Conv Shape:$o=\frac{I+2P-D(K-1)-1}{S}+1$, Receptive field: $R_k=R_{k-1}+Dia_k(K_t-1)\times\Pi_{i=1}^{k-1}S_i$, Logistic: $h_\theta(x)=\frac{1}{1+e^{-\theta^T x}}$

Image Segmentation Noise（噪声）**Partial Volume Effects**（部分容积效应）**Intensity Inhomogeneities**（强度不均匀）**Anisotropic Resolution**（各向异性分辨率）不同轴 spatial resolution differs **Imaging Artifacts**（成像伪影）Non-physiological elements or distortions in an image eg motion artifacts or metal artifacts, by equipment or movement. **Limited Contrast**（对比度受限）不同组织类似物理特性，类似 intensity **Morphological Variability**（形态学变异）形态各异

Segmentation Eval **Ground Truth**: Reference or standard against method can be compared,e.g. the optimal transformation,or true segmentation boundary. usually only available for: Synthetic or simulated phantoms(模体/幻影体),Physical phantom (如凝胶模体，仿造脑等器官结构)

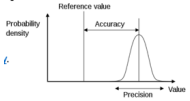
Gold Standard **Expert**: 人类观察着手工标注 ▪ **Disadvantage**: 需要训练&tedious & time-consuming

Intra-observer Variability: 同观测者不同结果, Inter-observer Variability: 不同观测者不同结果（Disagree）

▪ **Remedy**: 多次分割，多个专家分割，Quantify (dis)agree

Assess performance: **Precision/Positive Predictive Value:** $PPV:\frac{TP}{TP+FP}$, random errors, statistical variability, the repeatability, or reproducibility of the measurement

Accuracy$(ACC=\frac{TP+TN}{P+N}, P=TP+FN, N=TN+FP)$:



▪ 第一种说法: Acc = 系统误差的大小(Bias, trueness)

▪ 第二种说法: Acc = random+systematic -> high precision and high trueness.

简单点：高准确度 = 偏差小 + 随机波动小。$F_\beta=\frac{(1+\beta^2)PR}{\beta^2P+R}=\frac{(1+\beta^2)TP}{(1+\beta^2)TP+FP+\beta^2FN}$

Robustness: ▪ degradation in performance with respect to varying noise levels or other imaging artefacts

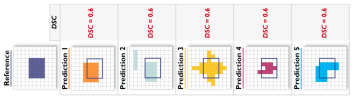
Confusion matrix **Conditional Positive/Neg** P/N: the number of real pos/neg cases in the data

▪**TP**/Hit ; **TN**/Correct Rejection ; **FP**/False Alarm/Type I E | **FN**/Miss/Type II E

▪**Recall/Sensitivity/Hit Rate/TP Rate**: $TPR=\frac{TP}{P}=\frac{TP}{TP+FN}$ ▪**Specificity/TN Rate**: $TNR=\frac{TN}{N}=\frac{TN}{TN+FP}$

▪**Precision/Pos Pred Val:** $PPV=\frac{TP}{TP+FP}$

Effect of structure shape



Overlap Measures: **Jaccard Index** (IoU): $JI=\frac{|A\cap B|}{|A\cup B|}$

Dice's Coefficient: $DSC=\frac{2|A\cap B|}{|A|+|B|}=\frac{2TP}{2TP+FP+FN}=F_1$

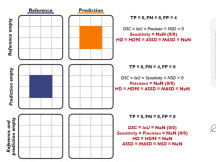
Other measures & Surface Dist. Measure

▪ **Volume similarity:** $VS=1-\frac{||A|-|B||}{|A|+|B|}=1-\frac{|FN-FP|}{2TP+FP+FN}$

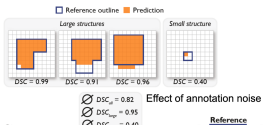
▪ **Hausdorff distance:** $HD=\max\left(h(A,B),h(B,A)\right), h(A,B)=\max_{a\in A}\min_{b\in B}||a-b||$

▪ **(Symmetric) Average surface distance:** $ASD=\frac{d(A,B)+d(B,A)}{2}, d(A,B)=\frac{1}{N}\sum_{a\in A}\min_{b\in B}||a-b||$

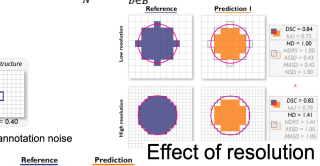
Pitfalls in Seg Eval Effect of structure size



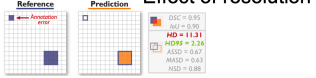
Effect of "empty" labelmaps



Effect of annotation noise



Effect of resolution

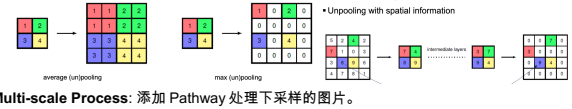


Segmentation Methods

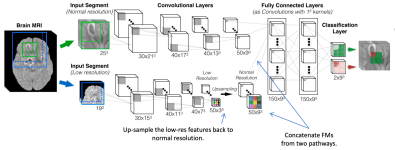
▪ **Intensity-based (e.g., thresholding τ)**: 选择 (UL: a lower and upper) τ. ✅: simple, fast, ❌:regions must be homogeneous & distinct -difficulty in finding consistent thresholds across images.-leakage, isolated pixels & 'rough' boundaries

▪ **Region-based (e.g., region growing):** 从用户选择的 seed ptr 生长区域 ✅ relatively fast, yields connected region (from a seed point) ❌: ▪ regions must be homogeneous, leakages and 'rough' boundaries likely, requires (user) input for seed points.

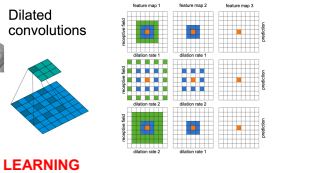
▪ **Atlas-based:** majority voting. Seg 对于 N 个标注好的 atlas，register 到新图片。然后 Fusion。✅ robust and accurate (like ensembles), yields plausible segmentations, fully automatic❌: computationally expensive, cannot deal well with abnormalities, not suitable for tumour segmentation.



Multi-scale Process: 添加 Pathway 处理下采样的图片。



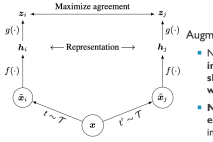
Dilated convolutions



SELF-SUPERVISED LEARNING (SSL) & CONTRASTIVE LEARNING

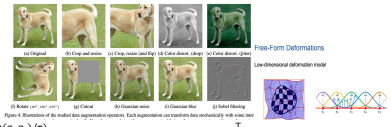
$ItoW:\begin{pmatrix}X\\Y\\1\end{pmatrix}=\begin{bmatrix}1&0&t_x\\0&1&t_y\\0&0&1\end{bmatrix}\begin{bmatrix}dxx&dyx&0\\dxy&dyy&0\\0&0&1\end{bmatrix}\begin{bmatrix}\alpha_x&0&0\\0&\alpha_y&0\\0&0&1\end{bmatrix}\begin{pmatrix}x\\y\\1\end{pmatrix}$

- Traditional interpolation-based upsampling (NN, bi-linear) can be implemented as a convolutional layer with fixed weights:



Augmentation pipeline

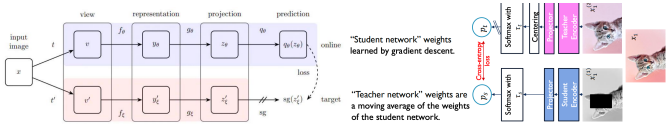
- Needs to reflect **what information the model should disregard** and **what it should focus on**
- **Needs to be hard enough**, otherwise trivial information is learned



SimCLR (Req Large $\mathcal{B}$): $\ell_{i,j}=-\log\frac{\exp(\text{sim}(z_i,z_j)/\tau)}{\sum_{k=1}^{2N}\mathbb{1}[k\neq i]\exp(\text{sim}(z_i,z_k)/\tau)}$ **NT-Xent**: $sim(u,v)=\frac{u^Tv}{\|u\|\|v\|}$. Low τ penalise hard negatives more (negative pairs wrongly mapped closed to each other).

Triplet loss: $\ell(x,x_+,x_-)=\sum_{x\in X}\max\left(0,\|f(x)-f(x_+)\|_2^2-\|f(x)-f(x_-)\|_2^2+\varepsilon\right)$, ε: margin param between x_+,x_-. **Linear Probing:** Froze Model, Train Head. **FT:** Train Model+Head.

BYOL (移除 neg 对，降低 B): 优化正对 emb 的 sim→ 降低需要大 B，更 robust to smaller B than SimCLR。

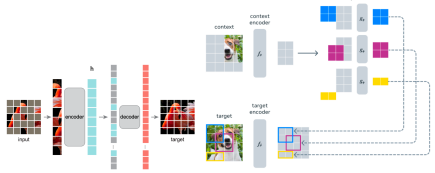


DINO: 使用 softmax+CE 替换 cos sim。Centering 替代q_θ

创建多个 local view (small crops) & 2 global view (big crops). All L & G crops 通过 student，老师只看见 G。

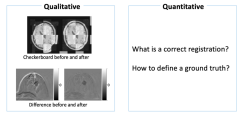
GENERATIVE APPROACHES TO SELF-SUPERVISED LEARNING; MAE: Loss: $MSE=\sum\left(\hat{x}^i-x^i\right)^2$

CL has some drawbacks: large B sizes, design of the augmentation pipeline etc.



Evaluation of image registration

▪ Two types of evaluation: **Qualitative** and **Quantitative** evaluation.



JOINT EMBEDDING PREDICTION (I-JEPA): 训练完整 Rec 很贵。

Loss: pred patch-level representation $\hat{s}_y(i)$, tgt ptch-lvl rep $s_y(i)$. Loss=$l_2(\hat{s},s)$

$Loss=\frac{1}{M}\sum_{i=1}^{M}L_2\left(\hat{s}_y(i),s_y(i)\right)=\frac{1}{M}\sum_{i=1}^{M}\sum_{i\in B_i}\left\|\hat{s}_{y_j}-s_{y_j}\right\|_2^2$

Image Registration

Images: - $f:R^{H,W,C}\to R^1$. **Meta Info**: Scale: element spacing (e.g. in mm), Orientation:main axes's dir, Position: image origin.

Deformations: LowDim Deform Model. Control Point. Finite Element. Dense Displacements field.

Applications in medical imaging: Multi-modal image fusion, Detection of change, Correction of motion, Motion estimation, Segmentation using Registration

Intensity-based Registration

Objective/Cost/Energy fx: $C(T)=D(I\circ T,J)$, $I\circ T$ moving image, J:fixed img, **Optimisation**: $\hat{T}=\arg\min_T C(T)$

Mono-modal vs Multi-modal: Mono-modal(intensities are related 简单函数), Multi: (复杂函数或统计关系)

(Dis)similarity Measures: Intensity differences:

Sum of Squared Differences (SSD):$D_{SSD}(I\circ T,J)=\frac{1}{N}\sum_{i=1}^{N}\left(I(T(x_i))-J(x_i)\right)^2$

Sum of Absolute Differences (SAD):$D_{SAD}(I\circ T,J)=\frac{1}{N}\sum_{i=1}^{N}|I(T(x_i))-J(x_i)|$. Assume: iid

Correlation Coefficient: $D_{CC}(I\circ T,J)=-\frac{1}{N}\sum_{i=1}^{N}(I(T(x_i))-\mu_I)(J(x_i)-\mu_J)\left\{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(I(T(x_i))-\mu_I)^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(I(T(x_i))-\mu_J)^2}\right\}^{-1}$

Assumption: linear relationship between intensity distributions

Intensity distributions: $p(i,j)=\frac{h(i,j)}{N}$, N is number of pixels in one image. $p(i)=\sum_i p(i,j)$

Shannon Entropy: $H(I)=-\sum_i p(i)\log(p_i)$ **Joint entropy:** $H(I,J)=-\sum_i\sum_j p(i,j)\log p(i,j)$

Mutual information: $MI(I,J)=H(I)+H(J)-H(I,J)$,describes how well one image can be explained by another image, can be rewritten in terms of marginal & joint prob $M(I,J)=\sum_i\sum_j p(i,j)\log\frac{p(i,j)}{p(i)p(j)}$,

$D_{MI}(I\circ T,J)=-MI(I\circ T,J)$, **Normalised MI:** $NMI(I,J)=\frac{H(i)+H(j)}{H(i,j)}$ is independent of the amount of overlap between images. Dissimilarity measure $D_{NMI}(I\circ T,J)=-NMI(I\circ T,J)$, Assumption: statistical relationship between intensity distributions.

Image Overlap (Dis)similarity measures are evaluated in the overlapping region of the two images. (easily have Local Optima). Solve: Successively increase degrees of freedom, Gaussian image pyramids

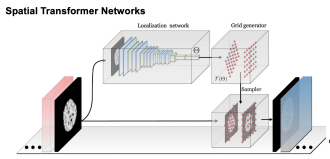
Optimisation Strategies: GD,SGD, Downhill-simplex, Bayesian/Discrete/Convex optmz

Qualitative Eval: Visual assessment

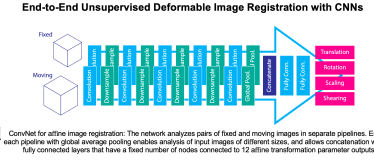
Quantitative Eval: req GT / surrogate measures

Evaluation needs to be independent of registration features or cost function!

Spatial Transformer Networks



End-to-End Unsupervised Deformable Image Registration with CNNs



ConvNet for affine image registration: The network analyzes pairs of fixed and moving images in separate pipelines. Ending each pipeline with global average pooling enables analysis of input images of different sizes, and allows concatenation with the fully connected layers that have a fixed number of nodes connected to 12 affine transformation parameter outputs.

Causality

Would my grades have been better had I studied more?, How effective is a treatment in preventing a disease?

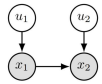
Causality is the relationship between cause and effect. Simpson's Paradox: Correlations may reverse depending on how we aggregate data and its subpopulations

Predictive Modelling Given an image X, train a model to predict some label Y, $P(Y|X)$

Ladder of Causation	Activity	Questions
1. Association *How would seeing X change my belief in Y?* Modeling correlations $P(Y\|X)$	Seeing, Observing	"What if I see …? How are the variables related? How would seeing X change my belief in Y?"
2.Intervention *What will Y be if I do X?* -Use experiments to identify causal effects -Crucial for planning and policy making	Doing, Intervening	What if I do …? How? (What would Y be if I do X? How can I make Y happen?)
3.Counterfactual *What if X had not occurred?* - Counterfactual reasoning - Deduce causes for observed events	Imagining, Retrospection, Understanding	What if X had done …? Why? Was it X that caused Y? What if X had not occurred? What if I had acted differently?

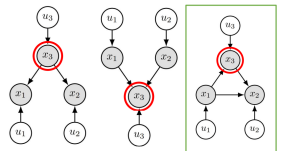
Structural Causal Models (SCM) is a triple: $M:=\langle X,U,F\rangle$, observed $X=\{x_1,\ldots,x_N\}$ and unobserved, $U=\{u_1,\ldots,u_N\}$, ▪ causal mechanisms: $F=\{f_1,\ldots,f_N\}$, ▪ The value of each variable is a function of its **parents** (direct causes): $x_k:=f_k(pa_k,u_k)$, $l=1,\ldots,N$

x_1,x_2 are **endogenous** whereas u_1,u_2 are **exogenous**



$x_1:=f_1(u_1),\quad u_1\sim\mathcal{N}(0,1)$

$x_2:=f_2(x_1,u_2),\quad u_2\sim\mathcal{N}(0,1)$



Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
initialize w_0
for each round $t=1,2,\ldots$ **do**
$m\leftarrow\max(C\cdot K,1)$
$S_t\leftarrow$ (random set of m clients)
for each client $k\in S_t$ **in parallel do**
$w_{t+1}^k\leftarrow\text{ClientUpdate}(k,w_t)$
$w_{t+1}\leftarrow\sum_{k=1}^{K}\frac{n_k}{n}w_{t+1}^k$

ClientUpdate(k,w): // *Run on client k*
$\mathcal{B}\leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
for each local epoch i from 1 to E **do**
for batch $b\in\mathcal{B}$ **do**
$w\leftarrow w-\eta\nabla\ell(w;b)$
return w to server

Observational Distribution:

➔ SCMs with jointly **independent** exogenous noises are Markovian, 如果一个 SCM 的所有外生噪声相互独立，那么称该 SCM 是 Markovian。也就是说，可以把外生变量的联合分布写成各自分布的乘积: $P(u_i,\ldots,u_N)=\prod_{k=1}^{N}P(u_k)$

➔ Markovian SCM induce unique joint **observational distribution** over the endogenous variables, 在这种模型里，每个内生变量只依赖自己的直接父节点，所以整体的观测分布可以用"各变量在给定其父节点下的条件分布"的乘积表示: $P_M(x_1,\ldots,x_N)=\prod_{k=1}^{N}P_M(x_k|pa_k)$, 其中的直接父节点（即直接的因）。➔ Each variable is independent of its non-descendants given its direct causes (causal Markov condition), "鉴于其 direct causes，每个变量都与其非后代变量无关。

Interventional Distribution:

➔ SCMs predict the causal effects of actions via interventions, ➔ Interventions answer causal questions like: E.g. what would be if we set $x_1:=c$? ➔Interventions replace one or more of the structural assignments and are denoted with the do operator: $do(x_k:=c)$,

➔ This induces a submodel M_c and its entailed distribution 被认为 the interventional distribution: $P_{M_c}(X|do(c))$

➔ 这样做会替换掉原先的结构方程，让系统变成一个新的子模型M_C; 我们就能计算干预后的概率分布: $P_{M_c}(X|do(c))$

Counterfactuals:

➔ SCMs can consider hypothetical scenarios: Given that we observed (x_1,x_2), what would $x_2$have been had $x_1$been c? All else being equal, would I have been late had I not missed the train?

➔ Counterfactual inference involves three steps: 1. **Abduction**: Update $P(U)$ given observed evidence, i.e. infer $P(U|X)$

2. **Action**: Intervene by e.g. $do(\widetilde{x_k}:=c)$ and obtain the submodel M_c, 3. **Prediction**: Use $\langle M_c,P(U|X)\rangle$ to compute counterfactuals.

Example: Computing Counterfactuals $\boldsymbol{x_1:=f_1(u_1)=1+u_1, x_2:=f_2(x_1,u_2)=3x_1+u_2}$

Q: Given we observed , what would have been had been 5?

1. Abduction: $x_1:=2=1+u_1\Rightarrow u_1=1, x_2:=4=3\cdot 2+u_2\Rightarrow u_2=-2$

2. Action: $\widetilde{x_1}:=5$, **3.Prediction:** $\widetilde{x_2}=3\widetilde{x_1}+u_2=3\cdot 5-2=13$

Deep Structural Causal Models

➔ Leverage **deep generative models** to learn SCM mechanisms: $x_k:=f_k(pa_k,u_k)$

➔ Tractably estimate causal effects of interventions and perform counterfactual inference, i.e. answer "what if…?" type questions

➔ Abduction is **challenging** in complex problems, e.g. medical imaging.

Morpho-MNIST, Brain Imaging,Chest X-ray Imaging

Evaluating Counterfactuals: Axiomatic Properties 公理角度来评估反事实推理

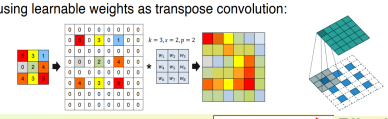
➔ **Soundness Theorem** 所有因果模型中,composition,effectiveness,reversibility 这 3 种性质是必需的。

➔ **Completeness Theorem** 这三种性质也是充分的。

➔ We can measure **Counterfactual Soundness** using these axiomatic properties

Conv Shape:$o=\frac{I+2P-D(K-1)-1}{S}+1$, Receptive field: $R_k=R_{k-1}+Dia_k(K_t-1)\times\Pi_{i=1}^{k-1}S_i$,

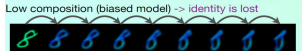
Or using learnable weights as transpose convolution:



Composition

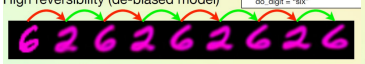
High composition (de-biased model)

Low composition (biased model) -> identity is lost

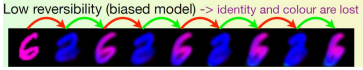


Reversibility

High reversibility (de-biased model)

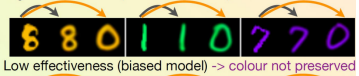


Low reversibility (biased model) -> identity and colour are lost

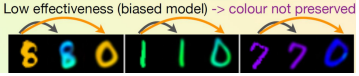


Effectiveness

High effectiveness (de-biased model)



Low effectiveness (biased model) -> colour not preserved



Conclusion & Outlook

➔ It is crucial to account for the **data-generating process** and **causality** in our modelling to avoid biased predictions

➔ Deep SCMs can generate plausible high-fidelity image **counterfactuals**

➔ Latent mediator models enable estimation of **direct**, **indirect** and **total** causal effects for high-dimensional variables Limitations: - Only consider Markovian SCMs; although Markovianity is a common assumption in causality literature, it is strong in most cases - Measuring counterfactual effectiveness relies on separately trained classifiers

Inverse Problem : $y = Ax + n$, Goal: Recover x from y. Inpainting • Deblurring • Denoising • Super-resolution

• Image Reconstruction (Medical imaging). X:原图, Y：观测到的图片

Example deblurring: Least Squares: $\underset{x}{\operatorname{argmin}} D(Ax, y)$, $D(Ax, y) = \frac{1}{2}\|Ax - y\|_2^2$

Least Squares Sol: $\hat{x} = (A^TA)^{-1}A^Ty$, Regularisation: $\arg\min_x D(Ax, y) + \lambda R(x)$, $R(x) = \frac{1}{2}\|x\|_2^2$

Wth Tikhonov regularisation:$\hat{x} = (A^TA + \lambda I)^{-1}A^Ty$, Better conditioned (and noise suppression)

Data Consistency term: $D(Ax, y)$, **Regularisation term (encoding prior knowledge on x):** $R(x)$, **Pram** λ

Common Regularisers: $R(x)$, Tikhonov: $\|\nabla x\|_2^2$,Total variation:$\|\nabla x\|_{2,1} = \sum_{i=0}^{n}\sqrt{\left(\sum_d \left(|\nabla x|_i^d\right)^2\right)}$, Wavelet $\|\Psi x\|$

Regularisation in inverse problems: $y \to \arg\min_x D(Ax, y) + \lambda R(x) \to \hat{x}$, Instead of choosing $\mathcal{R}$ a-priori based on asimple (geometric or other) model of image, learn $\mathcal{R}$ from training data.

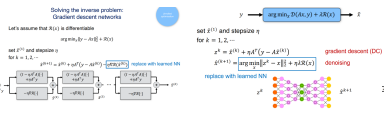
Solving Inverse Problems with Deep Learning

Model agnostic (ignores forward model): $x \to [Blurring] \to [Downsampling] \to y \to NN(y) \to x$

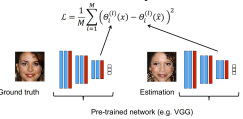
(partly) model agnostic: $x \to [Blurring] \to [Downsampling] \to y \to \tilde{F}^{-1} \to NN(\tilde{y}) \to x$, $\tilde{F}^{-1}$: Up-sample using (linear, cubic) interpolation **Proximal GD** $\boldsymbol{z^k = \hat{x} + \eta A^T(y - A\hat{x}^{(k)})}$.

Deep proximal gradient Decoupled (First learn, then reconstruct) ->

Unrolled Optimisation-> Problem Formulation:



-> Perceptual loss



- Upsample (LR) image to high-resolution (HR/SR) image, -Forward model (going from high- to low-resolution) is straightforward and involves some image degradation followed by downsampling. -Inverse model: e.g. interpolation-based models

Super-Resolution framework: **Post-upsampling SR** • Directly upsamples LR image into SR• Requires learnable upsampling layers• Advantages: – Fast and low memory requirements• Disadvantages: – Network has to learn entire upsampling pipeline – Network typically limited to a specific up-sampling factor $y \to n * conv \to upsample \to \hat{x}$

Pre-upsampling SR: Two stage process: 1 先上采样(e.g. linear interpolation) 2. refining upsampled 用 DNN (usually a CNN). • **ADV**:– Upsampling operation is performed using interpolation, then correct smaller details– Can be applied to a range of upscaling factors and image sizes

• **DISADV**: – Operates on SR image, thus requires 更多计算和内存$y \to upsample \to y_{up} \to n * conv \to \hat{x}$

Progressive upsampling SR: Multi-stage process: Use a cascade of CNNs to progressively reconstruct higher-resolution images.– At each stage, the images are upsampled to higher resolution and refined by CNNs• **ADV**:– Decomposes complex task into simple tasks – Reasonable efficiency **DISADV**: difficult to train very deep models $y \to m * [n * conv \to upsample] \to \hat{x}$

Iterative up-and-down sampling SR: • Approach:– Alternate between upsampling and downsampling (back-projection) operations – Mutually connected up- and down-sampling stages. • **ADV**:– Has shown superior performance as it allows error feedback– Easier training of deep networks. $y \to conv \to n * (upsample + downsample + residual) \to upsample \to \hat{x}$

How to implement upsampling?

$bilinear: \alpha = \frac{x - x_1}{x_2 - x_1}, \beta = \frac{y - y_1}{y_2 - y_1} f(x, y) = (1-\alpha)(1-\beta)I(x_1, y_1) + \alpha(1-\beta)I(x_2, y_1) + (1-\alpha)\beta I(x_1, y_2) + \alpha\beta I(x_2, y_2)$

Loss functions for SR: Pixel-wise loss function (either L1 or L2): $L_p = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x_i}|^p$

Alternative: **Huber loss function**: $L_p = \frac{1}{N}\sum_{i=1}^{N} d(x_i - \hat{x_i})$, $d(a) = 0.5a^2$ if $|a| \le 1$ otherwise $|a| - 0.5$

Perceptual loss: Computes loss on the output θ of an intermediate layer l of a pre-trained network

Total variation: $L = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_d \left(|\nabla x|_i^d\right)^2}$, Assumption: Absolute value of gradient of the image is low, e.g. image is piecewise constant. **GAN LOSS**: $\min_G \max_D E_{x \sim p_{data}(x)}[logD(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))]$

Aim: To detect: – Which images are (real) high-resolution? R(x)=0 if x on image manifold, otherwise infty

• Use discriminator output as loss function: $L = -logD(G(I_{LR}))$. R(x)=0 if x in range(G) otherwise infty

Parameterizing images via generative networks

• Consider a generator network Ψ with a fixed input z_0)

• The network parameters w can 被认为 a parameterization of Images $w \to x = \Psi(z_0; w)$, $z_0 \to VAE_{w_1, w_2, ..., w_n} \to x$, optimization problem: $\min_w \|x - \Psi(z_0; w)\|^2$

Deep Image Prior: 对于大多数生成网络 fitting naturally looking images is easier/faster than fitting others

DIP: Application to inpainting: For inpainting we only reconstruct the visible pixels, implicitly inferring the pixels that are masked out by mask m: $min_w \|m \odot (x - \Psi(z_0; w))\|^2$

X-ray computed tomography (CT): H contrast – H spatial resolution – fast acquisition – but ionising radiation

Magnetic Resonance (MR):H contrast –H spatial res –no ionising radiation –but slow acquisition process

Cost Function Synonym: Loss function, error function, similarity measure • Quantifies how well the model prediction matches targets • Needs to be selected according to the underlying task • Is optimized during training → needs to be differentiable! Eg: MSE

Training Phase Learning model: $\hat{y} = f_\theta(x_n)$, • The parameters θ of the mapping function f_θ are optimized under a cost function

• The cost function quantifies how well : $\hat{y} = f_\theta(x_n)$ is predicted given x_n, The parameters θ by minimizing the cost function $\mathcal{L}$ with learning rate τ: $\theta^{k+1} = \theta^k - \tau \frac{d\mathcal{L}}{d\theta}|_{\theta=\theta^k}$ **Testing Phase (Inference):** • Apply $f_\theta$using the optimized θ to the test set. Generalization: Ability to correctly predict unseen examples

Trustworthy AI/ML AI/ML: The need for data

• The power and effectiveness of AI/ML is critically dependent on the data that is used to train our AI/ML models. – Quality of the data is one of the important aspects that determines the effectiveness of AI/ML models: • Curation of the data (and the associated annotations) • Representativeness of the data (avoiding bias) – Quantity of data • In general, the more data is available for training, the more accurate and robust the resulting AI/ML models become. – Data sharing is more important, not only for training AI/ML models but also for evaluating solutions in multi-institutional/multi-national trials

What are the hurdles 障碍 to getting more data? Human and societal challenges:Cost and effort for collecting and annotating data – Incentives for data sharing (money, fame, other benefits) **Technical challenges**– Data quality– Data annotation– Data exchange formats **Legal challenges**: What is allowed? What consent is required? Regulation (e.g. GDPR) **Privacy challenges** –Ethical –Trust (risks such as privacy breaches, data leaks and re-identification)

Secure and privacy-preserving ML: • Optimal privacy preservation requires implementations that are secure by default so-called privacy by design • **Requirements**: – Minimal or no data transfer **Federated learning**: train a ML model across decentralized clients with local data, without exchanging them – Provision of theoretical and/or technical guarantees of privacy Differential privacy: perturb the data so that information about the single individual is reduced while retaining the capability of learning

Federated Learning 1. Model sent to each client for training on local data 2. Local model updates encrypted and sent back to server for aggregation 3. Aggregated model sent back to local client and model owner (server) 4. Back to 1

In federated learning: • Suppose N training samples are distributed to K clients, P_k is the set of indices of samples at client k, and $n_k = |P_k|$: $L(\theta) = \sum_{k=1}^{K}\frac{n_k}{N}L_K(\theta)$, $L_K(\theta) = \frac{1}{n_k}\sum_{i \in P_k} L(x_i, y_i, \theta)$, $E_{p_k}[L_K] = L\,(iid)$, , $E_{p_k}[L_K] \neq L\,(non-iid)$,

• Suppose a C fraction of clients are selected at each round: C = 1: full-batch (non-stochastic) gradient descent C < 1: stochastic gradient descent (SGD)

Federated SGD: Loop $Client: \nabla L_k(\theta) \to Server: \theta_k^{j+1} = \theta_k^j + \eta\sum_{k=1}^{K}\frac{n_k}{N}\nabla L_k(\theta)$

Federated Averaging: • First, model is randomly initialized on the central server

• For each round t: – A random set of clients are chosen; – Each client performs local GD steps – The server aggregates model parameters submitted by the clients $\theta_k^{j+1} = \sum_{k=1}^{K}\frac{n_k}{N}\theta_k^{j+1}$

Challenges for federated learning

• **Non-IID data** – Training data for a given client is typically site specific, hence the site's local dataset will not be representative of the distribution of training samples.

• **Unbalanced data** – Sites may have a lot or little training data, leading to varying amount of local training data across different sites.

• **Massively distributed data** – There may be extreme scenarios where each site only has very training samples (in the limiting case one example)

• **Communication costs**–Communication between clients and servers occurs communication overheads. The amount of overhead will depend on the number of clients and the frequency of updates from/to server.

• **Privacy protection** – No formal security/privacy provided – Prone to adversarial influence (server or client)

• Federated learning addresses the data sharing problems from the position of data governance; it allow the data owner to choose who is the direct consumer of their data

现有 privacy-preserving 方法

k-anonymity • Idea: 如果 ID 泄露，不应该连接到 individuals Example: [id,John,Male,1956,+ve] -> [id,Male,1950-1960,+ve]

Adversarial priors: assume the attacker has no other information about the targets, not real

Homomorphic Encryption: Example Alice encrypts two values x and y and sends them to Bob, who adds the encrypted values (via "homomorphic addition") to produce an encrypted sum. When Alice receives it back, she decrypts to get x+y. Bob never sees the plaintext x or y. This way, data remains private throughout the entire computation.

Secure multi-party computation: • The individual contributions are 'sharded' and only the individual shards are shared with the participants • If and only if they are all combined together, the final result is revealed • But not the individual parts

• Confidentiality: neither knows the real value• Shared Governance: The value can only be disclosed if everyone agrees

Example Compute average score of exam

Differential Privacy Randomized responses • Enables draw statistical conclusion from datasets without revealing information about individual data points. • Realised by adding a controlled amount of noise • Differential privacy formalizes how we define, measure and track the privacy protection afforded to individual as functions of factors like randomization probabilities and number of times surveyed.

$\mathbb{P}[A(D_1) = O] \le e^\epsilon \mathbb{P}[A(D_2) = O]$, $D_{1,2}$差一个个体的 DB。A 算法。O 是输出。ε隐私预算参数

Concrete attack example: Gradient-based model inversion1. The adversary randomly generates an image-gradient pair 2. The adversary captures the gradient update submitted by the victim 3. Using a suitable cost function (often cosine similarity), the adversary minimises the difference between the captured and the generated updates by perturbing the image they control 4. The algorithm is repeated until the final iteration is reached.

DP-SGD:(1). Compute gradients for each individual sample (they represent independent clients) (2). Clip the calculated gradients to obtain a known sensitivity (3). Add the noise scaled by the sensitivity from step 2 (4). Perform the gradient descent step

Interpretability and Explainability Why Important: – Complexity and prevalence! – Safety and robustness is critical – Generating knowledge • Debugging machine learning models: Data during deployment has noise

• To use machine learning responsibly we need to ensure that – Our values are aligned – Our knowledge is reflected

Interpretability: Common Misunderstandings

•Simple ML models (e.g. linear models or decision trees are interpretable)

• Trust, fairness and interpretability are all the same thing – Interpretability can help to formalize concepts such as fairness or trust– Once formalized it may not be need anymore

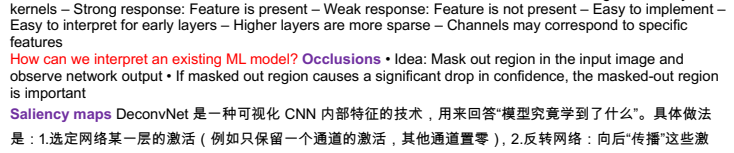
How can we interpret an existing ML model? • Ablation test: How important is a data point or feature? – Train without certain data or features and observe/study the impact – Difficult and expensive • Fit functions (use first derivatives) – Sensitivity analysis – Saliency maps• Visualization and attribution: – Identify input features responsible for model decision • Direct visualization of filters - Easy to implement - Limited practical value – First layers are easy to interpret (mostly low-level features) – Higher layers are more difficult to interpret (non-interpretable features)

• **Problem**: Visualization of filters has limited value • **Solution**: Instead visualize activations generated by kernels – Strong response: Feature is present – Weak response: Feature is not present – Easy to implement – Easy to interpret for early layers – Higher layers are more sparse – Channels may correspond to specific features

How can we interpret an existing ML model? **Occlusions** • Idea: Mask out region in the input image and observe network output • If masked out region causes a significant drop in confidence, the masked-out region is important

Saliency maps DeconvNet 是一种可视化 CNN 内部特征的技术，用来回答"模型究竟学到了什么"。具体做法是：1.选定网络某一层的激活（例如只保留一个通道的激活，其他通道置零），2.反转网络：向后"传播"这些激活，通过"unpooling" 重现原先的空间结构，同时根据池化时记录的"开关"位置（max locations），将特征图逐步还原回输入空间。这样无需重新训练网络，只需要一个几乎与普通反向传播相同的过程（主要在 ReLU 处理上稍有差异），即可重建模型在该通道激活下对应的输入图像模式。由此我们能得到一个可视化的"显著图"（saliency map），帮助理解 CNN 哪些像素特征触发了特定的激活。

• **Question**:– Which pixels are most significant to a neuron?– How would they need to change to most affect the activation of the neuron?• **Solution**:– Use back propagation but differentiate activation with respect to input pixels, not weights



DeepDream / Inceptionism: • Attempt to understand the inner workings of the network • Optimize with respect to image • Idea – Arbitrary image or noise as input – Instead of adjusting network parameters, tweak image towards high "X" where "X" can be • Neuron/Activation map/Layer • Logits/Class probability – Search for images that are "interesting" – Different layers enhance different features• Algorithm – Forward propagate to layer n – No minimization of loss. Instead maximize L2 norm of activations of a particular NN layer – Backpropagate to input layer • Resulting image will show learned features

Robustness: Adversarial Methods Example 57% cofidence + 0.007 noise = "gibbon"(99% confidence)

Adversarial attacks – Perturbation: Assume a linear classifier: $\theta^T x$• We can think of an adversarial example that contains a small, nonperceivable perturbation to the input. Let's denote the perturbation as η: $\tilde{x} = \boldsymbol{x + \eta}$ • Then, the logits of the classifier would be $\theta^T = \theta^T(x + \eta) = \theta^T x + \theta^T\eta$ • Given a small perturbation η, the effect of the perturbation on the logits of the classifier is given by $\theta^T\eta$.

• Idea: – Find a η that causes a change that is non-perceivable and ostensibly innocuous to the human eye, yet destructive and adverse enough for the classifier to the extent that its predictions are no longer accurate. – An adversarial example is one that which maximizes the value of $\theta^T\eta$ to sway the model into making a wrong prediction

• **Problem**: – Need a constraint on η; otherwise, one could just apply a large perturbation to the input • **Solution**: – Apply a constraint such that $\|\eta\|_\infty \le \epsilon$, • Assume a perturbation: $\eta = \epsilon \cdot sign(\theta)$ • What are the bounds of this perturbation? $\eta = \epsilon \cdot sign(\theta) = \theta^\top\eta = \epsilon \cdot \theta^\top sign(\theta) = \epsilon\|\theta\|_1 = \epsilon mn$, m: avg magnitude of an element of θ • This means that the change in activation given by the perturbation increases linearly with respect to n (or the dimensionality). • If n is large, one can expect even a small perturbation capped at ϵ to produce a perturbation big enough to render the model susceptible to an adversarial attack. • Remember that for images n = no. of pixels • Such perturbed examples as referred to as **adversarial examples**

Adversarial attacks - Fast Gradient Sign Method

Key idea: • Perform GD in order to max the loss (the goal of adversarial attack). • Consider the input image x to be a trainable parameter and compute the gradient with respect to the input image to create a perturbation. $\eta = \epsilon \cdot sign(\nabla_x L(\theta, x, y))$

• An adversarial example can be created as: $\tilde{x} = x + \epsilon \cdot sign(\nabla_x L(\theta, x, y))$

How can you use adversarial attacks? 1. Generate adversarial examples 2. Add the generated adversarial examples to the training set

3. Retrain model using training set. **Adversarial Data Augmentation**

-不同对比度或不同加权(MRI T1 vs. T2/FLAIR)时，互信息(Mutual Information) 常被用于度量图像对齐效果，因为它不要求两幅图像的强度具有线性或固定对应关系。-同样是不同序列/不同物理含义(如 T1 和 DWI)的图像，可使用互信息或相关比(Correlation Ratio)，因为它们能够适应不同分布的强度值。-如果是相同模态(如同是 T1 加权，或仅做轻微亮度调整)，可以使用如均方误差(SSD) 或 (归一化)交叉相关(NCC) 等简单度量，因为两幅图在强度上直接对应。

A differentially private algorithm ensures that its output distribution does not change "too much" when a single record in the input dataset is altered. Formally, for two datasets differing by exactly one entry, the ratio of the probabilities of any output is bounded by e^ϵ . The parameter ϵ (often called the privacy budget) controls how much the output can vary; smaller ϵ means stronger privacy guarantees but typically more noise is added to protect individuals' data. MODEL CAP UP, BIAS DOWN, VAR UP

Regularisation penalizes model complexity (using L1, L2, etc.), which reduces variance by preventing overfitting, though it may increase bias. Tuning the regularisation strength helps balance this trade-off.

GANs can serve as a learned prior or manifold constraint for inverse problems. The generator produces images that look realistic, while the discriminator enforces that these images match the true data distribution. By requiring the reconstruction to both fit the observed data and be judged "real" by the discriminator, the GAN effectively regularises the solution to lie on the manifold of valid images.

Multi-atlas label propagation Advantages• robust and accurate (like ensembles)• yields plausible segmentations• fully automatic Disadvantages• computationally expensive • cannot deal well with abnormalities • not suitable for tumour segmentation