Medical Image Analysis with Deep Learning
-
Uncertainties

M. Sdika [1]

[1] Univ.Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS

UMR 5220, U1206, F-69100, LYON, France

Deep learning on Medical Images: Specificities

- Priors
  → anatomy, pixel size, medical expertise, physics of image formation

- small annotated datasets
  → legal aspects, annotation cost

- decision on human people health → need safeguard
  - interpretation, explaination, localization
  - uncertainties estimation

Michaël Sdika
CREATIS -
CNRS

Source of uncertainties:

Data / Aleatoric uncertainties: randomness, task

Feature Overlap

Input Noise:

- Acquisition
- Pre/Post Processing

Annotation Noise

Model / Epistemic uncertainties: learning setup

Architecture

Parameters:

- improper training strategy
- lack of training instances
- local optimal solution

Distribution Shift:
- train vs test

Wenchong He and Zhe Jiang. 2023. A Comprehensive Survey on Uncertainty Quantification for Deep Learning.ACM Comput. Surv.

Source of uncertainties:

Data / Aleatoric uncertainties: randomness, task

Model / Epistemic uncertainties: learning setup

Input Noise:
- Acquisition
- Pre/Post Processing

Feature Overlap

Annotation Noise

Architecture

Parameters:
- improper training strategy
- lack of training instances
- local optimal solution

Distribution Shift:
- train vs test

Wenchong He and Zhe Jiang. 2023. A Comprehensive Survey on Uncertainty Quantification for Deep Learning.ACM Comput. Surv.

Domain Shift

Alzheimer vs Healthy

Covariate shift:

- network input
- change acquisition parameters

Semantic shift:

- network output
- input HIV+ patient
- OOD

Dataset biases, Confounder

Dataset biases, Confounder

Dataset biases, Confounder

Do the "Healthy" and "Patho" data match in:

- imaging parameters: scanner brand ? acquisition parameters ? ...
- population: age ? morphology ? ...
- something else ...

Content

Uncertainties
- Calibration
- Uncertainty Estimation for deep networks
- Conformal Prediction

OOD: Out of Distribution Detection

Right for the Right Reason ?
- Attribution Maps
- Confounder

Michaël Sdika
CREATIS - CNRS

Content

Uncertainties
- Calibration
- Uncertainty Estimation for deep networks
- Conformal Prediction

OOD: Out of Distribution Detection

Right for the Right Reason ?
- Attribution Maps
- Confounder

Michaël Sdika
CREATIS - CNRS

Content

Uncertainties
- Calibration
- Uncertainty Estimation for deep networks
- Conformal Prediction

OOD: Out of Distribution Detection

Right for the Right Reason ?
- Attribution Maps
- Confounder

Michaël Sdika
CREATIS - CNRS

Uncertainties approaches

Gawlikowski etal, Artificial Intelligence Review 2023, A survey of uncertainty in deep neural networks

Michaël Sdika
CREATIS - CNRS

Scoring Rule

Scoring rule: $S(p, y)$
loss or penalty when for the output prediction p and the observation y

Strictly proper scoring rule:
the best score is achieved when the output p is the true distribution of y

Example:
- negative log likelyhood NLL (CE for categorical pb):

$$-\log(p_y)$$

- Brier Score:

$$\frac{1}{C}\sum_c (p_c - y_c)^2$$

Michaël Sdika
CREATIS - CNRS

The network ouput as an uncertainty measure ??

$\longrightarrow$ Classifier $\longrightarrow p$

Is it the probability to give a correct output ?

$\longrightarrow$ **network confidence**

Reliability Diagram

Reliability Diagram

Reliability Diagram

Reliability Diagram

Michaël Sdika
CREATIS - CNRS

Reliability Diagram

Expected Calibration Error (ECE)

$$\text{ECE} = \frac{1}{B}\sum_{b=1}^{B} |\text{conf}_b - \text{acc}_b|$$

Multi-class, either:

- avg accuracy VS avg confidence
- classwize reliability diagram

Reliability Diagram

Michaël Sdika
CREATIS - CNRS

Reliability Diagram

under-confident **calibrated** **over-confident**

Michaël Sdika
CREATIS - CNRS

Modern deep networks are not calibrated

cause:

- depth/width/batch norm
- Cross Entropy loss:
  - $-\log(p_y)$

Guo *et. al.* 2017, On Calibration of Modern Neural Networks

Michaël Sdika
CREATIS - CNRS

Label Smoothing

$\longrightarrow$ Net $\longrightarrow p$

y GT label

correct with probability $1 - \epsilon$

- Last layer: softmax with K classes
- $\Rightarrow$ still use CE
- in the one hot vector replace 0/1 by $\frac{\epsilon}{k-1}$ / $1 - \epsilon$

Szegedy et al, CVPR 2016, Rethinking the inception architecture for computer vision

Deep Learning, Goodfellow et al

Michaël Sdika
CREATIS - CNRS

Spatially Varying Label Smoothing

SVLS

- gaussian smooth GT label maps
- use it in CE

NACL

$$\min CE + P(\|l - \tau\|)$$

Islam & Glocker, IPML 2021. Spatially varying label smoothing: Capturing uncertainty from expert annotations
Murugesan ... Dolz , MEDIA 2024, Neighbor-Aware Calibration of Segmentation Networks with Penalty-Based Constraints

Michaël Sdika
CREATIS - CNRS

Temperature scaling

Softmax :

$$\frac{e^{z_k}}{\sum e^{z_i}} \longrightarrow \frac{e^{z_k/T}}{\sum e^{z_i/T}}$$

On the validation dataset:

$$\min_T \mathrm{NLL}(T)$$

Guo *et. al.* 2017, On Calibration of Modern Neural Networks

Michaël Sdika
CREATIS - CNRS

Focal Loss

Lin etal, ICCV 2017. Focal loss for dense object detection
Mukhoti etal Neurips 2020. Calibrating deep neural networks using focal loss

Michaël Sdika
CREATIS - CNRS

Focal Loss

Lin etal, ICCV 2017. Focal loss for dense object detection
Mukhoti etal Neurips 2020. Calibrating deep neural networks using focal loss

Michaël Sdika
CREATIS - CNRS

Content

Uncertainties
- Calibration
- Uncertainty Estimation for deep networks
- Conformal Prediction

OOD: Out of Distribution Detection

Right for the Right Reason ?
- Attribution Maps
- Confounder

Michaël Sdika
CREATIS - CNRS

Bayesian Networks

Inference: MC sampling

- D: training dataset
- draw random $w_k \sim q(w)$
- $p(y|x, D) \approx \frac{1}{K} \sum_{k=1}^{K} p(y|x, w_k)$

Uncertainty Estimation:

- $p^c = p(y = c|x, w_k)$
- $H[y|x, D] \approx -\sum_c p^c \log(p^c)$

Michaël Sdika
CREATIS - CNRS

Dropout Regularization

(a) Standard Neural Net (b) After applying dropout.

Random set units to 0 during training

Test error with/without dropout

Srivastava *et. al.* 2014, Dropout: A Simple Way to Prevent Neural Networks from Overfitting

Michaël Sdika
CREATIS - CNRS

MC Dropout for uncertainty estimation

network: $x \rightarrow y = y(x)$

- dropout during inference: sample several outputs $(y_i(x))_{1,N}$
- Gal 2016: these samples are $\approx$ drawn from $p(\hat{y}|x^*)$
- estimate first and second moments of $p(y|x)$

$$E(y|x) \approx SampleMean\,(y_i)$$

$$Var(y|x) \approx \tau I + \text{SampleVar}(y_i)$$

Gal & Ghahramani, ICML 2016: Dropout as a bayesian approximation: representing model uncertainty in deep learning

Michaël Sdika
CREATIS - CNRS

Uncertainty measures for MS lesion segmentation

- comparison of several statistical measures from dropout samples
- thesholding on uncertainties for lesion segmentation
  $\rightarrow$ keep only certain lesions

Nair, *et. al.* MICCAI 2018, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation

Michaël Sdika
CREATIS - CNRS

Deep ensemble

- train a set of network
- stochasticity from random init
- in practice, K=5

Lakshminarayanan, *et. al.* NIPS 2017, Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles

Michaël Sdika
CREATIS - CNRS

Test Time Augmentation (TTA)

During test:
- generate multiple output
- prediction: mean / mode
- uncertainty: std-dev / entropy

aleatoric uncertainty related to image transformations and noise

Wang et al, 2019, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with CNN.

Michaël Sdika
CREATIS - CNRS

Uncertainty Estimations for Medical Image Segmentation

Unet
BRATS+skin lesion

Jungo & Reyes, MICCAI 2019, Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation

Michaël Sdika
CREATIS - CNRS

Uncertainty with respect to expert annotations

Lung CT images with segmentation made by 4 operators

Kohl *et. al.* NIPS 2018, A probabilistic u-net for segmentation of ambiguous images

Michaël Sdika
CREATIS - CNRS

Uncertainty with respect to expert annotations

Kohl *et. al.* NIPS 2018, A probabilistic u-net for segmentation of ambiguous images
Baumgartner etal. MICCAI 2019, Phiseg: Capturing uncertainty in medical image segmentation

Michaël Sdika
CREATIS - CNRS

Uncertainty with respect to expert annotations: Train

Kohl *et. al.* NIPS 2018, A probabilistic u-net for segmentation of ambiguous images

Michaël Sdika
CREATIS - CNRS

Uncertainty with respect to expert annotations: Results

Kohl *et. al.* NIPS 2018, A probabilistic u-net for segmentation of ambiguous images

Michaël Sdika
CREATIS - CNRS

Generating segmentation distrbution with diffusion models

Rahman etal, CVPR 2023, Ambiguous Medical Image Segmentation using Diffusion Models

Michaël Sdika
CREATIS - CNRS

Generating segmentation distrbution with diffusion models

Method	GED (↓)	CI (↑)	D_{max} (↑)
DDPM-det-Seg [59]	1.081	0.616	0.548
DDPM-Prob-Seg	0.417	0.683	0.689
CIMD (Ours)	**0.321**	**0.759**	**0.915**

Rahman etal, CVPR 2023, Ambiguous Medical Image Segmentation using Diffusion Models

Regression

$$x \longrightarrow \boxed{\text{Net}} \longrightarrow \hat{a}$$

$$\text{MSE} = \|\hat{a} - a\|^2$$

Nix & Weigend, ICNN 1994. Estimating the mean and variance of the target probability distribution

Michaël Sdika
CREATIS - CNRS

Regression

$$x \longrightarrow \boxed{\text{Net}} \longrightarrow \hat{a}$$

$$x \longrightarrow \boxed{\text{Net}} \longrightarrow \begin{matrix}\mu \\ \sigma\end{matrix}$$

MSE $= \|\hat{a} - a\|^2$

$A(x) \sim \mathcal{N}(\mu(x), \sigma(x))$

likelyhood: $\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\|\mu - a\|^2}{2\sigma^2}}$

homoscedastic ($\sigma =$ cst):

$\Rightarrow$ MSE = NLL

$\longrightarrow$ NLL $: \log(\sigma) + \frac{\|\mu - a\|^2}{2\sigma^2}$

Nix & Weigend, ICNN 1994. Estimating the mean and variance of the target probability distribution

Quantile Regression

$$\min_x \sum_i r(x - x_i)$$

Michaël Sdika
CREATIS - CNRS

Quantile Regression

$$\min_x \sum_i r(x - x_i)$$

$$r_\tau(x) = \left| \begin{array}{ll} -(1-\tau)x & \text{if } x < 0 \\ x\tau & \text{otherwize} \end{array} \right.$$

Michaël Sdika
CREATIS - CNRS

Quantile Regression

Loss: $r_{25}(q_{25} - y) + r_{50}(q_{50} - y) + r_{75}(q_{75} - y)$

Michaël Sdika
CREATIS - CNRS

Content

Uncertainties
- Calibration
- Uncertainty Estimation for deep networks
- Conformal Prediction

OOD: Out of Distribution Detection

Right for the Right Reason ?
- Attribution Maps
- Confounder

Michaël Sdika
CREATIS - CNRS

Conformal Prediction

{ fox squirrel 0.99 }

{ fox squirrel 0.82, gray fox, 0.03, bucket, 0.02, rain barrel 0.02 }

{ marmot, 0.30, fox squirrel, 0.22, mink, 0.18, weasel, 0.16, beaver, 0.03, polecat 0.01 }

for a test image I, find a set of labels $C(I)$ s.t.:

$$P\left(y \in C(I)\right) \geq 1 - \alpha$$

Angelopoulos & Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv 2021

Michaël Sdika
CREATIS - CNRS

Conformal Prediction: Classification

Settings:

- classifier $\hat{f}(x) \in [0,1]^K$
- calibration data $(x_i, y_i)_{i \in [1,n]}$
- error rate $\alpha \in [0,1]$
- CALIB / TEST: exchangeable, same distribution

Problem:

- $(x, y) \in$ TEST
- find a set of labels $C(x)$ s.t.

$$1 - \alpha \leq P\left(y \in C(x)\right) \leq 1 - \alpha + \frac{1}{n+1}$$

Angelopoulos & Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv 2021

Michaël Sdika
CREATIS - CNRS

Conformal Prediction: Solving the problem

- heuristic uncertainty measure for the model ex: $\rightarrow \hat{f}(x)$
- Define a score function $s(x, y)$ (lower=better) ex: $\rightarrow s(x, y) = 1 - \hat{f}_y(x)$
- Define the set C_q:

$$C_q(x) = \{y \text{ s.t } s(x, y) < q\}$$

- Notice that

$$P\left(y \in C_q(x)\right) = P\left(s(x, y) < q\right) = F_S(q)$$

is the cumulative distribution function of $s(X, Y)$

$$\Rightarrow \hat{q} \approx (1 - \alpha)\text{-quantile}$$

Angelopoulos & Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv 2021

Michaël Sdika
CREATIS - CNRS

Conformal Prediction: Solution for the Example

Calibration data:

- conformal score:

$$s_i = 1 - f_{y_i}(x_i)$$

- $q \approx (1-\alpha)$-quantile of scores

$x \in$ TEST (y unknown):

$$C(x) = \left\{ y | 1 - \hat{f}_y(x) \leq q \right\}$$

Angelopoulos & Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv 2021

Michaël Sdika
CREATIS - CNRS

Adaptive prediction set

$$s(x,y) = \sum_{f_z(x) \geq f_y(x)} f_z(x)$$

Angelopoulos & Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv 2021

Michaël Sdika
CREATIS - CNRS

Conclusion on Conformal Prediction

- rigourously calibrate any model
- scalar prediction: quantile regression

No Magics !!

- bad model / bad heuristic uncertainties / small calib set
  $\rightarrow$ return all the labels
- metrics:
  - coverage: is the true label in the prediction set ?
  - size: how large is the prediction set ?

Angelopoulos & Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv 2021

Michaël Sdika
CREATIS - CNRS

Content

Uncertainties
- Calibration
- Uncertainty Estimation for deep networks
- Conformal Prediction

OOD: Out of Distribution Detection

Right for the Right Reason ?
- Attribution Maps
- Confounder

Michaël Sdika
CREATIS - CNRS

Uncertainties estimation is not robust to domain shift

Ovadia, et al. 2019 "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift."

Michaël Sdika
CREATIS - CNRS

ODD

Yang et al., arXiv, Generalized out-of-distribution detection: A survey

Michaël Sdika
CREATIS - CNRS

ODD

	CIFAR-10			CIFAR-100			ImageNet-200			ImageNet-1K		
	Near-OOD	Far-OOD	ID Acc.	Near-OOD	Far-OOD	ID Acc.	Near-OOD	Far-OOD	ID Acc.	Near-OOD	Far-OOD	ID Acc.
- Post-hoc Inference Methods												
OpenMax [2] (CVPR'16)	87.62±0.29%	89.62±0.19%	95.06±0.36%	76.41±0.28%	79.68±0.41%	77.25±0.13%	80.27±0.10%	90.20±0.17%	86.73±0.08%	74.77	89.26	76.18
MSP [1] (ICLR'17)	88.03±0.25%	90.73±0.45%	95.06±0.36%	80.27±0.11%	77.70±0.44	77.25±0.13%	83.34±0.06%	90.13±0.06%	86.73±0.08%	76.02	85.23	76.18
TempScale [4] (ICML'17)	88.08±0.31%	90.97±0.52%	95.06±0.36%	80.90±0.85%	78.74±0.51%	77.25±0.13%	**83.68**±0.04%	90.82±0.06%	86.73±0.08%	77.14	87.56	76.18
ODIN [5] (ICLR'18)	82.87±1.03%	87.96±0.61%	95.06±0.36%	79.90±0.51%	79.28±0.21%	77.25±0.13%	80.27±0.08%	91.71±0.16%	86.73±0.08%	74.75	89.47	76.18
MDS [11] (NeurIPS'18)	84.20±0.46%	89.73±1.30%	95.06±0.36%	58.69±0.08%	69.59±1.78	77.25±0.13%	61.93±0.51%	74.72±0.26%	86.73±0.08%	55.44	74.25	76.18
MDSEns [11] (NeurIPS'18)	60.43±0.26%	73.50±0.27%	95.06±0.36%	46.31±0.24%	66.03±0.08	77.25±0.13%	54.32±0.24%	69.27±0.57%	86.73±0.08%	49.67	67.52	76.18
RMDS [45] (arXiv '21)	89.80±0.26%	92.20±0.21%	95.06±0.36%	80.15±0.31%	**82.92**±0.42%	77.25±0.13%	82.57±0.25%	88.08±0.34%	86.73±0.08%	76.99	86.38	76.18
Gram [46] (ICML'20)	58.00±0.85%	71.73±0.36%	95.06±0.36%	51.60±0.75%	73.36±1.46	77.25±0.13%	67.67±1.07%	71.19±0.54%	86.73±0.08%	61.70	79.71	76.18
EBO [47] (NeurIPS'20)	87.58±0.46%	91.21±0.92%	95.06±0.36%	80.91±0.06%	79.77±0.61%	77.25±0.13%	82.50±0.05%	90.86±0.21%	86.73±0.08%	75.89	89.47	76.18
OpenGAN [48] (ICCV'21)	53.71±0.68%	54.61±0.51%	95.06±0.36%	65.98±1.26%	67.88±2.18%	77.25±0.13%	59.78±1.50%	73.15±1.07%	86.73±0.08%	N/A	N/A	N/A
GradNorm [44] (NeurIPS'21)	54.90±0.08%	57.55±0.22%	95.06±0.36%	70.13±0.45%	69.14±1.95%	77.25±0.13%	72.75±0.40%	84.26±0.07%	86.73±0.08%	72.96	90.25	76.18
ReAct [12] (NeurIPS'21)	87.11±0.61%	90.42±1.41%	95.06±0.36%	80.77±0.05%	80.39±0.48%	77.25±0.13%	81.87±0.08%	92.31±0.56%	86.73±0.08%	77.38	95.67	76.18
MLS [14] (ICML'22)	87.52±0.07%	91.10±0.09%	**81.05**±0.07%	79.07±0.07%	82.90±0.04%	91.11±0.16%	82.90±0.04%	91.11±0.16%	86.73±0.08%	76.48	89.57	76.18
KLM [14] (ICML'22)	79.19±0.00%	82.68±0.21%	95.06±0.36%	76.56±0.25%	76.24±0.53	77.25±0.13%	80.36±0.00%	88.53±0.11%	86.73±0.08%	76.64	87.60	76.18
VIM [15] (CVPR'22)	88.44±0.20%	**92.48**±0.21%	95.06±0.36%	74.98±0.11%	81.70±0.62%	77.25±0.13%	78.68±0.24%	91.26±0.19%	86.73±0.08%	72.08	92.68	76.18
KNN [16] (ICML'22)	**90.64**±0.30%	92.96±0.41%	95.06±0.36%	80.18±0.11%	82.40±0.17%	77.25±0.13%	81.57±0.17%	93.16±0.22%	86.73±0.08%	71.10	90.18	76.18
DICE [69] (ECCV '22)	78.34±0.79%	84.23±1.00%	95.06±0.36%	79.38±0.25%	80.01±0.18	77.25±0.13%	81.78±0.34%	90.80±0.31%	86.73±0.08%	73.07	90.95	76.18
RankFeat [70] (NeurIPS '22)	79.46±2.52%	75.87±2.06%	95.06±0.36%	61.80±1.28%	67.10±1.42	77.25±0.13%	56.92±1.50%	38.22±2.01%	86.73±0.08%	50.99	53.93	76.18
ASH [17] (ICLR'23)	75.27±1.06%	78.49±0.50%	95.06±0.36%	78.20±0.11%	80.58±0.60	77.25±0.13%	82.38±0.50%	**93.98**±0.27%	86.73±0.08%	**78.17**	**95.74**	76.18
SHE [18] (ICLR'23)	81.54±0.51%	85.32±1.45%	95.06±0.36%	78.95±0.18%	76.92±1.18	77.25±0.13%	80.18±0.25%	89.81±0.61%	86.73±0.08%	75.78	90.92	76.18
- Training Methods (w/o Outlier Data)												
ConfBranch [51] (arXiv '18)	89.84±0.26%	92.85±0.29%	94.88±0.05%	71.05±0.52%	68.98±1.19%	78.59±0.27%	79.93±0.26%	90.43±0.38%	85.82±0.07%	70.66	83.94	75.63
RotPred [52] (NeurIPS' 19)	**92.68**±0.27%	96.62±0.16%	**95.38**±0.31%	76.47±0.16%	**88.40**±0.17%	76.03±0.16%	81.59±0.20%	92.56±0.09%	**86.37**±0.36%	**76.52**	90.00	**76.55**
G-ODIN [32] (CVPR'20)	89.12±0.57%	95.51±0.31%	94.70±0.15%	77.15±0.28%	85.67±1.50	74.40±0.04%	77.28±1.30%	92.33±0.11%	84.59±0.20%	70.77	85.51	74.85
CSI [7] (NeurIPS' 20)	89.5±0.40%	92.08±0.91%	91.19±0.14%	71.45±0.25%	66.31±0.21	61.00±0.46%	N/A	N/A	N/A	N/A	N/A	N/A
ARPL [53] (TPAMI'21)	87.44±0.19%	89.31±0.52%	93.66±0.11%	74.94±0.95%	73.69±1.88	70.70±1.08	82.02±0.30%	89.23±0.113	83.95±0.32%	76.30	85.50	75.87
MOS [54] (CVPR'21)	71.45±1.00%	76.41±0.93%	94.83±0.11%	80.48±1.18%	80.17±1.21	76.98±0.28%	89.84±0.40%	80.48±0.92%	85.80±0.20%	72.85	82.75	72.81
VOS [6] (ICLR'22)	87.70±0.60%	90.83±0.91%	94.31±0.66%	**80.93**±0.28%	81.52±0.98	77.29±0.13%	82.51±0.15%	91.00±0.28%	86.25±0.08%	N/A	N/A	N/A
LogitNorm [55] (ICML'22)	92.33±0.06%	**96.74**±0.06%	94.30±0.19%	78.47±0.31%	81.53±1.26	76.38±0.15%	**82.66**±0.15%	93.04±0.21%	86.04±0.19%	74.62	91.54	76.45
CIDER [63] (ICLR'23)	90.71±0.16%	94.71±0.36%	N/A	73.18±0.38%	80.49±0.66%	N/A	80.58±1.31%	90.68±1.00%	N/A	68.97	**92.18**	N/A
NPOS [42] (ICLR'23)	89.78±0.53%	94.07±0.49%	N/A	78.35±0.37%	82.29±1.55%	79.40±0.36%	79.44±0.36%	**94.49**±0.47%	N/A	N/A	N/A	N/A
- Training Methods (w/ Outlier Data)												
OE [17] (NeurIPS' 18)	**94.82**±0.21%	**96.00**±0.11%	94.67±0.24%	**88.30**±0.18%	**81.41**±0.49	**76.84**±0.24%	**84.84**±0.30%	**89.02**±0.38%	85.82±0.17%	N/A	N/A	N/A
MCD [38] (ICCV'19)	91.03±0.12%	91.00±1.19%	94.95±0.04%	77.07±0.32%	74.72±0.78	75.83±0.04%	83.62±0.09%	88.94±0.80%	**86.12**±0.17%	N/A	N/A	N/A
UDG [39] (ICCV'21)	89.91±0.28%	94.08±0.09%	92.70±0.04%	78.02±0.18%	79.59±1.77%	71.59±0.64%	74.38±2.63%	82.09±2.76%	68.15±0.21%	N/A	N/A	N/A
MixOE [40] (WACV'23)	88.73±0.42%	91.93±0.60%	94.55±0.16%	80.95±0.28%	76.40±1.44	75.13±0.08%	82.62±0.67%	88.57±0.45%	85.73±0.67%	N/A	N/A	N/A

Yang et al., arXiv, Generalized out-of-distribution detection: A survey

Zhang et al, arxiv, OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection

Michaël Sdika
CREATIS - CNRS

ODD

ID / OOD
from Task Net internal

Michaël Sdika
CREATIS - CNRS

Confidence based OOD

Maximum Class Probability

ID : $\max_c p_c >$ thresh

MCP with better calibration:

- MCP Dropout
- MCP Deep ensemble

ODIN

- $\tilde{x}$: adv attack to increase confidence
- adjust temperature (G-ODIN:large)
- score : MCP $(\tilde{x})$

DUQ

- feature learning
- contrastive learning
- score : RBF kernel distance to centroid

KNN

- contrastive
- KNN

Mahalanobis

Berger, et al. UNSURE MICCAI 2021, Confidence-based out-of-distribution detection: a comparative study and analysis
Hendrycks & Gimpel, ICLR 2017, A baseline for detecting misclassified and out-of-distribution examples in neural networks
Liang etal ICLR 2018: Enhancing the reliability of out-of-distribution image detection in neural networks
Van Amersfoort etal, ICML 2020 : Uncertainty estimation using a single deep deterministic neural network
Sun, et al, ICML 2022, Out-of-distribution detection with deep nearest neighbors

Michaël Sdika
CREATIS - CNRS

Reconstruction-based OOD

Assumption: better reco for ID than OOD

score: $\|\hat{x} - x\|$

$$\|\hat{x} - x\| + \lambda(z - \mu)\Sigma^{-1}(z - \mu)$$

Denouden etal, arxiv 2018, Improving Reconstruction Autoencoder Out-of-distribution Detection with Mahalanobis Distance

Michaël Sdika
CREATIS - CNRS

Reconstruction-based OOD

Denouden etal, arxiv 2018, Improving Reconstruction Autoencoder Out-of-distribution Detection with Mahalanobis Distance

Michaël Sdika
CREATIS - CNRS

Content

Uncertainties
- Calibration
- Uncertainty Estimation for deep networks
- Conformal Prediction

OOD: Out of Distribution Detection

Right for the Right Reason ?
- Attribution Maps
- Confounder

Michaël Sdika
CREATIS - CNRS

Content

Uncertainties
- Calibration
- Uncertainty Estimation for deep networks
- Conformal Prediction

OOD: Out of Distribution Detection

Right for the Right Reason ?
- Attribution Maps
- Confounder

Michaël Sdika
CREATIS - CNRS

Attribution maps: WHERE, not WHAT

Post hoc: apply it once the network has been trained

blue: push toward negative
red : push toward positive

Michaël Sdika
CREATIS - CNRS

Lots of methods, several approaches

Perturbation // Gradient // Activation

Kumar & Kamath. "Attribution-based XAI methods in computer vision: A review." arXiv:2211.14736 (2022).

Michaël Sdika
CREATIS - CNRS

Occlusion

$$A(I,B)(x) = F(I) - F(I^x)$$

I^x: baseline value at pixel x otherwize I

- I^x might be OOD
- untractable $\Rightarrow$ patch, superpixel...

Zeiler & Fergus, ECCV 2014, Visualizing and Understanding Convolutional Networks

Michaël Sdika
CREATIS - CNRS

Saliency maps: the gradient as an attribution maps

Saliency Maps / Gradient:

$$A(I) = \nabla_I F$$

Input x Gradient:

$$A(I) = (I - B)\nabla_I F$$

saliency map

Simonyan et al, Deep inside convolutional networks: Visualising image classification models and saliency maps. ICLR 2014
Shrikumar et al, Learning important features through propagating activation differences. ICML 2017

Michaël Sdika
CREATIS - CNRS

Global vs Local Attributions

$$C = 1.05x_1 + 10x_2$$

$$x_1 = 1.000.000 \text{ and } x_2 = 10.000$$

Local:

$$R1 = \frac{\partial C}{\partial x_1} = 1.05$$

$$R2 = \frac{\partial C}{\partial x_2} = 10$$

How do you behave locally ?

Global:

$$R1 = x_1 \frac{\partial C}{\partial x_1} = 1.050.000$$

$$R2 = x_2 \frac{\partial C}{\partial x_2} = 10.000$$

How initial investments contributed ?

Ancona et al ICLR 2018, Towards better understanding of gradient-based attribution methods for deep neural networks

Michaël Sdika
CREATIS - CNRS

Retropagation rules: LRP, DeepLIFT, ...

Method	Attribution $R_i^c(x)$
Gradient * Input	$x_i \cdot \frac{\partial S_c(x)}{\partial x_i}$
Integrated Gradient	$(x_i - \bar{x}_i) \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\tilde{x})}{\partial(\tilde{x}_i)}\Big\vert_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$
ϵ-LRP	$x_i \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z)}{z}$
DeepLIFT	$(x_i - \bar{x}_i) \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z) - f(\bar{z})}{z - \bar{z}}$
Occlusion-1	$S_c(x) - S_c(x_{[x_i=0]})$

Example of attributions on MNIST

Ancona et al ICLR 2018, Towards better understanding of gradient-based attribution methods for deep neural networks

Michaël Sdika
CREATIS - CNRS

Path Methods: Integrated Gradient

$$A(I,B) = (I-B)\int_0^1 \frac{\partial F}{\partial I}\left(B+\alpha(I-B)\right)d\alpha$$

satisfy: Linearity w.r. model, Sensitivity, Implementation invariance
Completeness ($\sum_x A(I)_x = F(I) - F(B)$)

M Sundararajan, et al ICML 2017, Axiomatic attribution for deep networks

Michaël Sdika
CREATIS - CNRS

CAM

$$y = \sum_i w_i a_i$$
$$\propto \sum_{i,x,y} w_i f_i(x,y)$$
$$\propto \sum_{x,y} \sum_i w_i f_i(x,y)$$
$$\propto \sum_{x,y} A(x,y)$$

Zhou et al, CVPR 2016, Learning deep features for discriminative localization

Michaël Sdika
CREATIS - CNRS

CAM

Zhou et al, CVPR 2016, Learning deep features for discriminative localization

Michaël Sdika
CREATIS - CNRS

Grad-CAM

$$w_k = \text{avg}_X \left(\frac{\partial y}{\partial f_k} \right)$$

$$A(I) = \text{relu} \left(\sum_k w_k f_k(X) \right)$$

Can be used on any layer

(usually last conv)

GradCAM Guided GradCAM

Selvaraju et al, ICCV 2017, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Michaël Sdika
CREATIS - CNRS

Attribution maps constrained training

Constraint:
I is Healthy $\Rightarrow A(I)$ is blue

naive constrained

Michaël Sdika
CREATIS - CNRS

Content

Uncertainties
- Calibration
- Uncertainty Estimation for deep networks
- Conformal Prediction

OOD: Out of Distribution Detection

Right for the Right Reason ?
- Attribution Maps
- Confounder

Michaël Sdika
CREATIS - CNRS

Counfounder effect: Right for the wrong reason

TODO ?: add old vs young example
cf pres dl school 2018
/home/sdika/work/pres/deep-learning/archive-cpe-cnrs-dlmi/deep-learning-school/2018

Michaël Sdika
CREATIS - CNRS

Counfounder effect: Right for the wrong reason

Horse-picture from Pascal VOC data set

Source tag present

↓

Classified as horse

No source tag present

↓

Not classified as horse

Artificial picture of a car

Lapuschkin et al. "Unmasking Clever Hans predictors and assessing what machines really learn." Nature communications 2019

Michaël Sdika
CREATIS - CNRS

Obvious Counfounder in Medical Images

pacemaker on chest radiograph

ruler on skin lesion images

Kawahara et al. "Seven-point checklist and skin lesion classification using multitask multimodal neural nets." IEEE JBHI 22018.

Michaël Sdika
CREATIS - CNRS

Attribution maps comparison with artificial confounders

CS: % high att within confound area

NCC: pred. change ⇒ att change

Sun, Koch, Baumgartner, MICCAI 2023, Right for the Wrong Reason: Can Interpretable ML Techniques Detect Spurious Correlations?

Michaël Sdika
CREATIS - CNRS

The brain shape as a confounder

Wargnier-Dauchelle, ..., Sdika, IEEE-ISBI 2018, A More Interpretable Classifier for Multiple Sclerosis.
Wargnier-Dauchelle, Grenier, Sdika, MIDL 2024, An unexpected confounder: how brain shape can be used to classify MRI scans ?

Michaël Sdika
CREATIS - CNRS

The brain shape as a confounder

Classif MRI → P_{patho}

Classif PMAPS → P_{patho}

Classif Brain Mask → P_{patho}

IG-MRI IG-CSF IG-WM IG-GM

Classification task			TNR	TPR	BA
IXI	vs	BraTS	1.00	1.00	1.00
IXI	vs	OFSEP	0.54	0.70	0.62
IXI	vs	HCP	1.00	1.00	1.00
IXI	vs	IBC/kirby/MPI	1.00	0.86	0.93
HCP	vs	IBC/kirby/MPI	0.99	0.95	0.97
ADNI-CN	vs	ADNI-AD	0.35	0.74	0.55

Wargnier-Dauchelle, ..., Sdika, IEEE-ISBI 2018, A More Interpretable Classifier for Multiple Sclerosis.
Wargnier-Dauchelle, Grenier, Sdika, MIDL 2024, An unexpected confounder: how brain shape can be used to classify MRI scans ?

Michaël Sdika
CREATIS - CNRS

The brain shape as a confounder: assessment and solution

image

crop

reg

During test:
- Change an image from class A s.t. its brain shape is in class B
- Check if the performance decrease

Solution:
brain shape normalization

Michaël Sdika
CREATIS - CNRS

Confounder Free Training

CF-net architecture

- update FE+P
  estimate y

- update CP **on control only**
  estimate c
  avoid confounding of y on CP

- adversarial update FE:
  to fool CP
  $\Rightarrow$ F cannot predict c

Zhao et al, Nature Communications 2020, Training confounder-free deep learning models for medical applications

Michaël Sdika
CREATIS - CNRS

Confounder Free Training: results

Zhao et al, Nature Communications 2020, Training confounder-free deep learning models for medical applications

Michaël Sdika
CREATIS - CNRS

Conclusion

Thanks for your attention !!