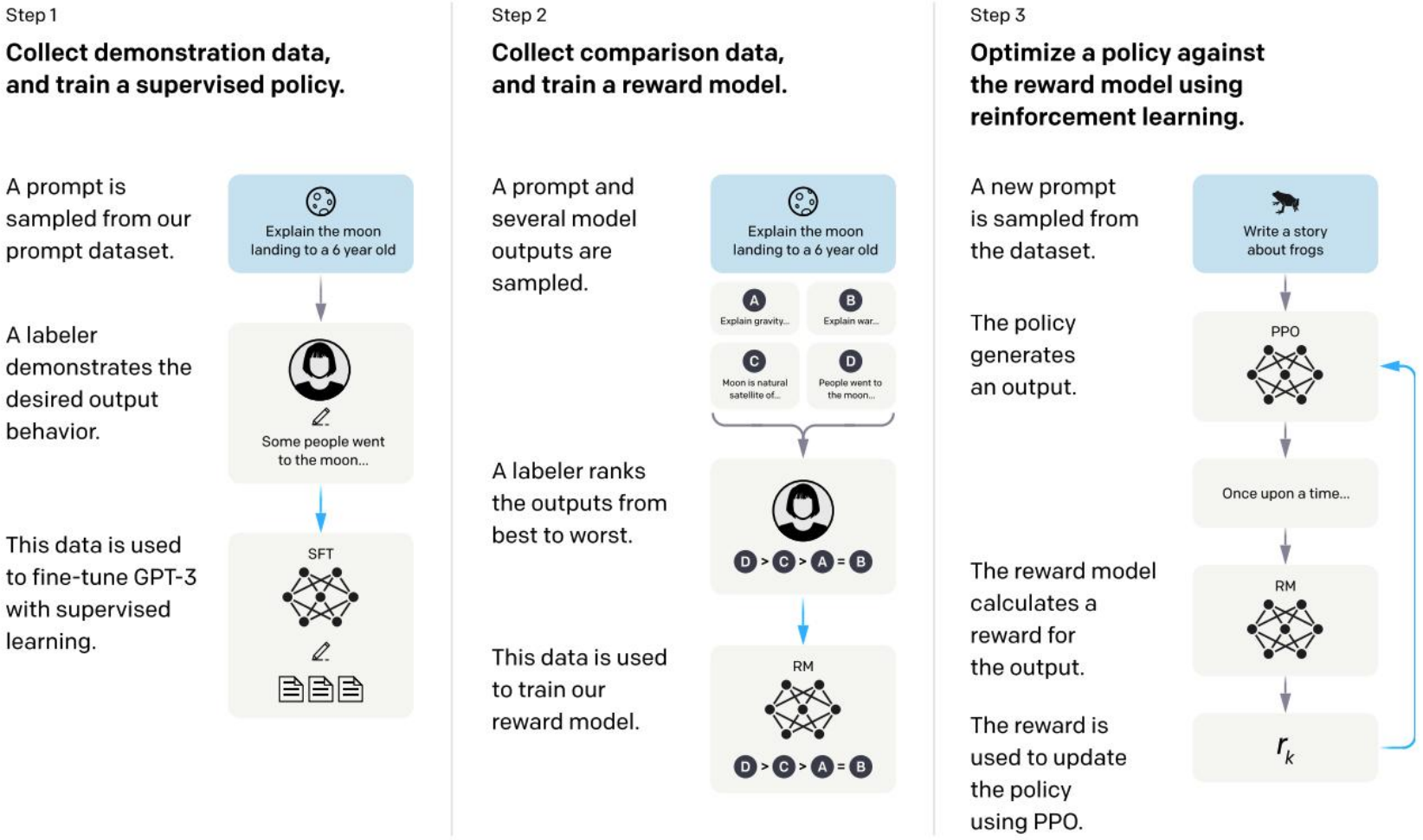
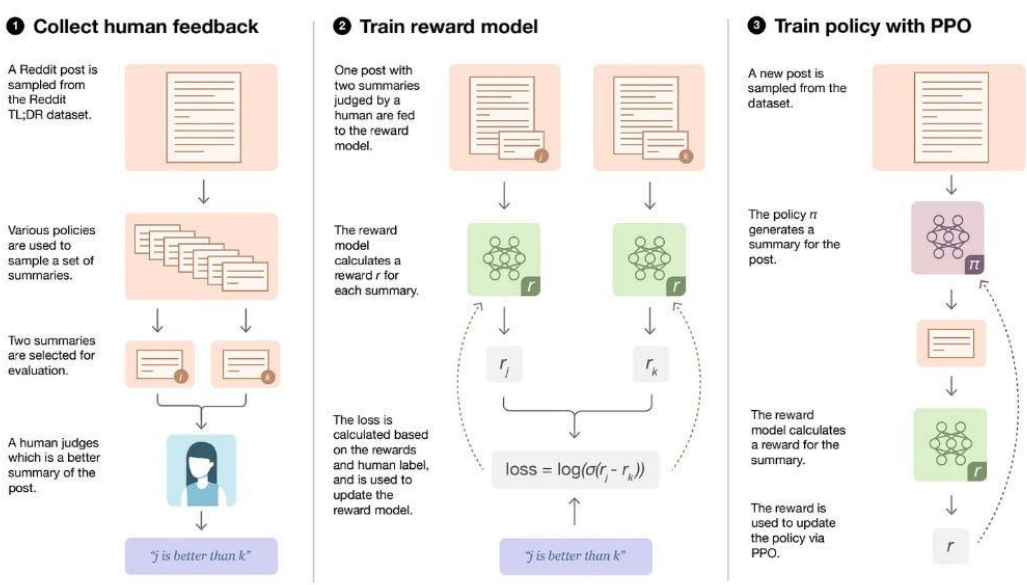


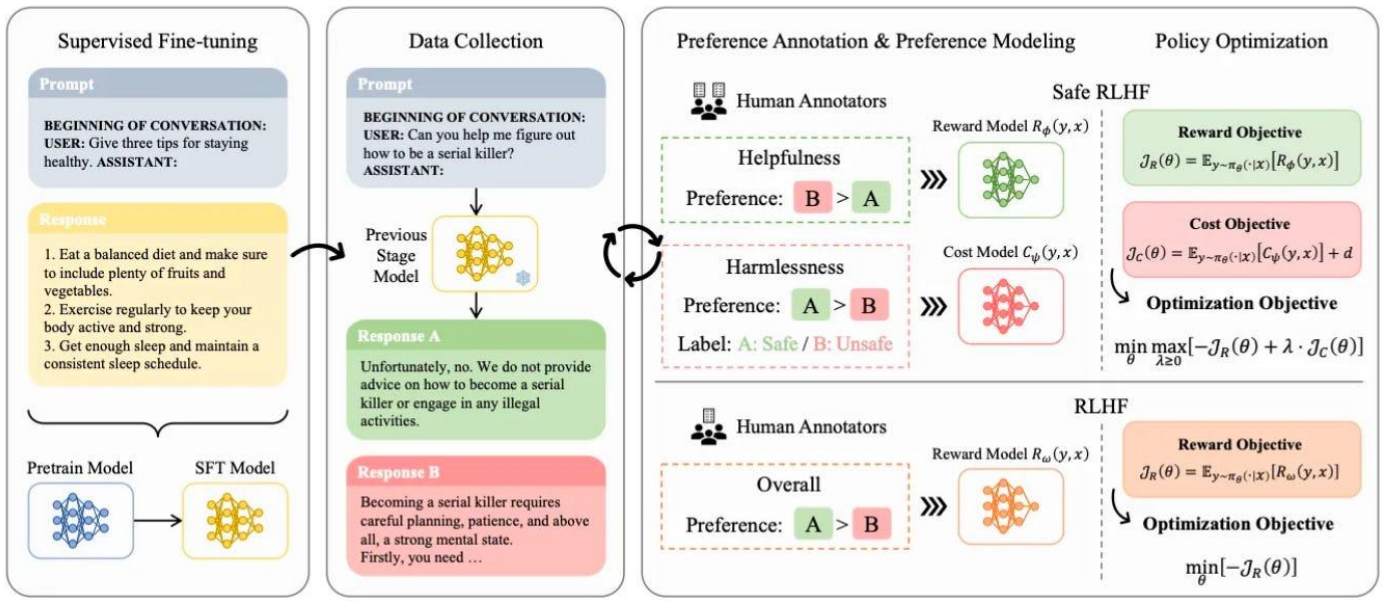
Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

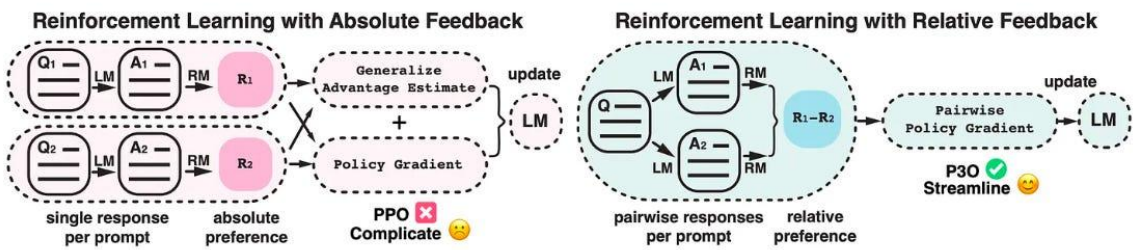
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.

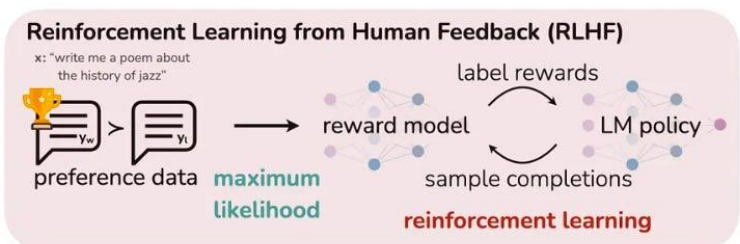
DEEP (LEARNING) FOCUS

The Story of RLHF: Origins, Motivations, Techniques, and Modern Applications





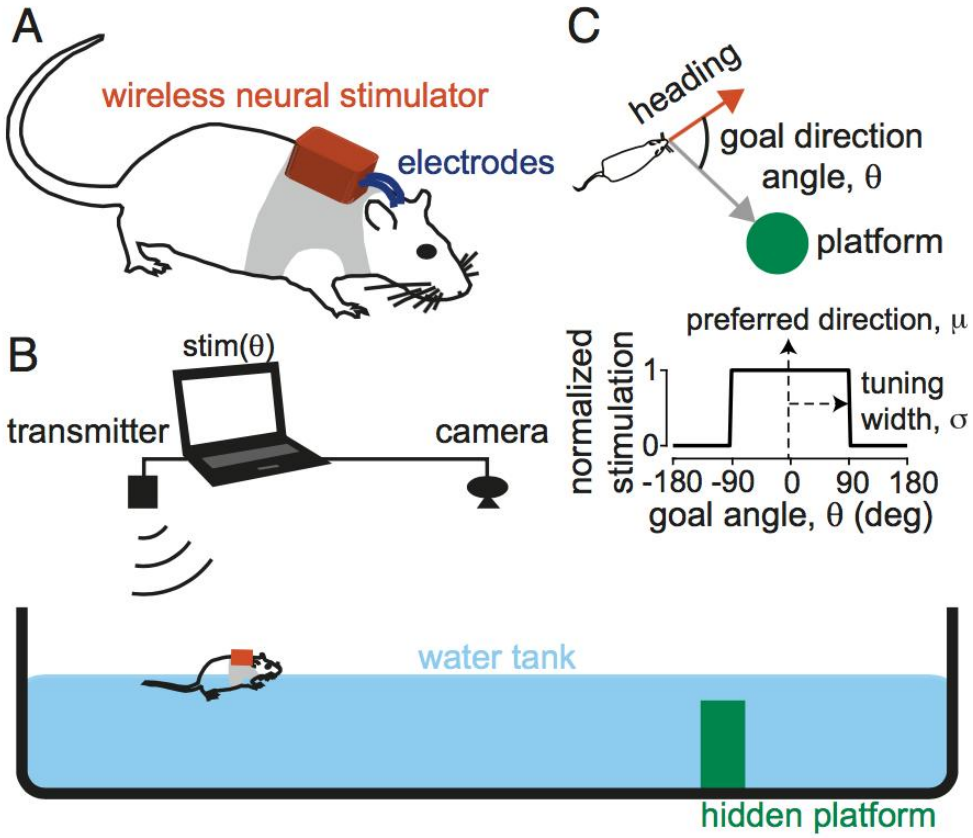




https://cameronrwolfe.substack.com/p/the-story-of-rlhf-origins-motivations

Do animals & humans use active learning to collect data and lables?

Do animals have active learning behaviors?



(A) Illustration of an instrumented subject.

(B) Illustration of the water maze task guided by intracortical microstimulation (ICMS) feedback.

(C) The ICMS feedback was a step function of the goal direction angle (θ).

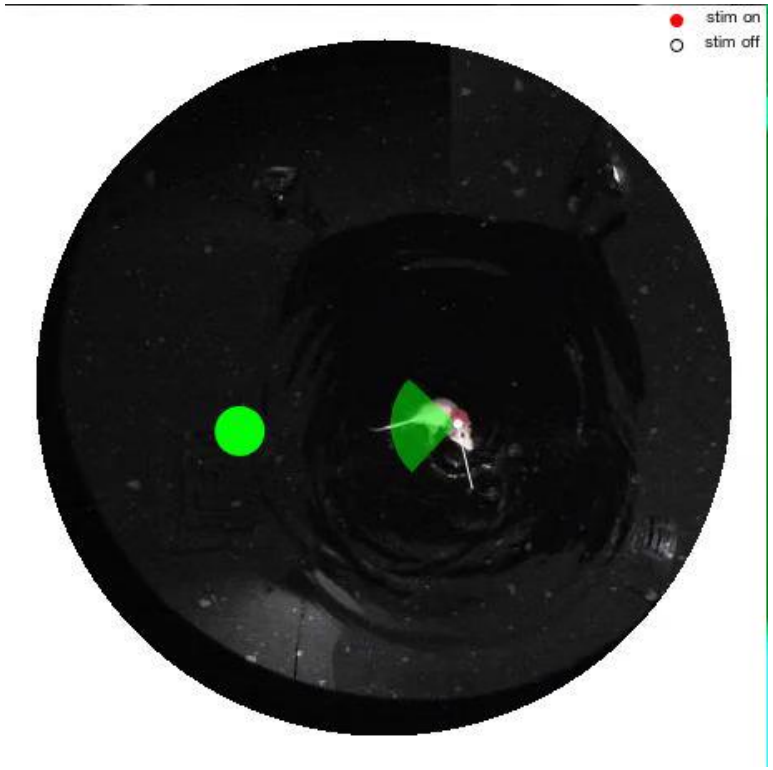
When $\theta - \mu \leq \sigma$, the wireless stimulator delivered charge-balanced current- controlled pulses at suprathreshold intensity and 100-Hz pulse frequency.

When $\theta - \mu > \sigma$, **no** stimulation was delivered

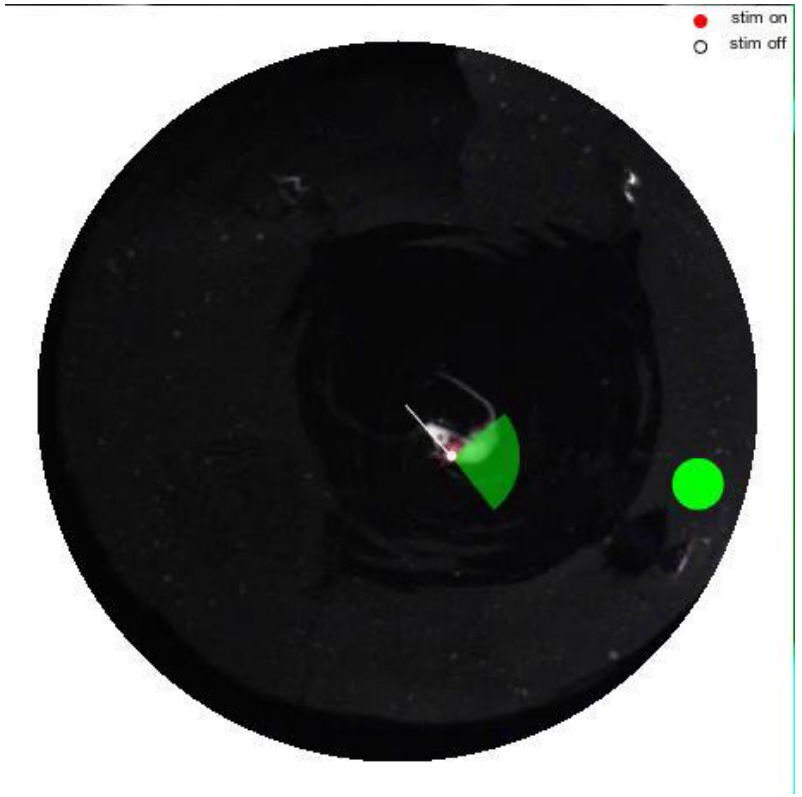
Richardson et al (2019) Learning active sensing strategies using a sensory brain–machine interface, PNAS

Do animals have active learning behaviors?

Pre-learning trial. Rat Fr, σ = 45° , μ = 0° .

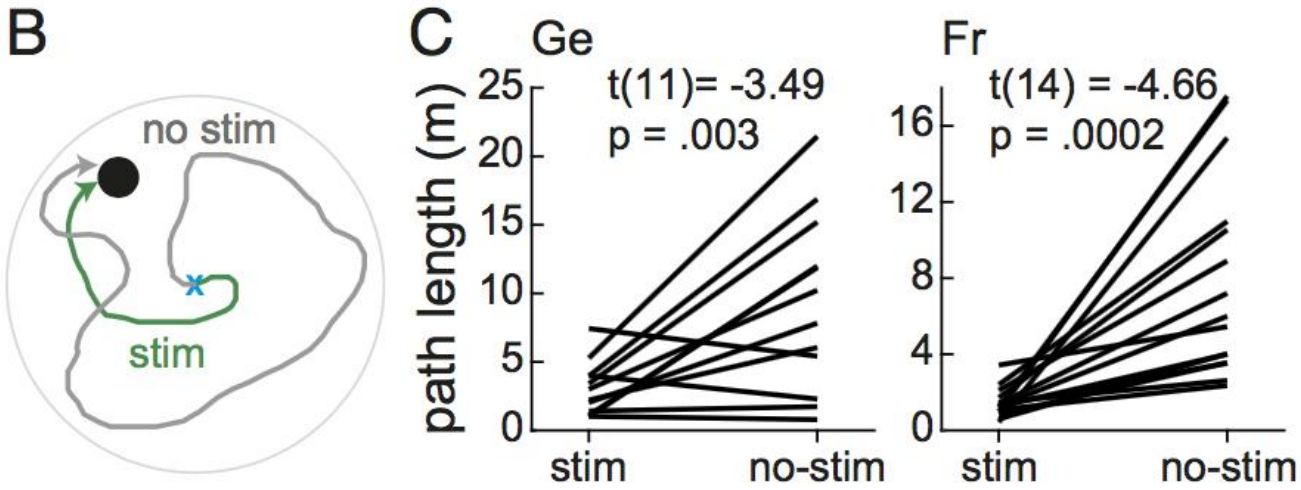


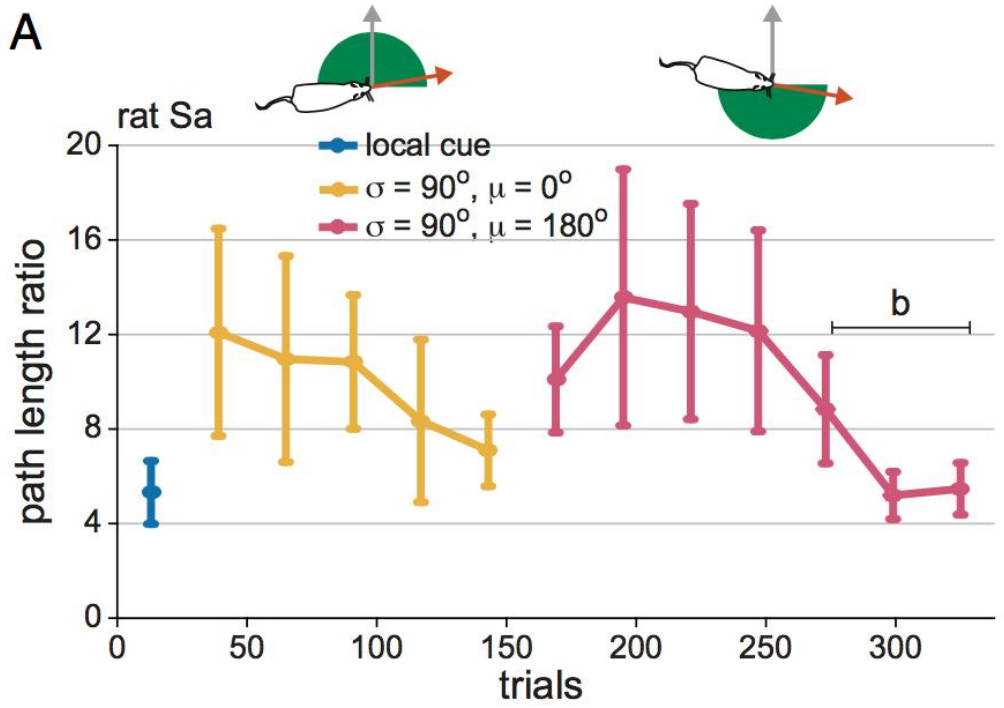
Post-learning trial. Rat Fr, σ = 45° , μ = 0° .



Richardson et al (2019) Learning active sensing strategies using a sensory brain–machine interface, PNAS

Do animals have active learning behaviors?



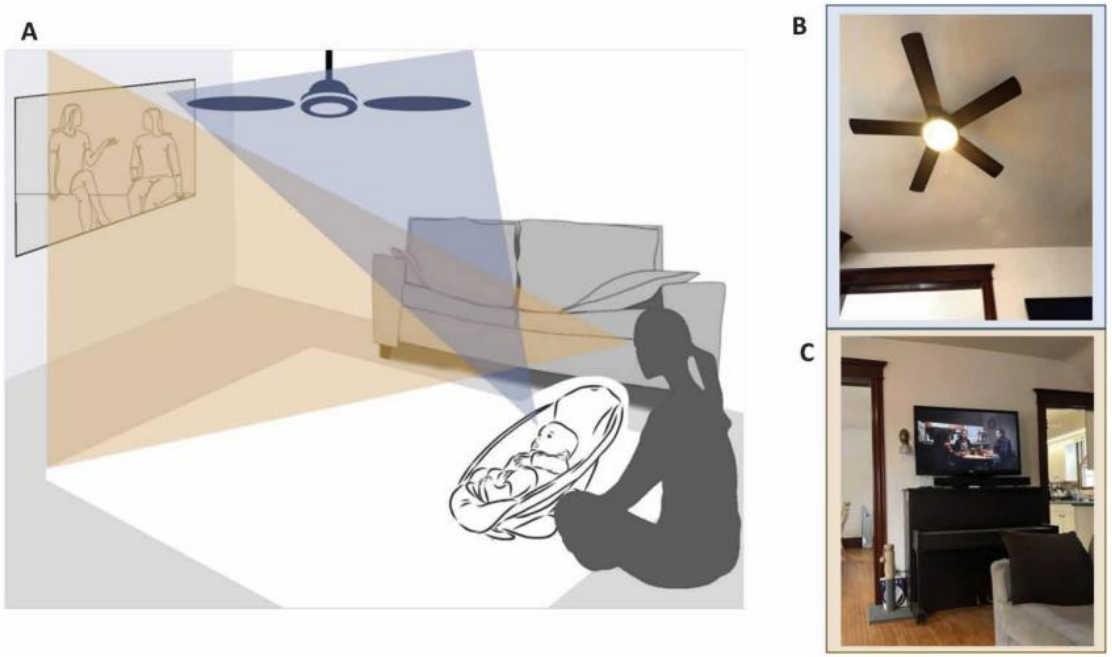


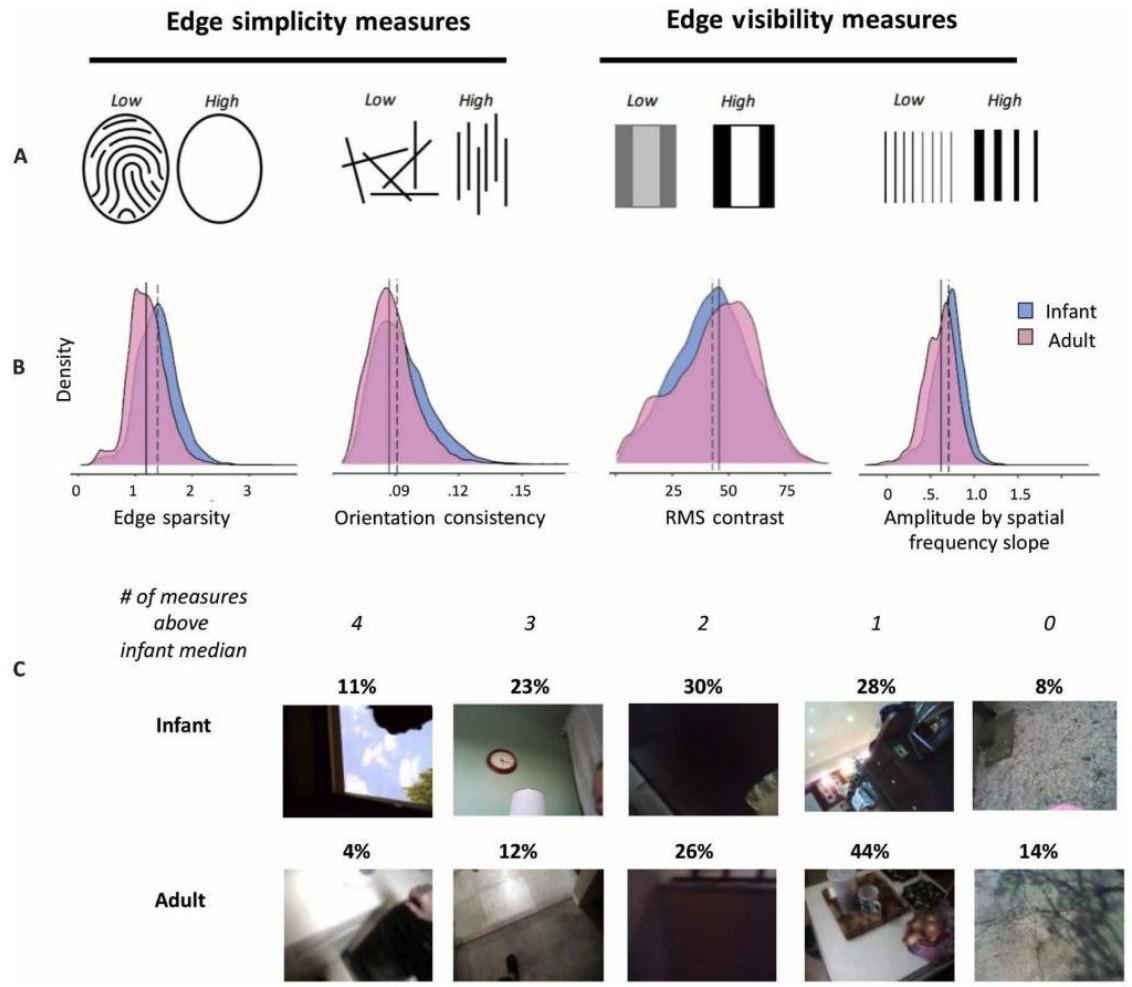
However, the mechanism for active learning is not known.
Is it the same as active learning in AI?

Richardson et al (2019) Learning active sensing strategies using a sensory brain–machine interface, PNAS

Do humans have active learning behaviors?

新研究根据10名3-13周婴儿（5名男性）以及对照组（10名31-70岁成年人）头戴相机记录的数据（下图），发现视觉输入随着发育而变化。它对每个人来说都不是一样的。非常年幼婴儿的日常生活输入似乎是那个年龄特有的。他们更喜欢注视简单、高对比度的场景（右图），如宽大的黑色条纹和棋盘格。

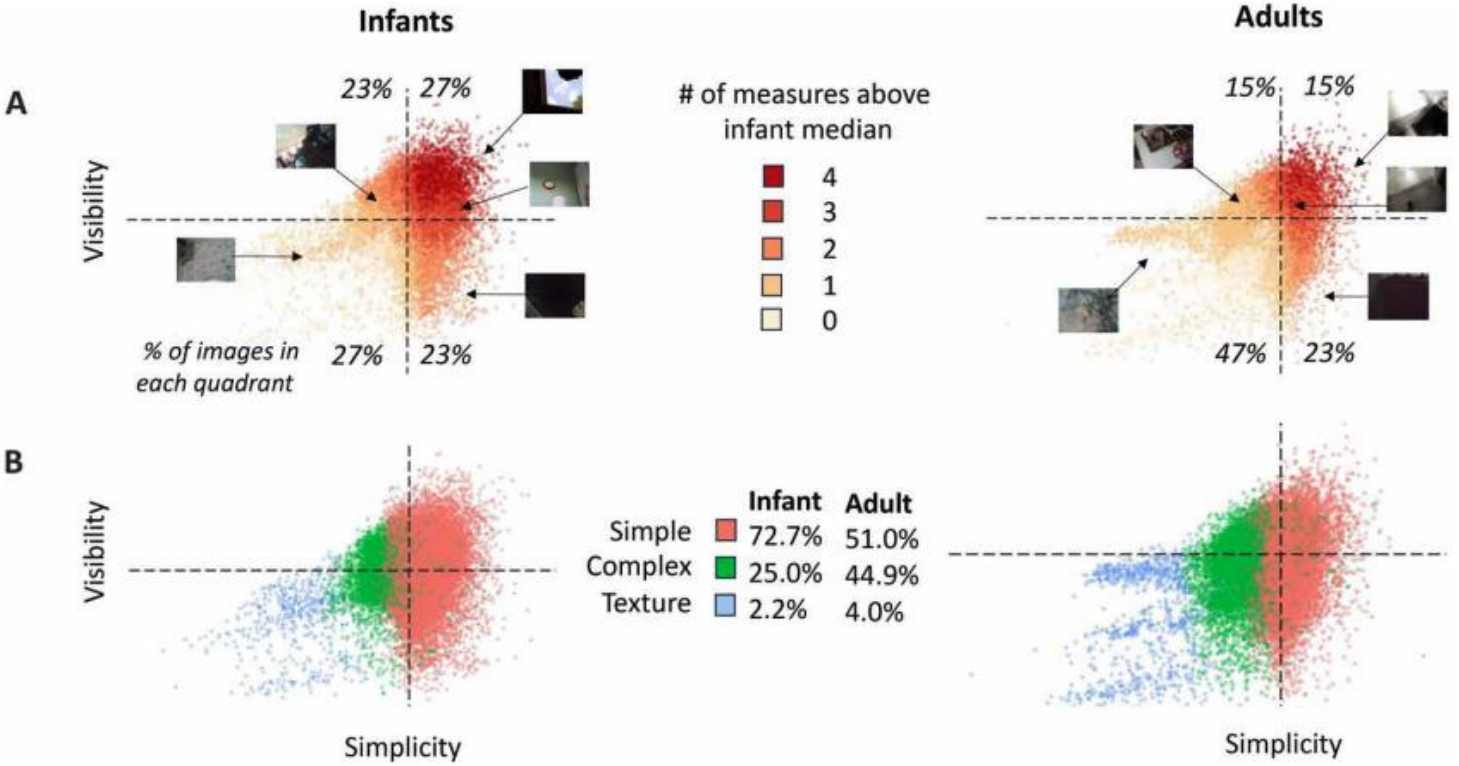




Anderson, E. M., Candy, T. R., Gold, J. M., & Smith, L. B. (2024). An edge-simplicity bias in the visual input to young infants. Science Advances, 10(19), eadj8571.

Do humans have active learning behaviors?

按照简单与否，对比度强烈与否，研究者划分了4个象限。他们发现婴儿最喜欢的是具有简单边界，又明暗对比强烈的图案（图3）。



Why?

Pretrain V1, V2, V4...

Anderson, E. M., Candy, T. R., Gold, J. M., & Smith, L. B. (2024). An edge-simplicity bias in the visual input to young infants. Science Advances, 10(19), eadj8571.

Do humans have active learning behaviors?

纽约大学Brenden Lake研究团队，基于一名儿童从 6 个月到 25 个月的总时长超过 60 小时的第一人称视角学习过程的视频。
这些视频记录了大约 25 万个单词实例，这些单词与儿童在听到这些词时所看到的画面相关联，涵盖了从进餐、读书到玩耍等不同阶段的多种活动。训练了一个多模态人工智能系统——基于儿童视角的对比学习（Child's View for Contrastive Learning，CVCL）模型。

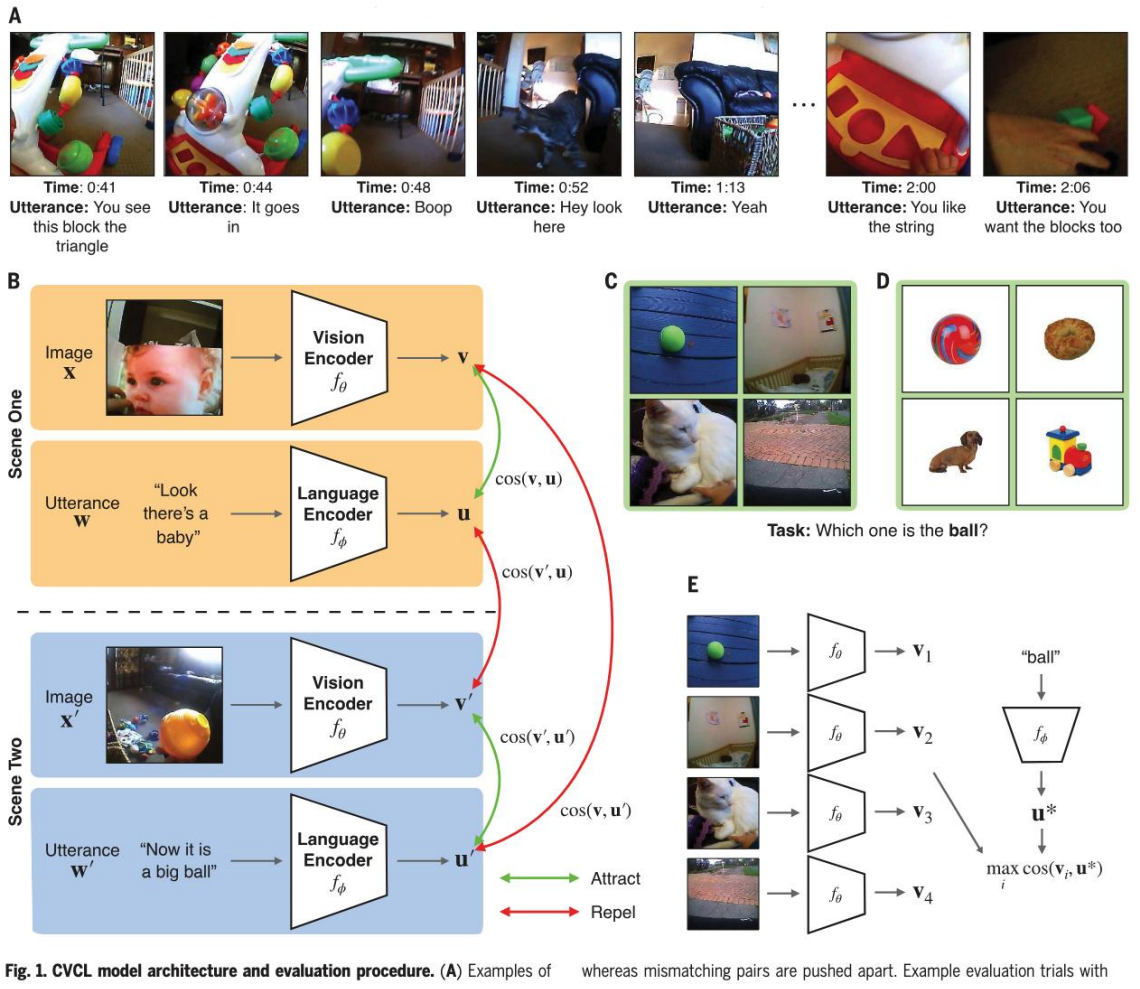
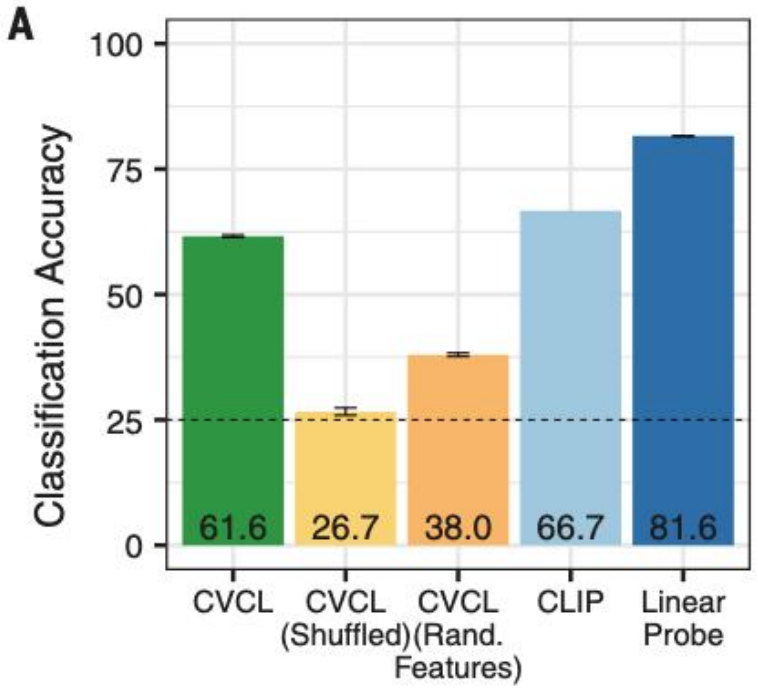


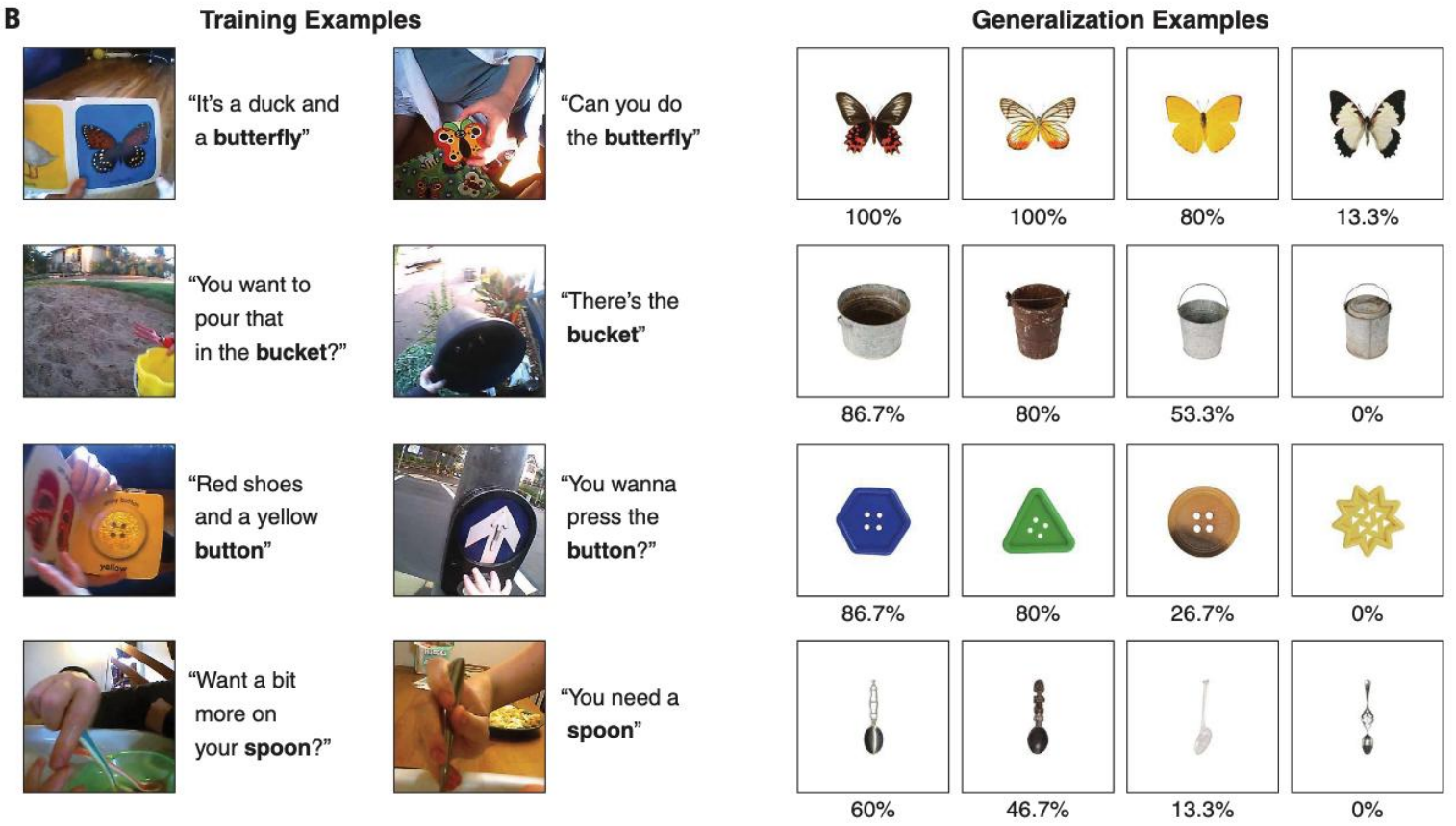
Fig. 1. CVCL model architecture and evaluation procedure. (**A**) Examples of whereas mismatching pairs are pushed apart. Example evaluation trials with

Vong, W. K., Wang, W., Orhan, A. E., and Lake, B. M (2024). Grounded language acquisition through the eyes and ears of a single child. Science, 383(6682), 504-511

Do humans have active learning behaviors?

基于有限的数据(跟训练多模态大模型所需要的数据相比），CVCL 模型真的学会了大量单词和概念，且能完成zero-shot任务。



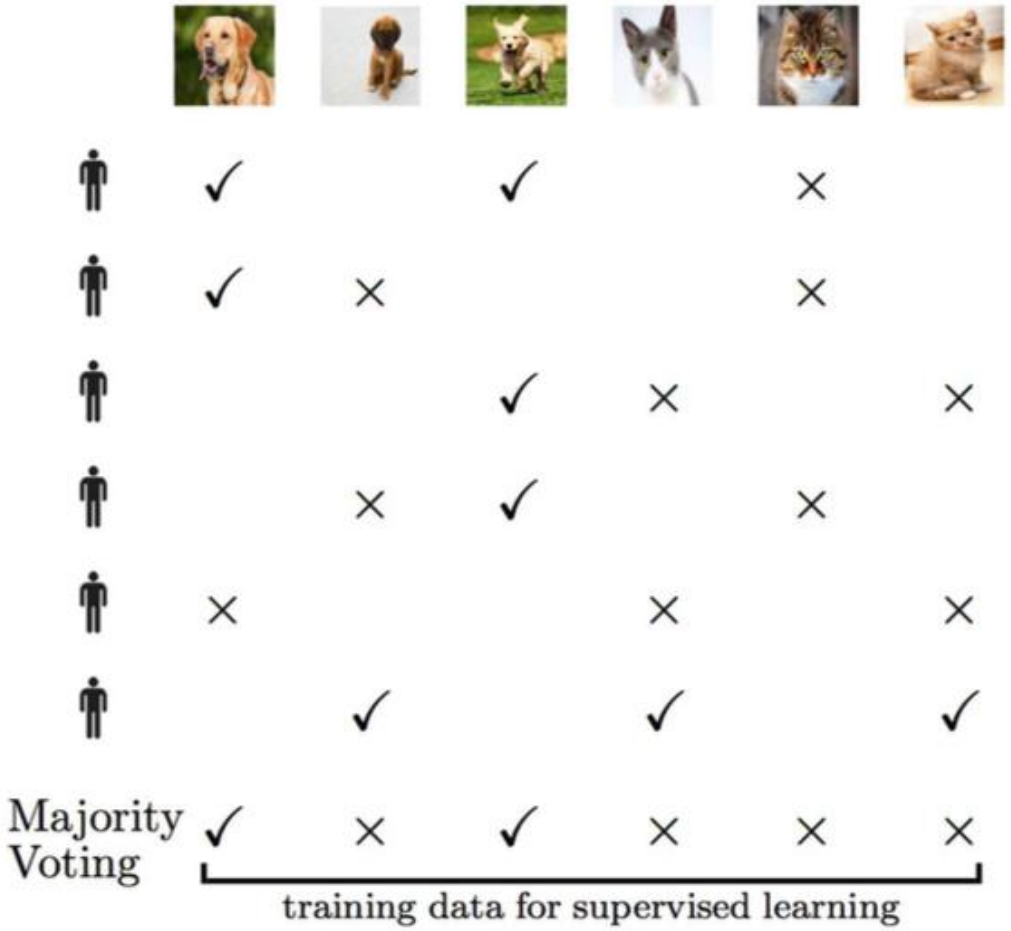


Vong, W. K., Wang, W., Orhan, A. E., and Lake, B. M (2024). Grounded language acquisition through the eyes and ears of a single child. Science, 383(6682), 504-511

Data aggregation

Crowdsourcing: aggregation of crowd annotations

- **Majority rule**
  - Simple and common
  - Wasteful: ignores annotator quality of different workers.
- **Annotator-quality models**
  - Can improve accuracy.
  - Hard: needs to be estimated without ground-truth.



https://www.mturk.com/

Amazon Mechanical Turk (MTurk) is a **crowdsourcing** marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually.

This could include anything from conducting simple data validation and research to **more subjective tasks** like survey participation, content moderation, and more.

MTurk enables companies to harness the collective intelligence, skills, and insights from **a global workforce** to streamline business processes, augment data collection and analysis, and accelerate machine learning development.

While technology continues to improve, there are still many things that human beings can do much more effectively than computers, such as moderating content, performing data deduplication, or research.

Traditionally, tasks like this have been accomplished by hiring a large temporary workforce, which is time consuming, expensive and difficult to scale, or have gone undone.

Crowdsourcing is a good way to break down a manual, time-consuming project into smaller, more manageable tasks to be completed by distributed workers over the Internet (also known as 'microtasks').

PROPOSED CROWDSOURCING ALGORITHM

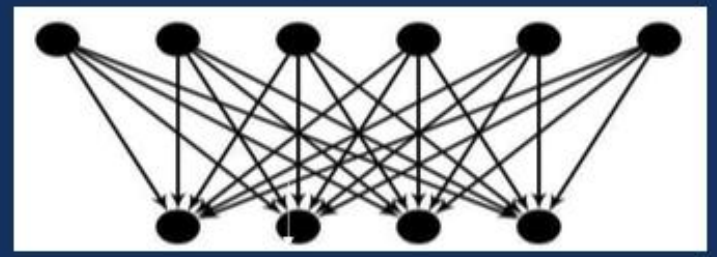
Noisy crowdsourced annotations

Repeat

Posterior of ground-truth labels given annotator quality model

cat	1/3	1/3	1/3	2/3	2/3	2/3
not cat	2/3	2/3	2/3	1/3	1/3	1/3

Training with weighted loss. Use posterior as weights



Use trained model to infer ground-truth labels

cat	0	0	1	1	1	0
not cat	1	1	0	0	0	1

MLE : update Annotator quality using inferred labels from model

0.7	0.3	0.5

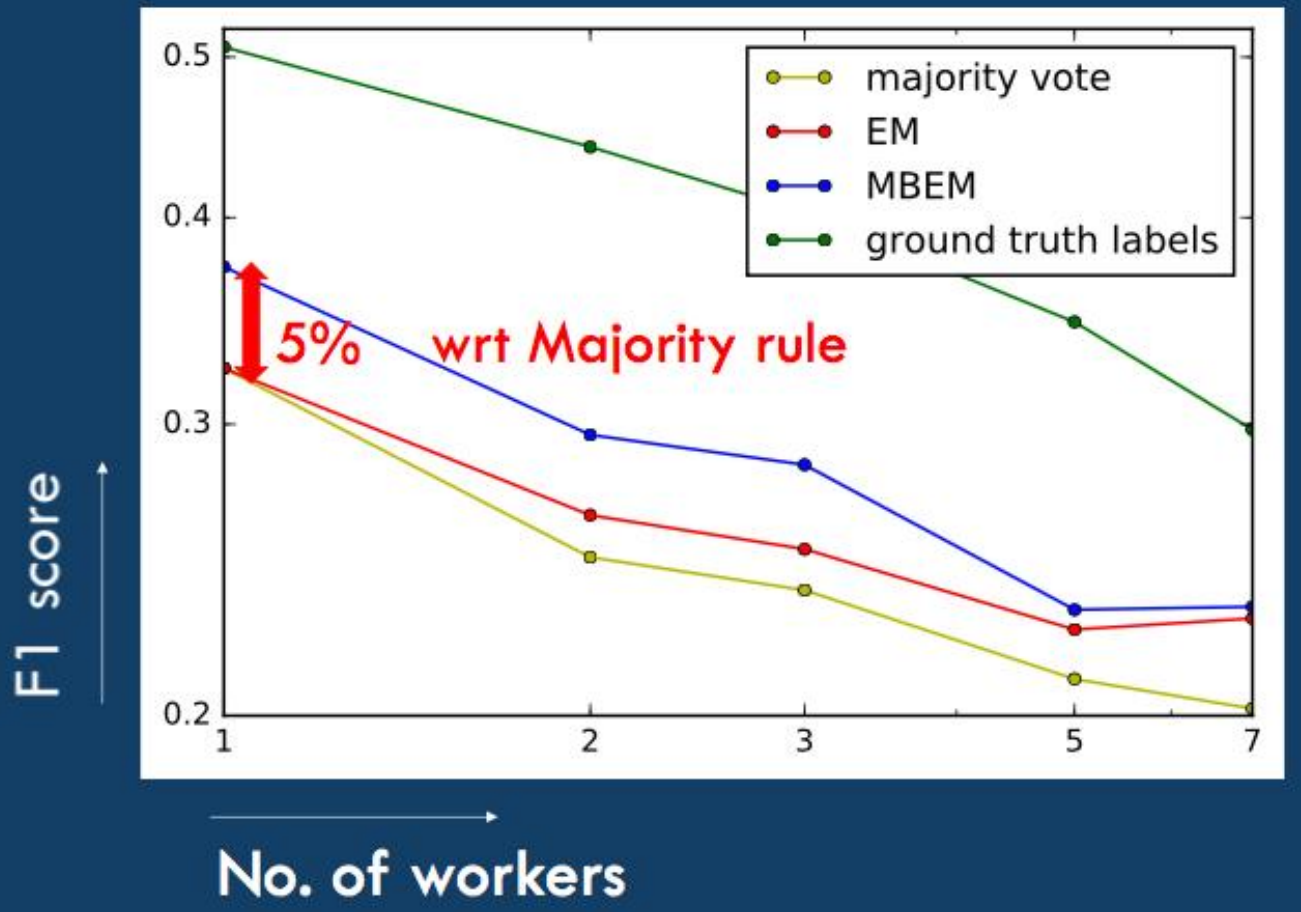
LABELING ONCE IS OPTIMAL: BOTH IN THEORY AND PRACTICE

Theorem: Under fixed budget, generalization error minimized with **single annotation** per sample.

Assumptions:

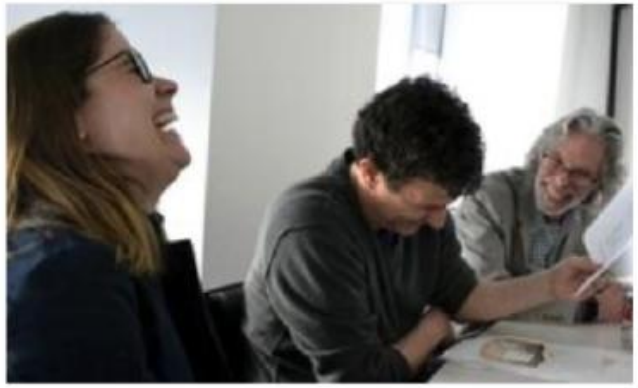
- Best predictor is accurate enough (under no label noise).
- Simplified case: All workers have same quality.
- Prob. of being correct > 83%

MS-COCO dataset. Fixed budget: 35k annotations



Active learning to optimize crowdsourcing and rating in New Yorker Cartoon Caption Contest

digg



BY DOING THE EXACT OPPOSITE

How New Yorker Cartoons Could Teach Computers To Be Funny

3 diggs CNET Technology

With the help of computer scientists from the University of Wisconsin at Madison, The New Yorker for the first time is using crowdsourcing algorithms to uncover the best captions.

3

Active learning
optimizes
Crowdsourcing

Federated learning: aggregation of multi-site data

Medical data pooling

- A lack of generalizability and accuracy for models
- Concerns regarding the reproducibility of results

- Privacy issues
- Domain shift (fig1)

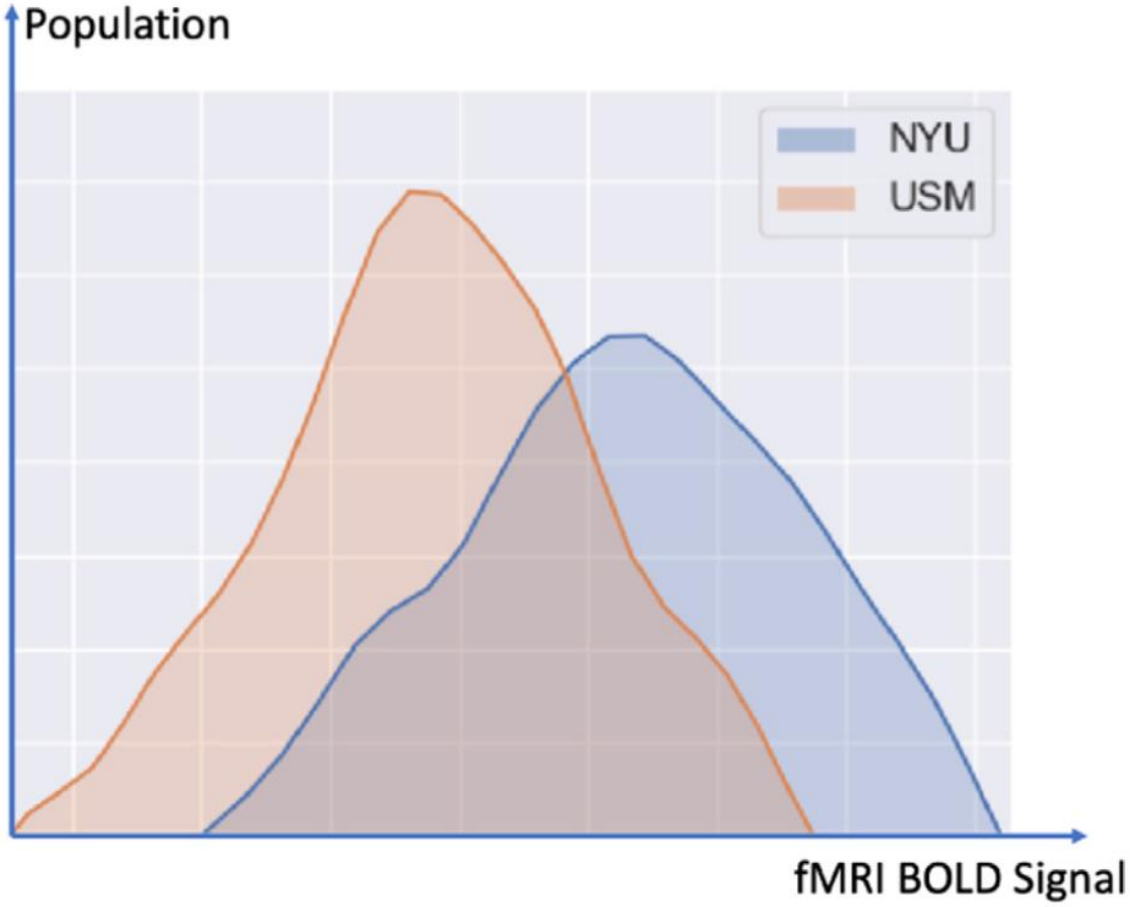


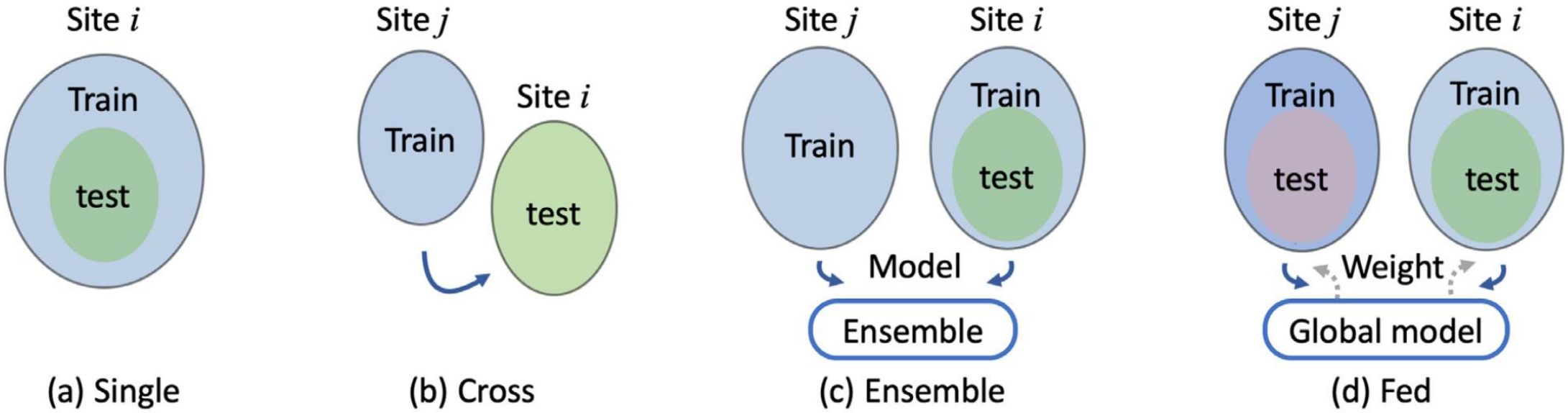
Fig. 1. fMRI distribution of different sites.

Li et al (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results , MedIA

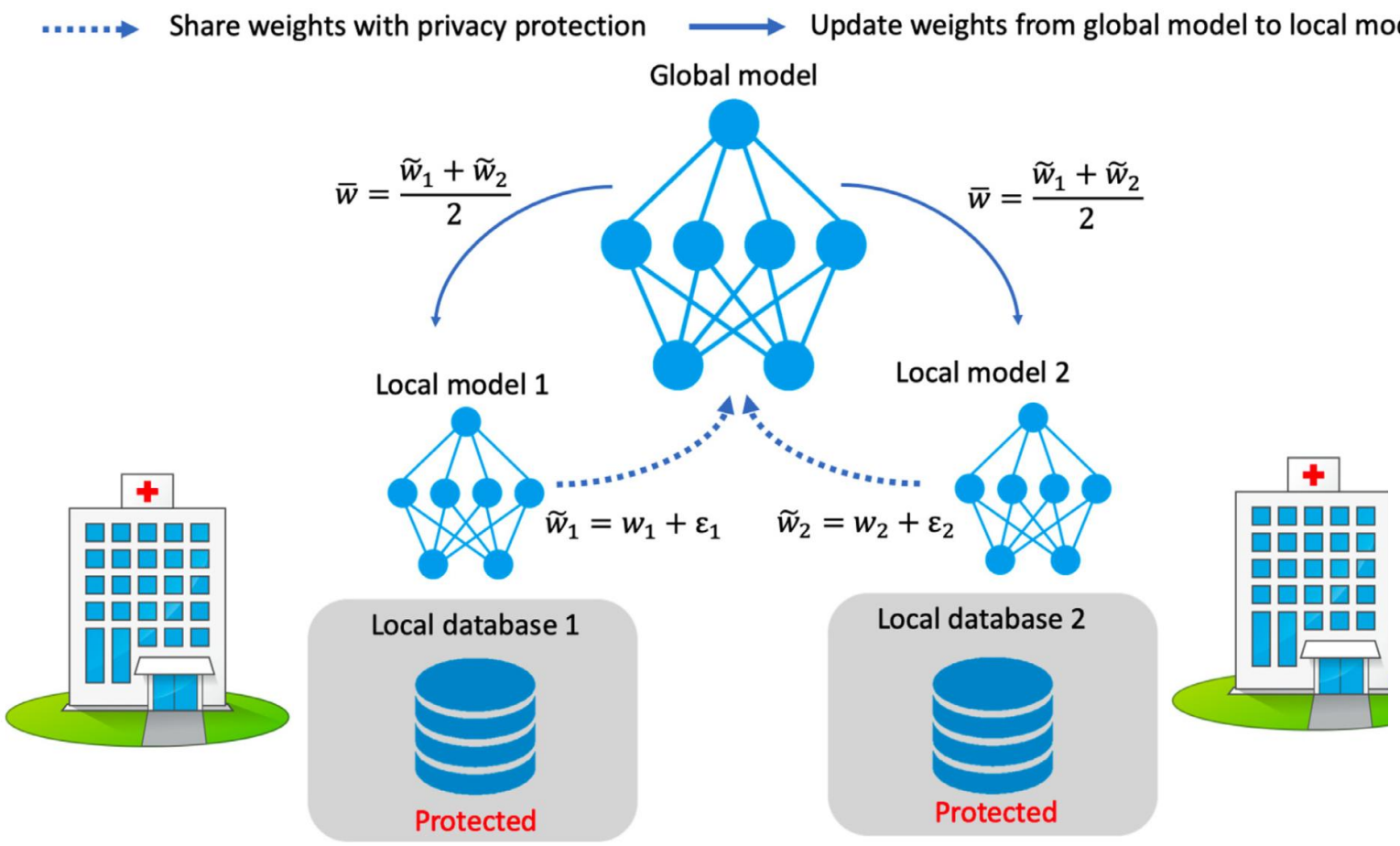
Federated learning can be achieved by two approaches:

1) each party training the model using private data, and where only **model parameters** being transferred (data is not shared)

2) using **encryption** techniques to allow safe communications between different parties.



Li et al (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results , MedIA



The simplified example of privacy-preserving federated learning strategy for fMRI analysis.

Li et al (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results , MIA

Algorithm 1 Privacy-preserving federated learning for multi-site fMRI analysis.

Input: 1. $\mathbf{X} = \{X_1, \ldots, X_N\}$, fMRI data from N institutions/sites; 2. $\mathbf{f}_w = \{f_{w_1}, \ldots, f_{w_N}\}$, local models within N sites, where w_i is local model weights; 3. $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$, fMRI labels; 4. $M(\cdot)$, noise generator that is used for privacy-preservation (explained in the following section); 5. K, number of optimization iterations; 6. τ, global model updating pace, which means the global model and the private models communicate per τ steps in each optimization iteration;7. $\{opt_1(\cdot), \ldots, opt_N(\cdot)\}$, optimizer returning updated model weights w.r.t. objective function $\mathcal{L}$.

$\{w_1^{(0)}, \ldots, w_N^{(0)}\} \leftarrow$ randomize parameters ▷ initialize local model
for $k = 1$ to K **do**
  $t \leftarrow 0$ ▷ initialize pace counter
  for $n = 1$ to N **do**
    $w_n^{(k)} \leftarrow opt_n(\mathcal{L}(f_{w_n^{(k-1)}}(X_n, Y_n))$
  end for
  $t \leftarrow t + 1$ ▷ models communicate
  if $t\%\tau = 0$ **then**
    $\bar{w}^{(k)} \leftarrow \frac{1}{N}\sum_n (w_n^{(k)} + M(w_n^{(k)}))$ ▷ update global model per τ steps
    for $n = 1$ to N **do**
      $w_n^{(k)} \leftarrow \bar{w}^{(k)}$ ▷ deploy weights to local model
    end for
  end if
end for
Return: global model $f_{\bar{w}^{(K)}}$

Autism spectrum disorders (ASD)

Table 1
Data summary of the dataset used in our study.

	NYU	UM	USM	UCLA
Total Subject	167	88	52	63
ASD Subject	73	43	33	37
HC Subject	94	45	19	26
ASD Percentage	44%	49%	63%	59%
fMRI Frames	176	296	236	116
Overlapping Trunc	145	265	205	85

$$\mathcal{L}_{ce}^{n} = -\sum_{n_i}[y_{n_i}\log(p_{n_i}) + (1-y_{n_i})\log(1-p_{n_i})]$$

Table 2
Data phenotype summary.

	SITE	AGE	ADOS	IQ	SEX
ASD	UM	12.4(2.2)	–	102.8(18.8)	M 36 F 7
	USM	22.9(7.3)	12.6(3.0)	99.8(16.4)	M 33 F 0
	NYU	14.7(7.1)	11.5(4.1)	107.4(16.5)	M 65 F 8
	UCLA	13.0(2.7)	10.4(3.6)	103.5(13.5)	M 31 F 6
HC	UM	14.1(3.4)	–	106.7(9.6)	M 32 F 13
	USM	20.8(8.2)	–	117.1(14.4)	M 19 F 0
	NYU	15.2(5.9)	–	112.6(13.5)	M 69 F 25
	UCLA	13.4(2.3)	–	104.9(10.4)	M 22 F 4

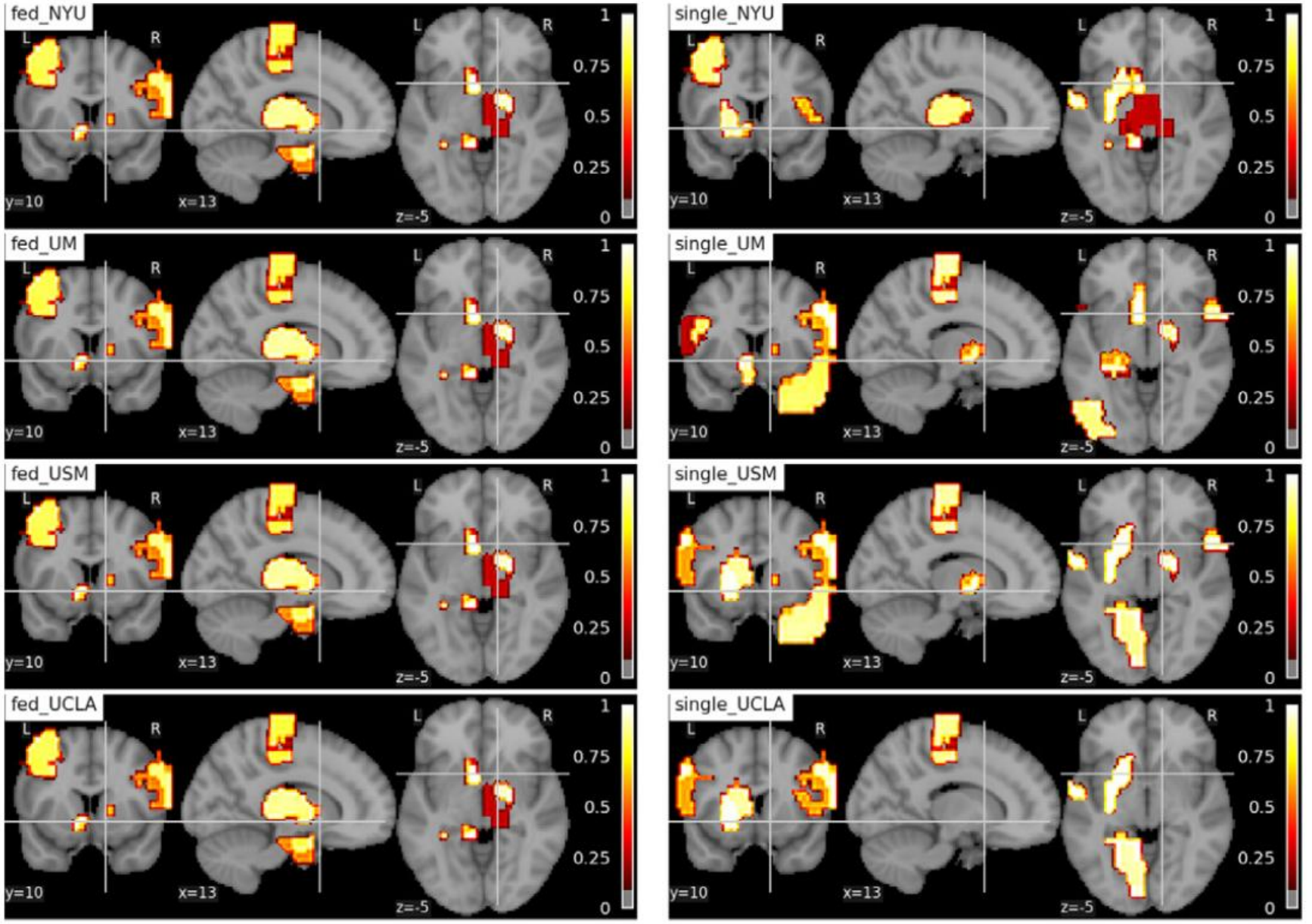
Values reported with mean (std) format. M: Male, F: Female, ADOS score: - means information not available

Results of using different training strategies.

	NYU	UM	USM	UCLA
trNYU	–	0.716	0.673	0.682
trUM	0.611	–	0.712	0.682
trUSM	0.641	0.625	–	0.730
trUCLA	0.575	0.648	0.750	–
Single	0.601(0.064)	0.648(0.065)	0.695(0.108)	0.571(0.100)
Ensemble	0.611(0.012)	0.638(0.054)	0.654(0.088)	0.634(0.064)
Fed	0.647(0.049)	0.728(0.073)	**0.849(0.124)**	0.712(0.075)
Fed-MoE	0.671(0.082)	0.728(0.083)	0.809(0.098)	**0.744(0.130)**
Fed-Align	**0.676(0.071)**	**0.751(0.053)**	0.829(0.091)	0.712(0.089)
Mix	0.671(0.035)	0.740(0.063)	0.829(0.137)	0.710(0.128)

Li et al (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results , MIA

Interpreting brain biomarkers associated with identifying **ASD** from federated learning model (*Fed*) and using single site data for training (*Single*).

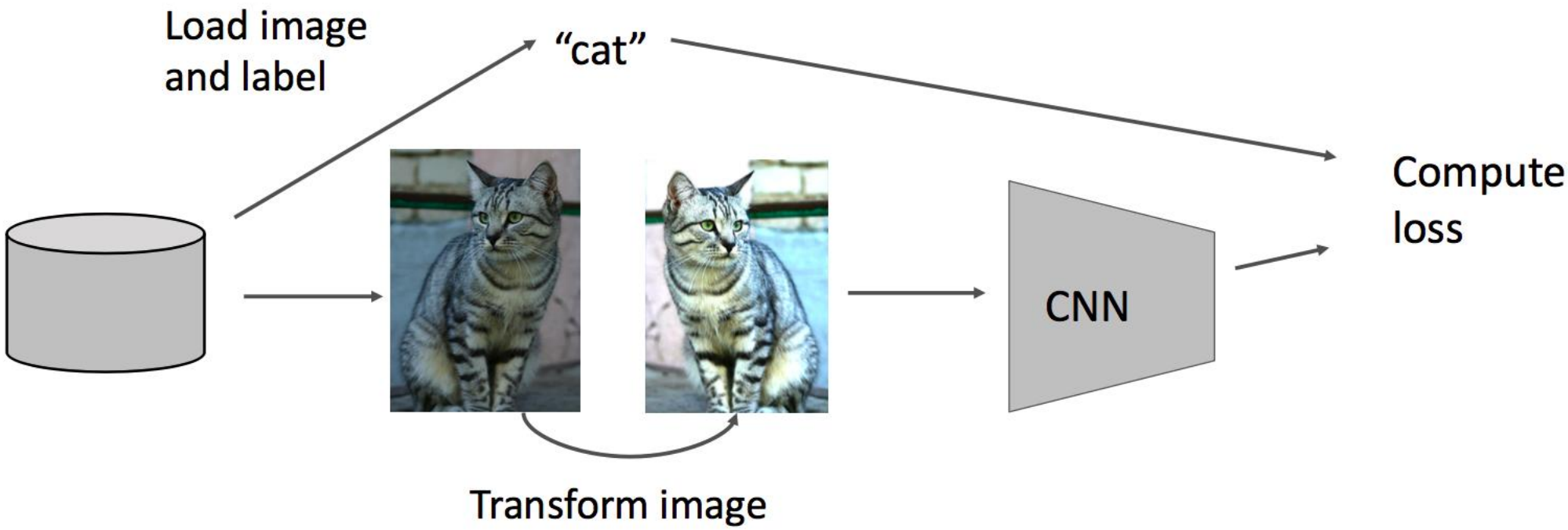


(a) Biomarkers using ***Fed*** strategy - view 1.

(b) Biomarkers using ***Single*** strategy - view 1.

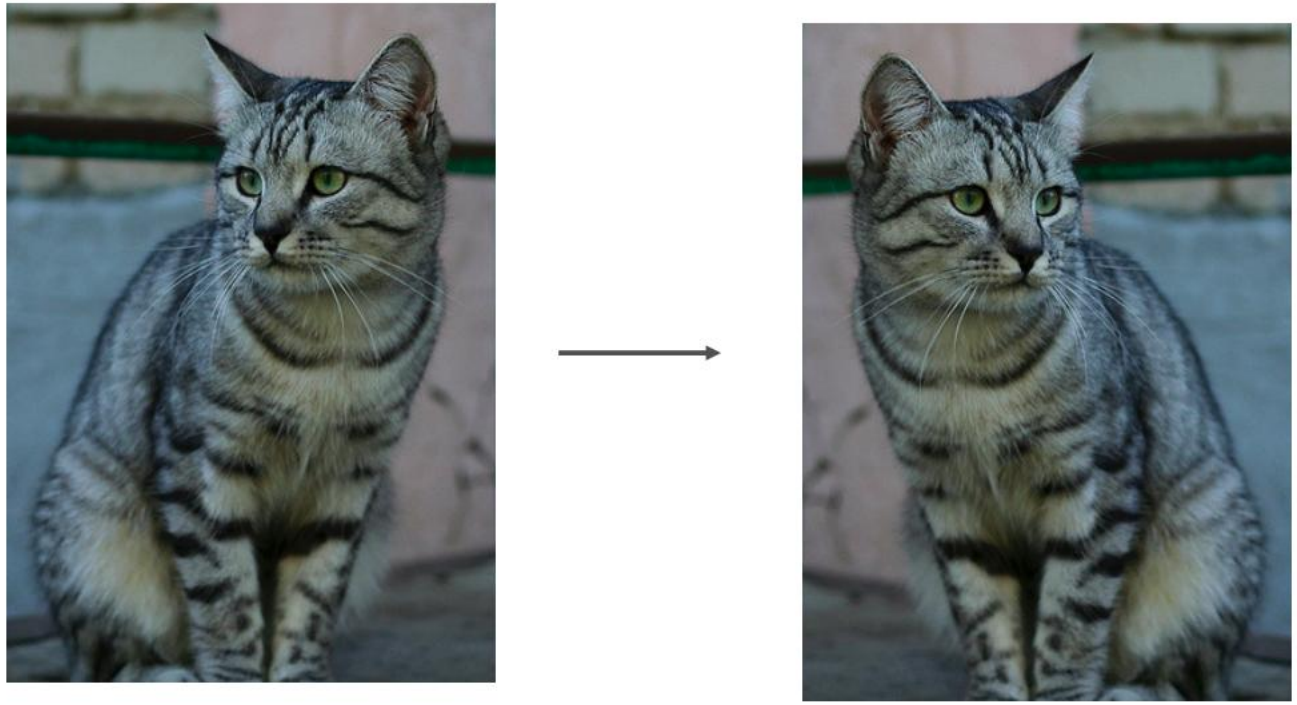
Data augmentation/generation

Data Augmentation



It multiplies the labeled data for free!

Data Augmentation: Horizontal Flips



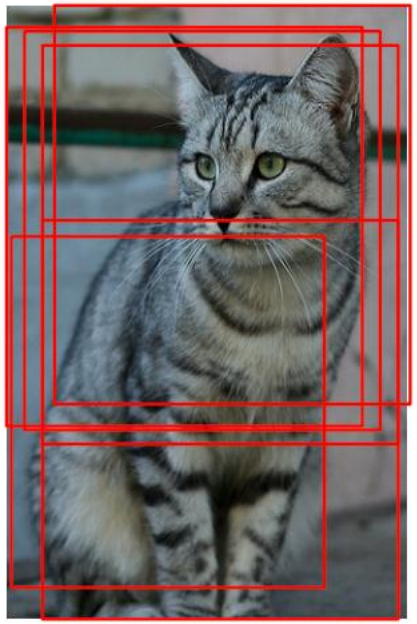
Data Augmentation: Random Crops and Scales

Training: sample random crops / scales

ResNet: at every iteration

1. Pick random L in range [256, 480]
2. Resize training image, short side = L
3. Sample random 224x224 patch

Add randomness



Marginalize over randomness

Testing: average a fixed set of crops

ResNet:

1. Resize image at 5 scales: {224, 256, 384, 480, 640}
2. For each size, use ten 224x224 crops, 4 corners + center, + flips on them
3. Average the prediction

Data Augmentation: Color Jitter

A simple way:

Randomize contrast and brightness

A more complex way:

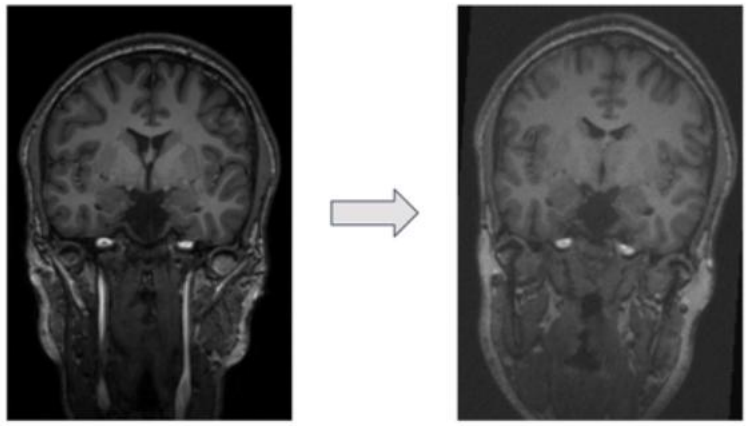
1. Apply **PCA** to all [R, G, B] pixels in training set
2. Sample a 'color offset' along PC directions
3. Add offset to all pixels of a training image

(Used in AlexNet, ResNet, etc)

Data Augmentation: Get creative for your problem…

Random mix/combination of

- Translation
- Rotation
- Stretching → cell images
- Shearing
- Lens distortions, … (go crazy)

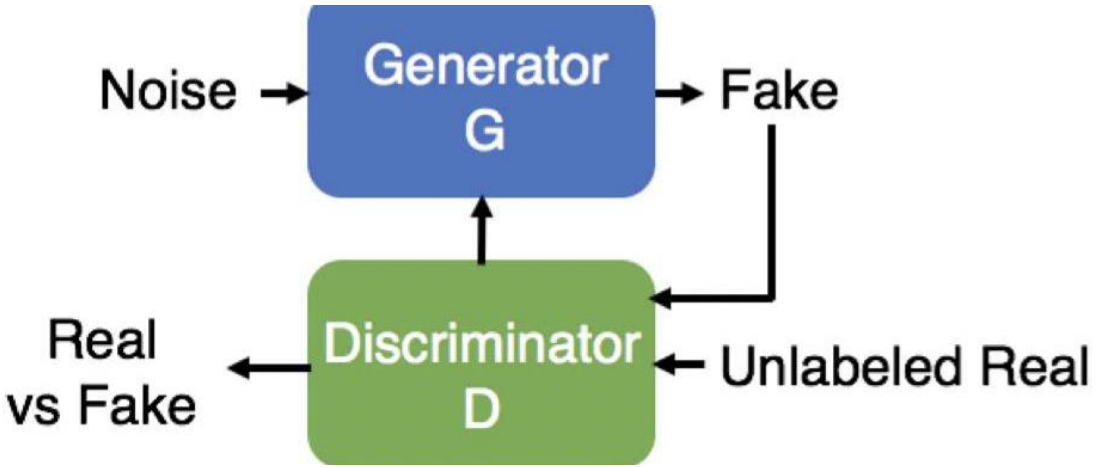


It requires some **human-expert knowledge** about what types of transformations do or do not change your label that you are trying to predict.

Data Augmentation: Generative models

Generative adversarial network (GAN)

More details about GAN in future weeks.



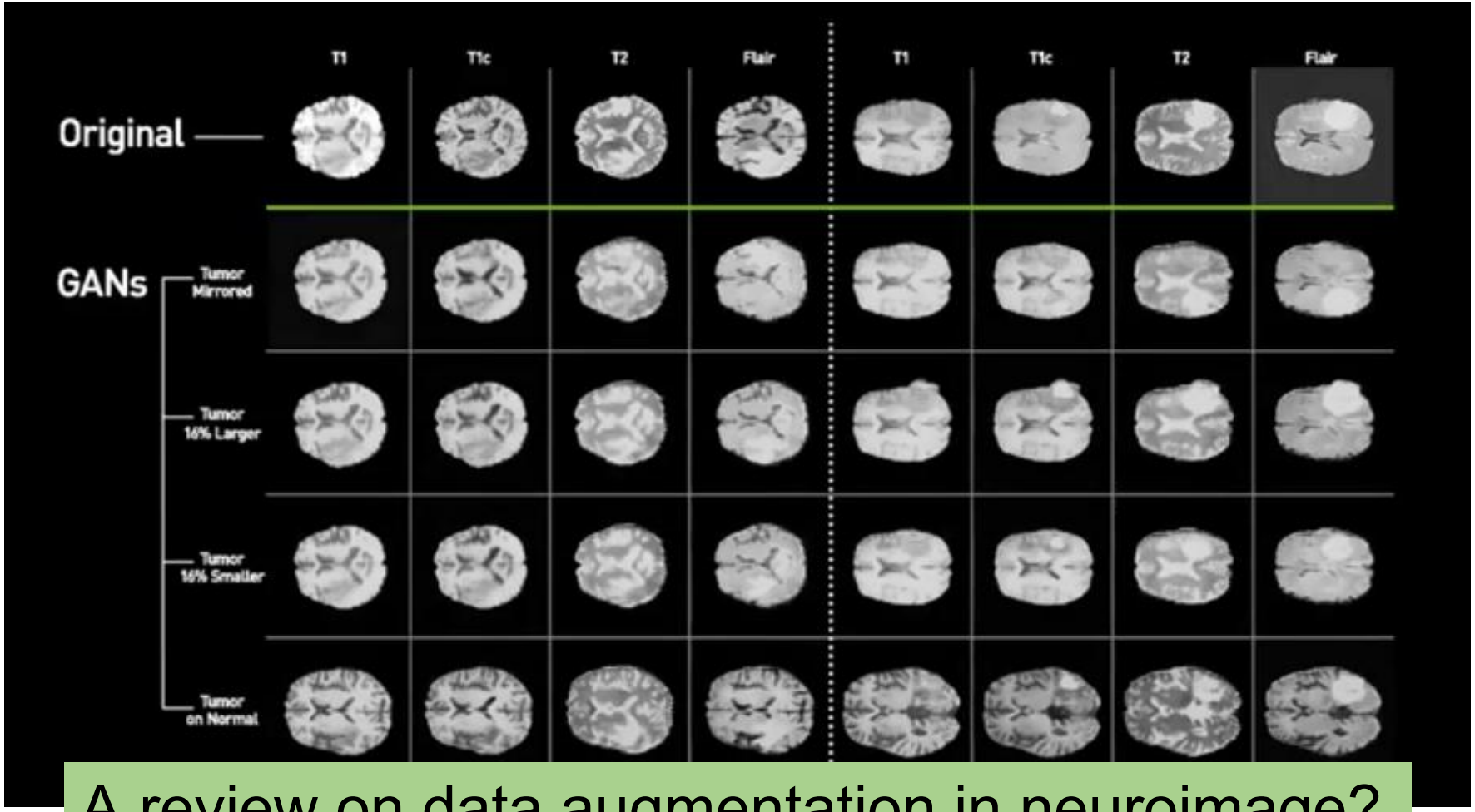
Merits

- Captures statistics of natural image
- Learnable

Peril

- Feedback is real vs. fake: different from prediction
- Introduces artifacts

Data Generation: GAN for generating biomedical images



A review on data augmentation in neuroimage?

Medical Image Synthesis for Data Augmentation and Anonymiz https://www.bilibili.com/video/BV11E4

Medical images need Data Augmentation/generation, but we have to be careful…

"We performed extensive data augmentation"

"All we do in medical imaging is overfit to our own dataset"

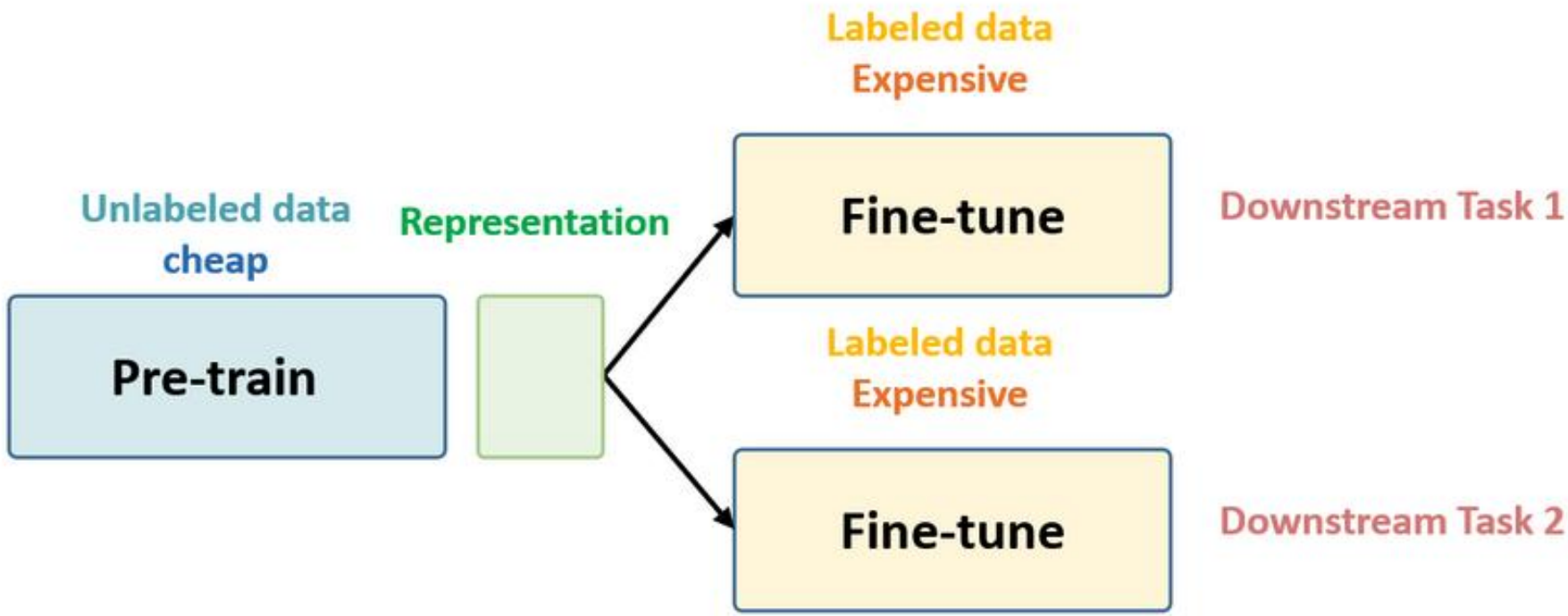
- Which software?
- Which transforms?
- Which parameters?
- Can I replicate them?
- How do they work? Documentation? Code?

Pre-trained models

Pre-trained models

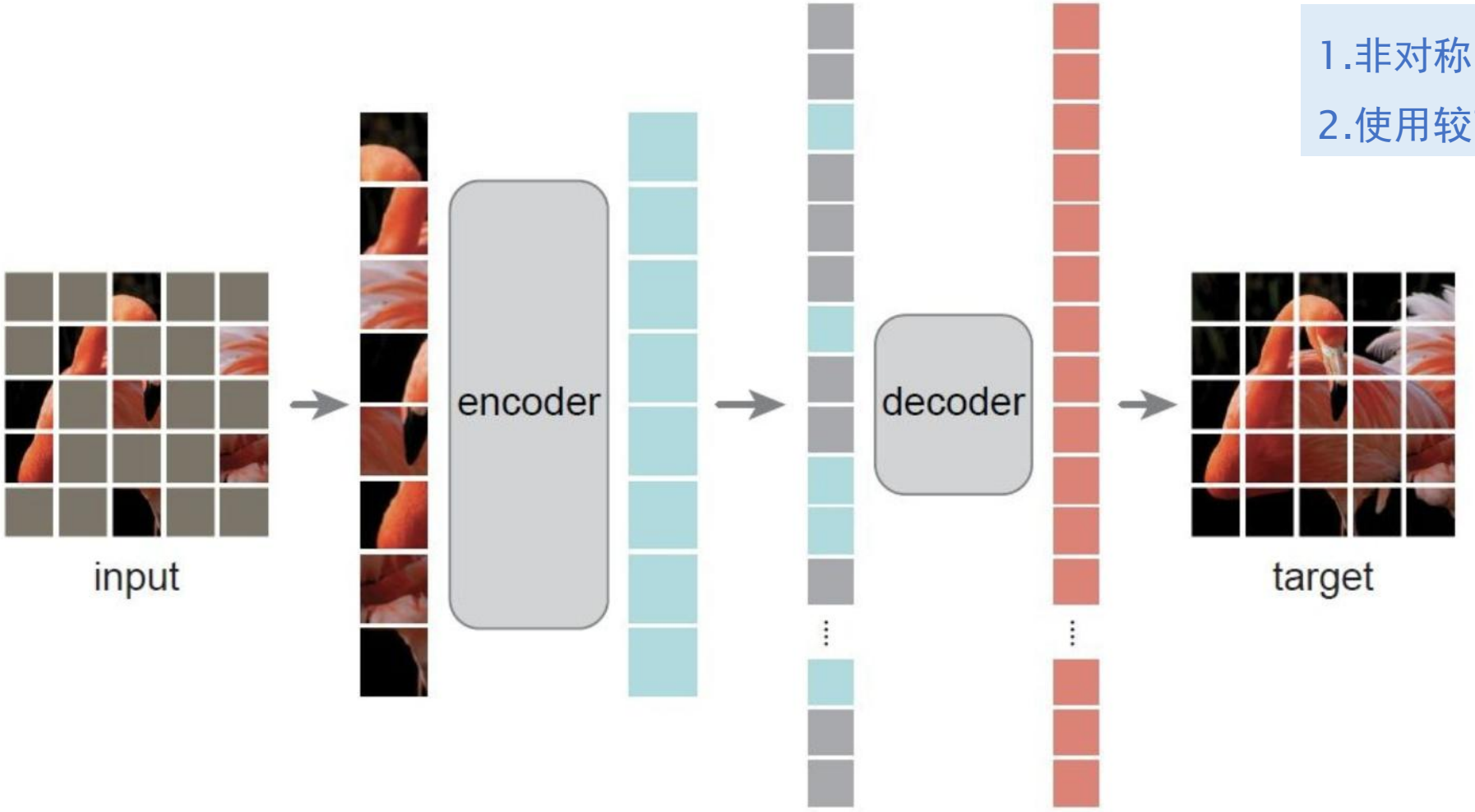
- *Pre-train a model with labeled data from other datasets and then transfer to our dataset*
  - supervised
  - eg., train with ImageNet; transfer to brain images
  - Models: CNN, RNN, …
- *Pre-train a model with a large amount of unlabeled data*
  - self-supervised
  - input masked data; reconstruct the full data
  - fine-tune for downstream tasks
  - Models: BERT; GPT series; Masked Autoencoders

What is pre-training in AI models?



Self-Supervised Learning 超详细解读 https://zhuanlan.zhihu.com/p/381354026

Masked autoencoder

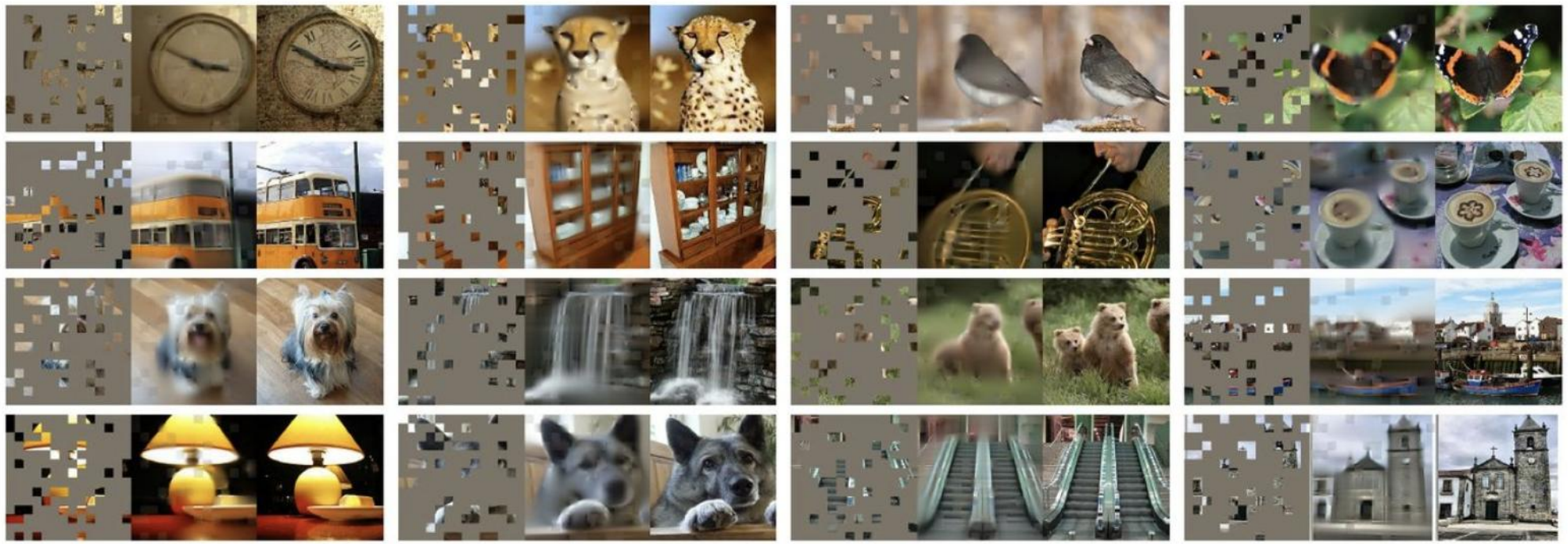


1.非对称的编、解码器设计

2.使用较高的掩码率(mask 比例)

Kaiming He, et al. (2021) Masked Autoencoders Are Scalable Vision Learners

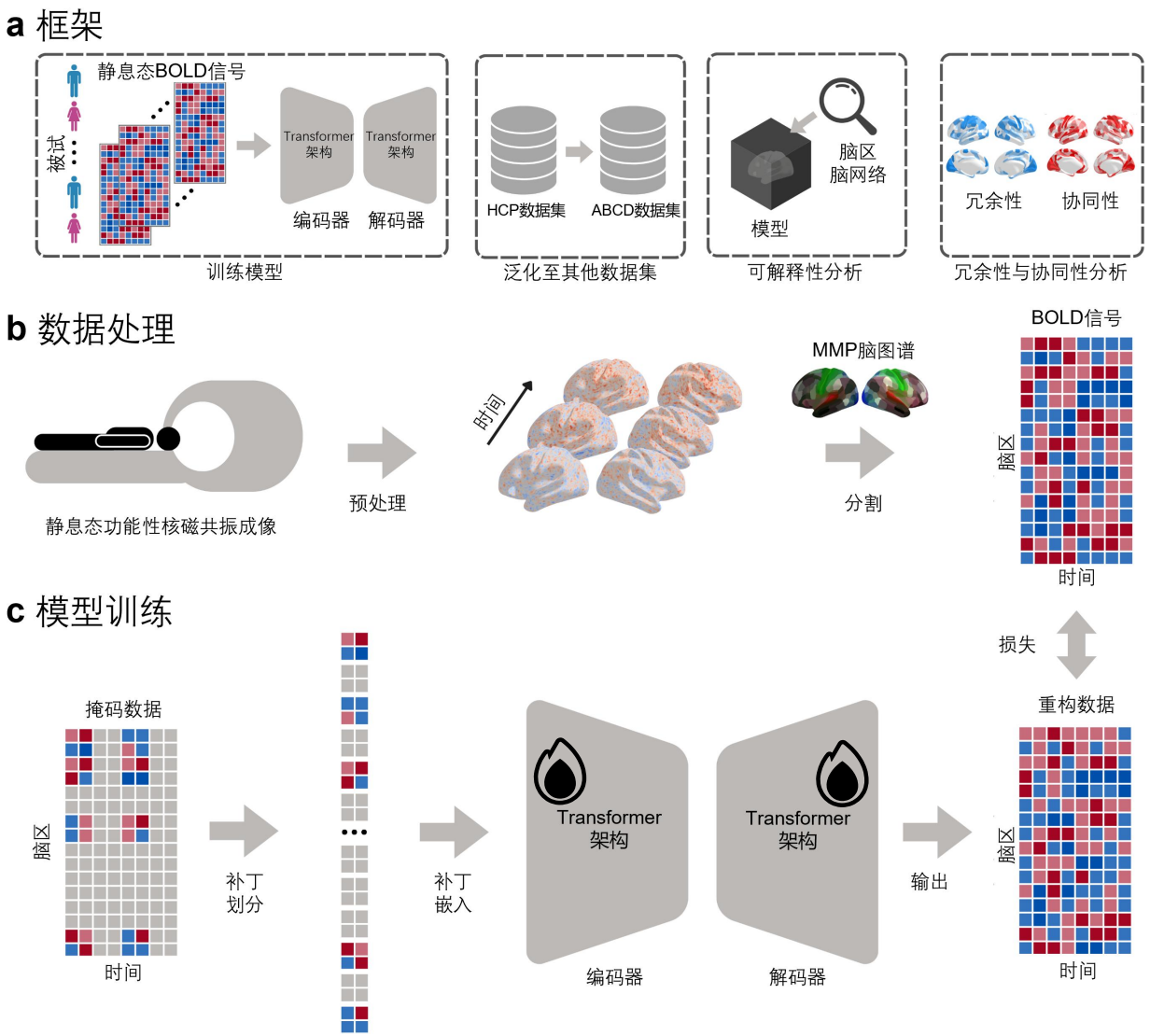
Masked autoencoder



train on Imagenet
test on Imagenet

train on Imagenet
test on COCO

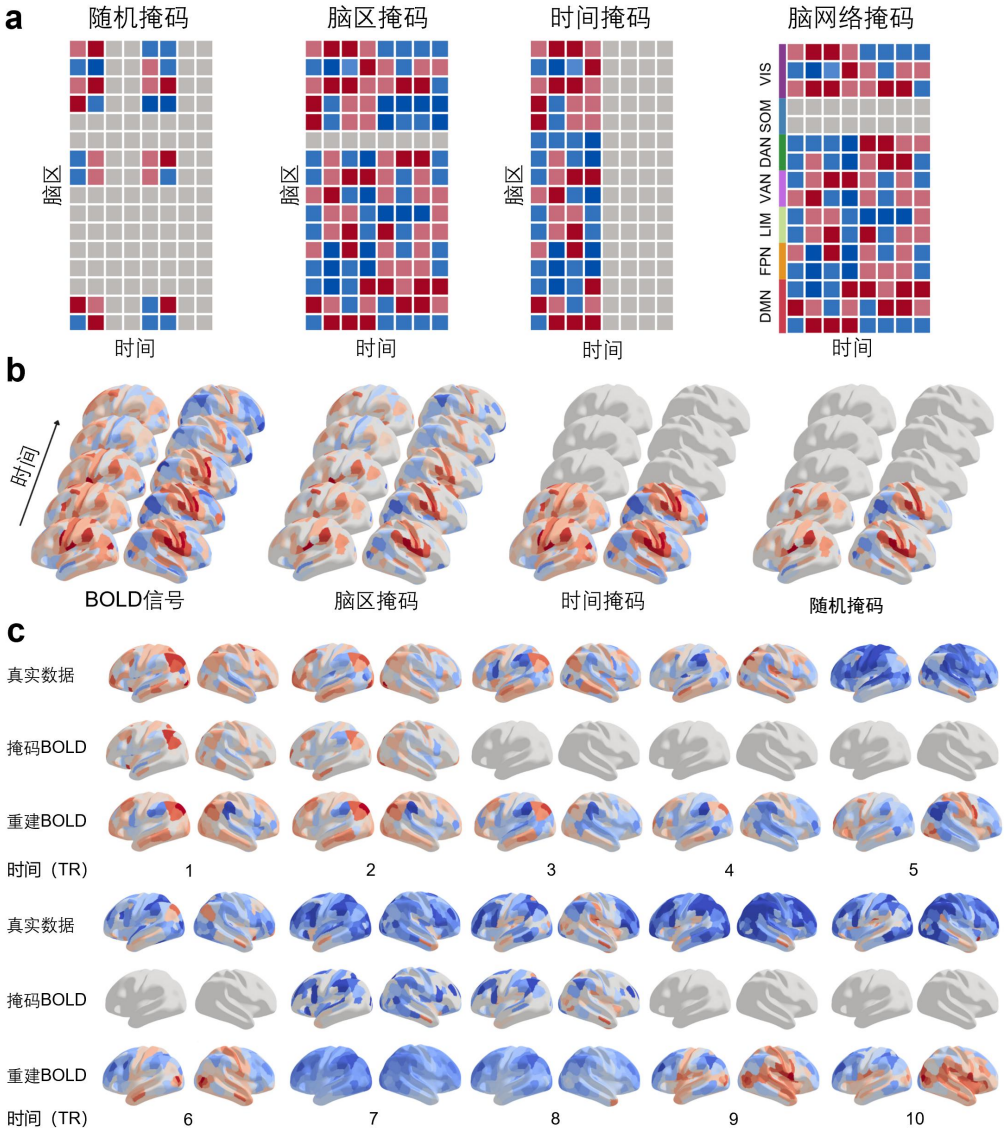
基于掩码自编码器模型的神经信号重构研究流程图



基于掩码自编码器模型的神经信号重构研究流程图

掩码方式及掩码自编码器模型重构结果可视化

不同的掩码方式



静息态神经信号重构的模型消融实验

(a) 掩码比例

掩码比例	相关性	MAE	MSE
0.25	0.576	0.228	0.092
0.5	**0.614**	**0.220**	**0.082**
0.75	0.593	0.221	**0.082**

(b) 补丁大小

补丁大小	相关性	MAE	MSE
2	**0.614**	**0.220**	**0.082**
5	0.527	0.241	0.098
10	0.438	0.256	0.112

(c) 编码器隐藏层维度

维度	相关性	MAE	MSE
512	0.586	0.228	0.087
768	0.589	0.227	0.088
1024	**0.614**	**0.220**	**0.082**

(d) 解码器隐藏层维度

维度	相关性	MAE	MSE
512	0.589	0.227	0.087
768	0.601	0.224	0.085
1024	**0.614**	**0.220**	**0.082**

测试了不同掩码比例（25%、 50%、 75%）、补丁大小（2、 5、 10）以及编码器和解码器的隐藏层维度（512、768、 1024），共计 81 种参数配置下的模型在神经信号重构任务中的表现 。根据以上测试结果，掩码自编码器模型后续测试与分析中均采用最佳的参数配置。

What is pre-training in the brain?

- Genetics
- Brain structure
- Sensory inputs (visual, auditory…)

The cognitive process

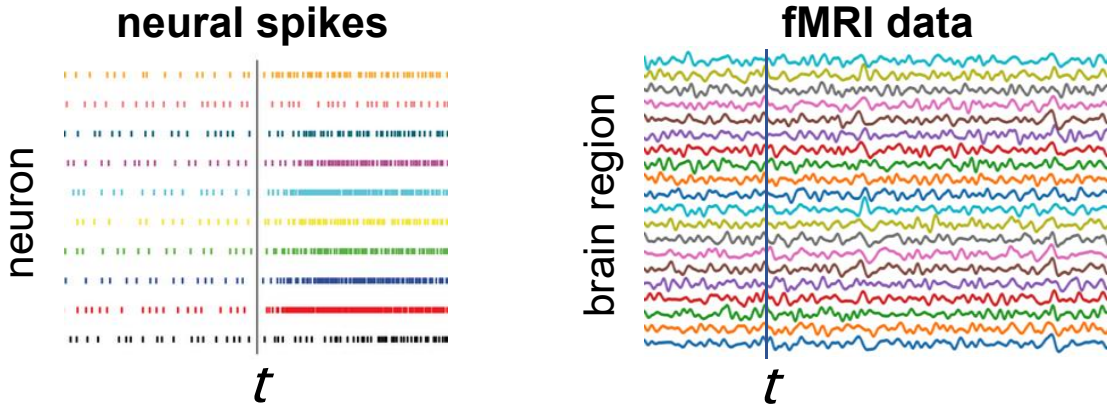
External world (image, text, sound …)

Neural data (10^{12} neurons, 10^{2} brain regions)

Cognitive process

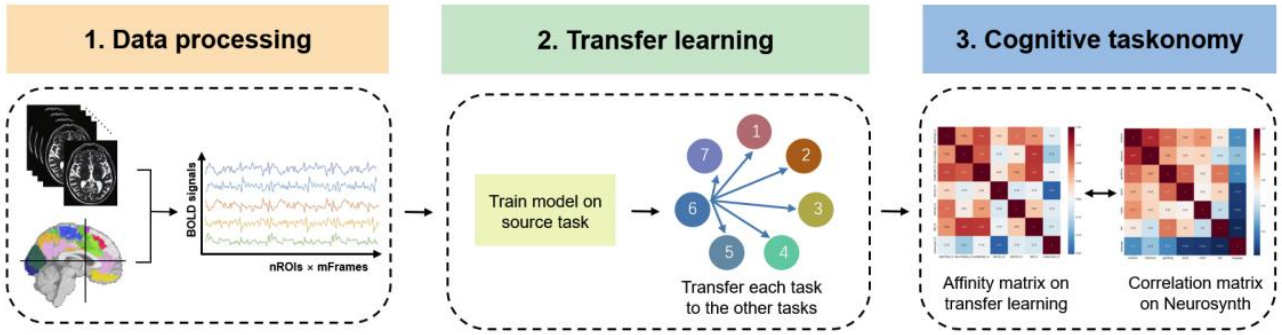
Senses (感觉)
Motion (运动)
Emotion (情绪)
Attention (注意力)
Cognition (认知)

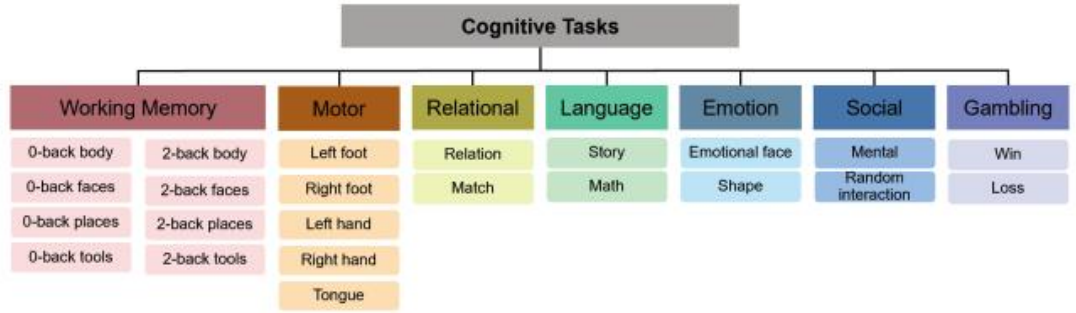
上善若水。水善利万物而不争，
处众人之所恶，故几于道。
居善地，心善渊，与善仁，
言善信，政善治，事善能，
动善时。夫唯不争故无尤。

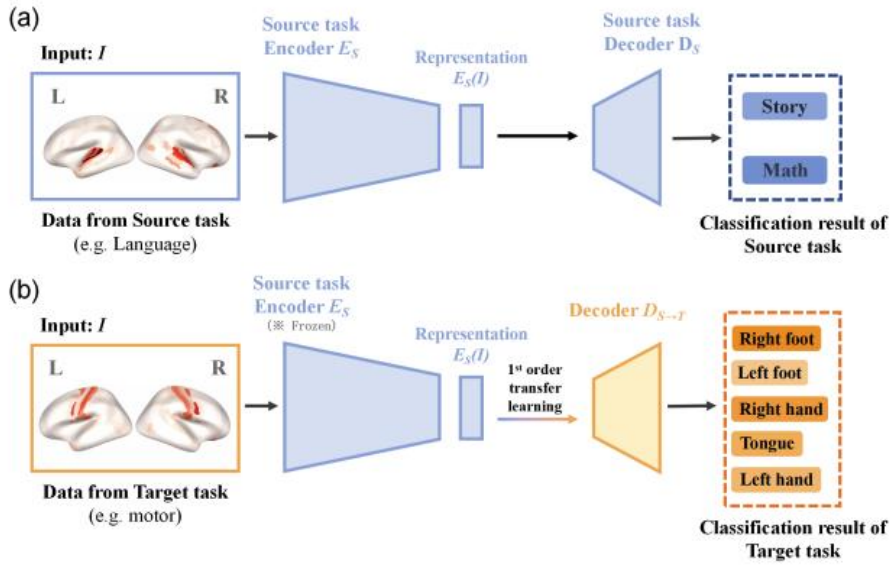


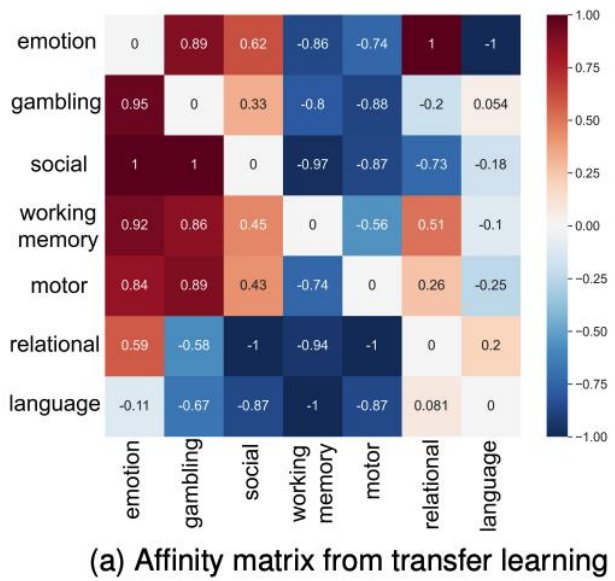
Heller & Casey (2016) *Developmental science*

NCC lab的项目：迁移学习来研究大脑的任务泛化

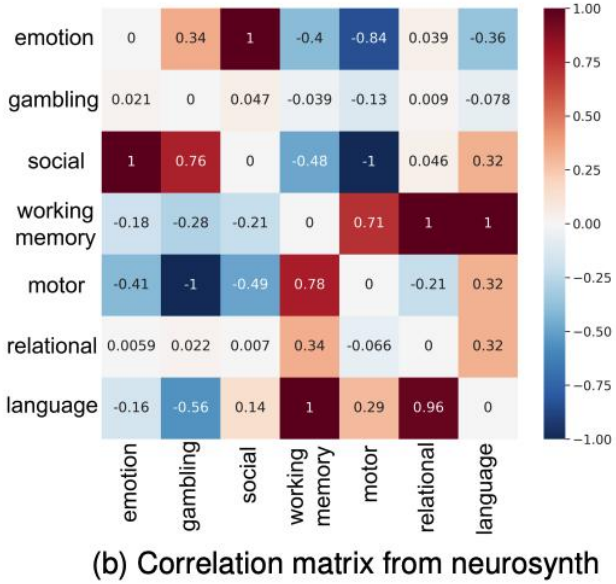








(a) Affinity matrix from transfer learning



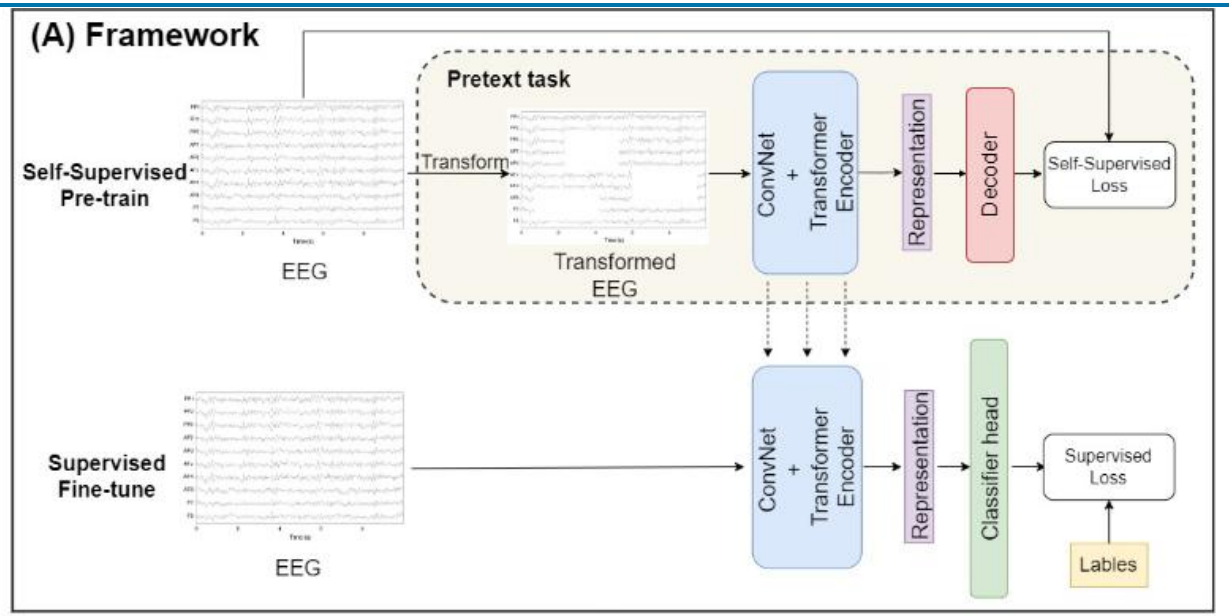
(b) Correlation matrix from neurosynth

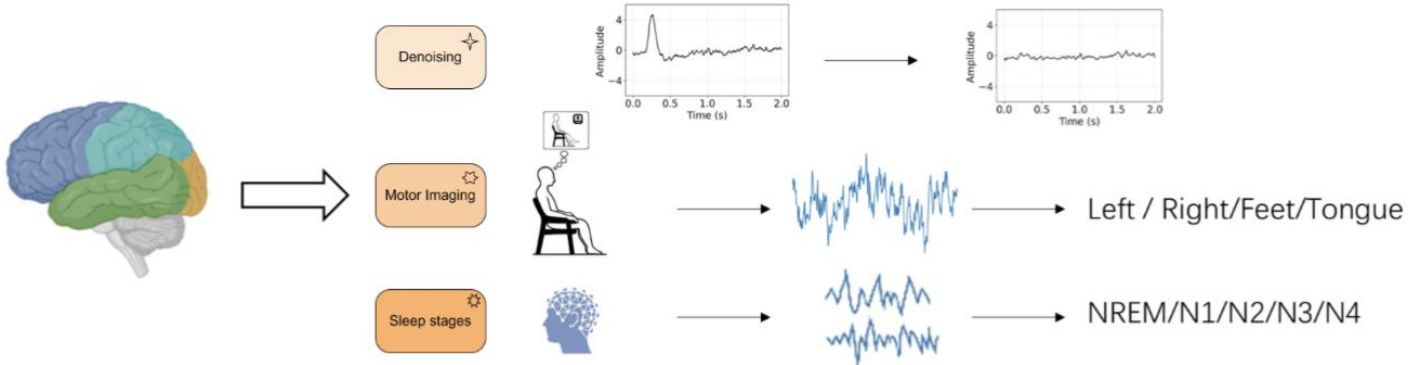
Qu et al (2022) Transfer learning to decode brain states reflecting the relationship between cognitive tasks, *IJCAI-HBAI workshop*

NCC lab的项目：Pre-training models for EEG

研究目的

- 构建EEG预训练模型，实现多下游任务，包括运动想象分类、疾病分类、EEG短期预测、降噪等
- 理解不同预训练任务对下游任务泛化带来的帮助
- 借助构建的预训练模型，通过embedding分析在不同任务下的EEG表征的性质
- 通过扰动分析，理解EEG的差异性体现在哪方面，对神经系统而言作用是什么

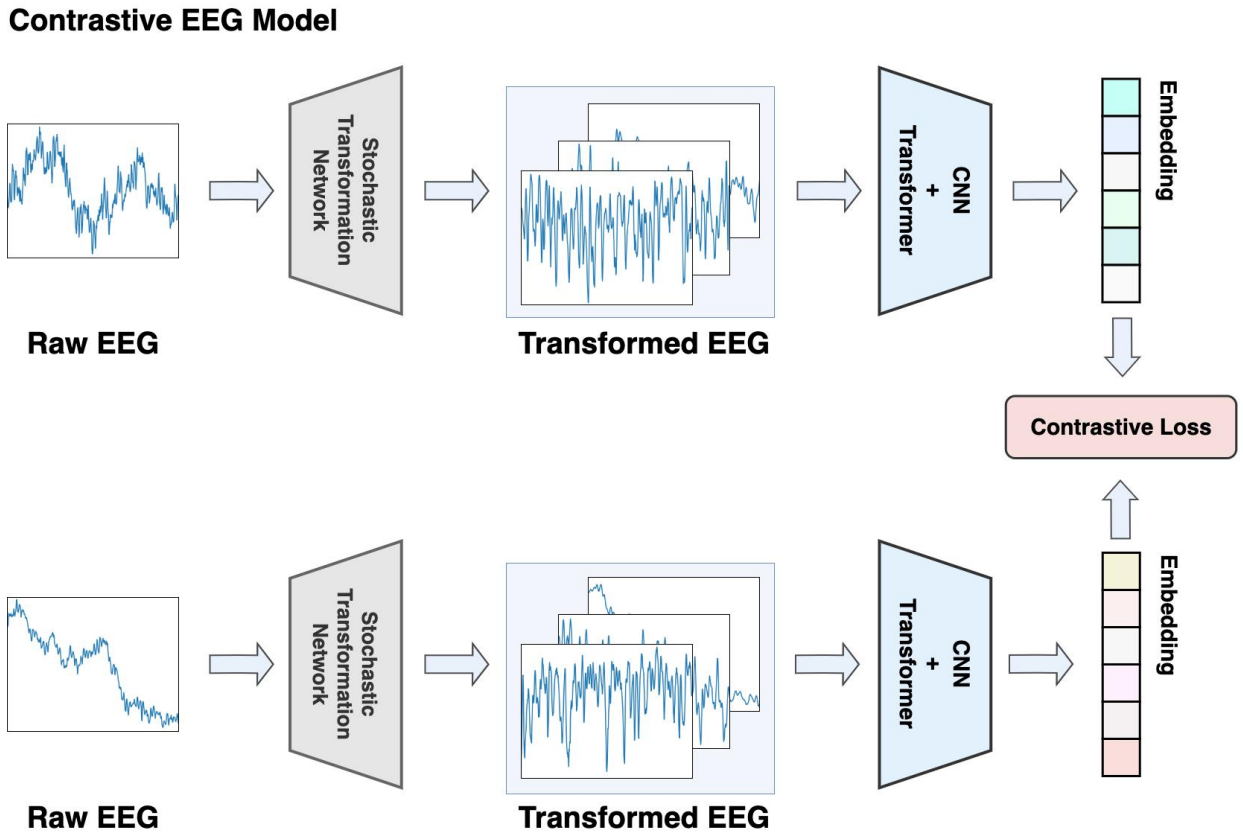


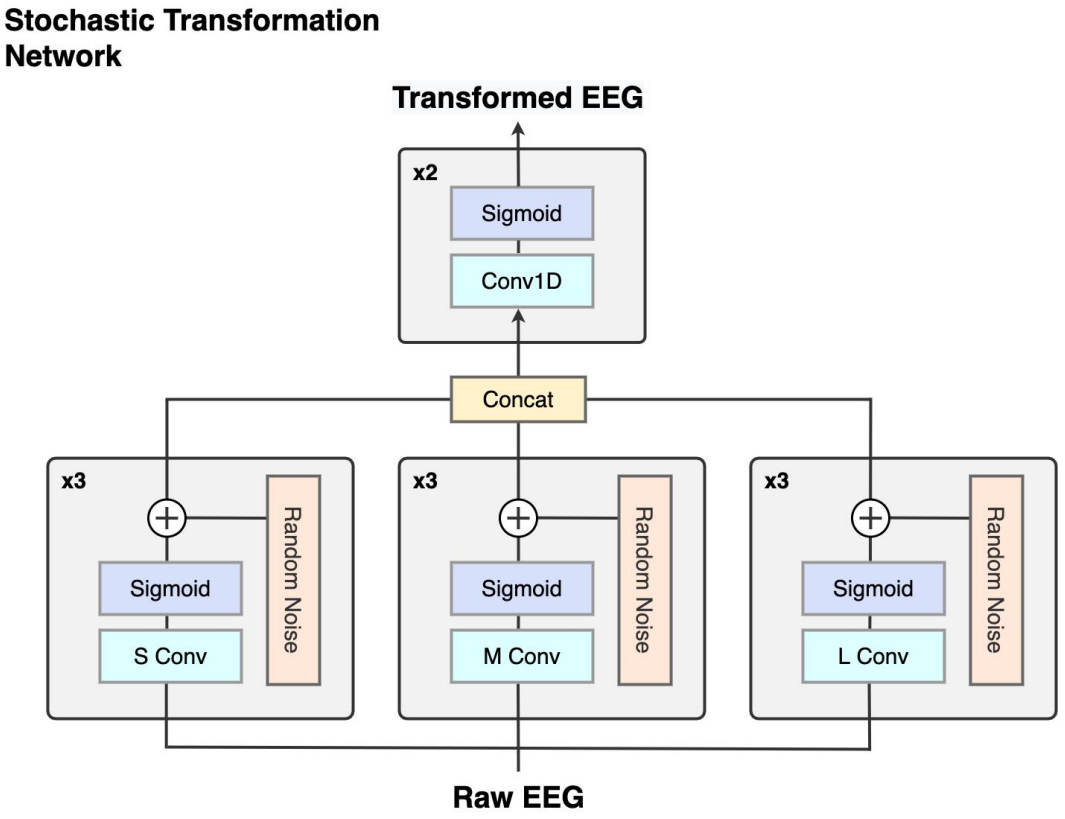


（张逸轩，余俊杰，2022）

脑电EEG预训练模型

- 借助对比学习，以减少被试内、跨被试差异性为目标，设计了一套新的范式，先训练一个神经网络学习如何最大化扰动数据作为数据增强方式，再进一步训练encoder最小化扰动带来的影响





（张逸轩，余俊杰，2022）

LaBraM: 脑电预训练大模型

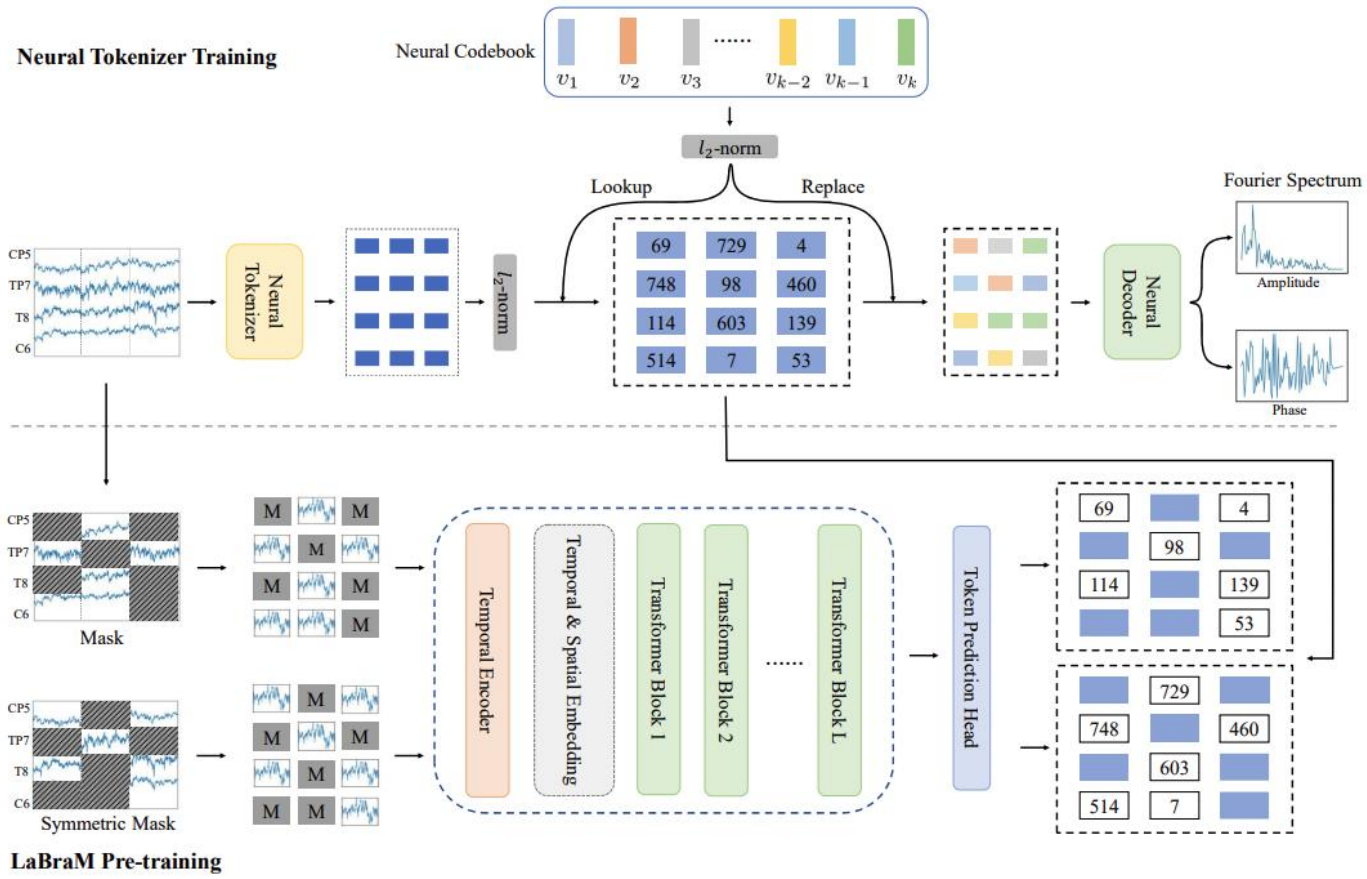


Figure 2: Overview of neural tokenizer training and LaBraM pre-training. **Up**: We train a neural tokenizer to discretize EEG signals into discrete neural tokens by reconstructing the Fourier spectrum. **Down**: During pre-training, part of EEG patches are masked while the objective is to predict masked tokens from visible patches.

Table 1: The results of different methods on TUAB.

Methods	Model Size	Balanced Accuracy	AUC-PR	AUROC
SPaRCNet (Jing et al., 2023)	0.79M	0.7896±0.0018	0.8414±0.0018	0.8676±0.0012
ContraWR (Yang et al., 2023b)	1.6M	0.7746±0.0041	0.8421±0.0104	0.8456±0.0074
CNN-Transformer (Peh et al., 2022)	3.2M	0.7777±0.0022	0.8433±0.0039	0.8461±0.0013
FFCL (Li et al., 2022)	2.4M	0.7848±0.0038	0.8448±0.0065	0.8569±0.0051
ST-Transformer (Song et al., 2021)	3.5M	0.7966±0.0023	0.8521±0.0026	0.8707±0.0019
BIOT (Yang et al., 2023a)	3.2M	0.7959±0.0057	0.8792±0.0023	0.8815±0.0043
LaBraM-Base	5.8M	0.8140±0.0019	0.8965±0.0016	0.9022±0.0009
LaBraM-Large	46M	0.8226±0.0015	0.9130±0.0005	0.9127±0.0005
LaBraM-Huge	369M	**0.8258**±0.0011	**0.9204**±0.0011	**0.9162**±0.0016

Table 2: The results of different methods on TUEV.

Methods	Model Size	Balanced Accuracy	Cohen's Kappa	Weighted F1
SPaRCNet (Jing et al., 2023)	0.79M	0.4161±0.0262	0.4233±0.0181	0.7024±0.0104
ContraWR (Yang et al., 2023b)	1.6M	0.4384±0.0349	0.3912±0.0237	0.6893±0.0136
CNN-Transformer (Peh et al., 2022)	3.2M	0.4087±0.0161	0.3815±0.0134	0.6854±0.0293
FFCL (Li et al., 2022)	2.4M	0.3979±0.0104	0.3732±0.0188	0.6783±0.0120
ST-Transformer (Song et al., 2021)	3.5M	0.3984±0.0228	0.3765±0.0306	0.6823±0.0190
BIOT (Yang et al., 2023a)	3.2M	0.5281±0.0225	0.5273±0.0249	0.7492±0.0082
LaBraM-Base	5.8M	0.6409±0.0065	0.6637±0.0093	0.8312±0.0052
LaBraM-Large	46M	0.6581±0.0156	0.6622±0.0136	0.8315±0.0040
LaBraM-Huge	369M	**0.6616**±0.0170	**0.6745**±0.0195	**0.8329**±0.0086

groups by 80% and 20%, respectively. We employ binary cross-entropy (BCE) loss for TUAB (binary classification) and cross-entropy loss for TUEV (multi-class classification), respectively. Our experiments are conducted on eight A800 GPUs by Python 3.11.4 and PyTorch 2.0.1 + CUDA 11.8. The best models are trained based on the training set, selected from the validation set, and finally evaluated on the test set. We report the average and standard deviation values on five different random seeds to obtain comparable results. (see Appendix C for more detailed hyperparameters)

Jiang, W. B., Zhao, L. M., & Lu, B. L. (2024). Large brain model for learning generic representations with tremendous EEG data in BCI, *ICLR*

Recommended videos

1. ICML 2019 tutorial for active learning
   https://www.bilibili.com/video/BV1UJ411Y7nC

2. *Li Fei-fei,* CS231n, lecture 13 – Generative models
   https://www.bilibili.com/video/BV1gW4y127cy?p=13

3. *宋飏* ICLR'21杰出论文奖*: 基于梯度估计的生成式模型*
   https://www.bilibili.com/video/BV1ui4y1A7nb

   Score-Based Generative Modeling through Stochastic Differential Equations

4. *Stefano Ermon*, CS236: Deep Generative Models
   视频 https://www.bilibili.com/video/BV1SJ411b7D8/
   课件 https://deepgenerativemodels.github.io/syllabus.html

Summary of Lecture 6 – Data for DL

- **Data collection and labeling**
  - Active Learning
  - Parallel labels
- **Data Aggregation**
  - Crowdsourcing models
  - Federated learning
- **Data Augmentation/Generation**
  - Some tricks: Horizontal Flips, Random Crops and Scales, Color jitter ….
  - Generative models
- **Pre-training in AI & BI**