HELMHOLTZ AI

UNCERTAINTY QUANTIFICATION IN MACHINE LEARNING

A Primer

Peter Steinbach, Till Korten, Sebastian Starke, Steve Schmerler

Helmholtz-Zentrum Dresden-Rossendorf / 2025-10-02

TODAY'S AGENDA

- Uncertainties!
- UQ For Regression
  - MCDropout
  - Uncertainty Calibration
  - DeepEnsembles
  - UQ for Instance Segmentations
- UQ For Classification
- Conclusions

UNCERTAINTIES!

WHAT IS UNCERTAINTY?

Definition (Merriam-Webster today)

uncertainty: not known beyond doubt, not having certain knowledge, not clearly identified or defined, not constant, indefinite, not certain to occur, not reliable

WHAT IS UNCERTAINTY?

Definition (Merriam-Webster today)

uncertainty: not known beyond doubt, not having certain knowledge, not clearly identified or defined, not constant, indefinite, not certain to occur, not reliable

Definition (Wikipedia today)

Uncertainty refers to *epistemic* situations involving imperfect or unknown information. It applies to predictions of future events, to physical measurements that are already made, or to the unknown. Uncertainty arises in partially observable or stochastic environments, as well as due to ignorance, indolence, or both.
likely from [Russell and Norvig 2010]

WHAT IS UNCERTAINTY?

Definition (Merriam-Webster today)

uncertainty: not known beyond doubt, not having certain knowledge, not clearly identified or defined, not constant, indefinite, not certain to occur, not reliable

Definition (Wikipedia today)

Uncertainty refers to *epistemic* situations involving imperfect or unknown information. It applies to predictions of future events, to physical measurements that are already made, or to the unknown. Uncertainty arises in partially observable or stochastic environments, as well as due to ignorance, indolence, or both.
likely from [Russell and Norvig 2010]

...

UNCERTAINTIES IN ML FOR SCIENCE [TAN ET AL. 2023]

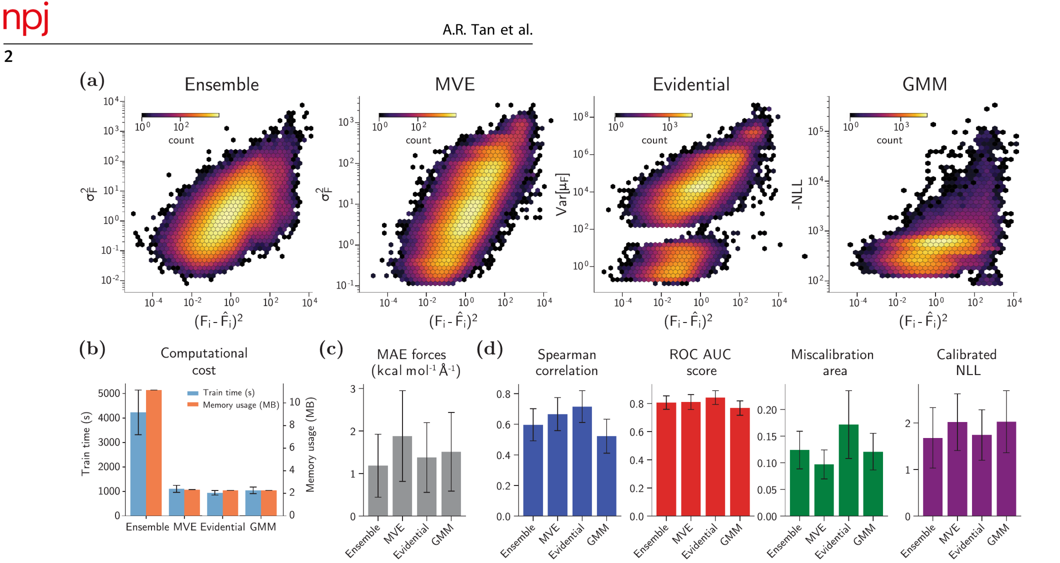


Fig. 1 Comparison of UQ methods on the rMD17 data set. a Hexbin plots showing (predicted) uncertainties versus squared errors of atomic

UNCERTAINTIES: A PARADIGM SHIFT

*Let's say I have data and I fit a linear model to predict y from x. ...So that would be the predictive take. The inferential take would be: can I say something about which features are significant in this (linear) model? Can I actually give under some assumptions **a confidence interval** for their coefficient in the linear model and so on.*

UNCERTAINTIES: A PARADIGM SHIFT

*Let's say I have data and I fit a linear model to predict y from x. ...So that would be the predictive take. The inferential take would be: can I say something about which features are significant in this (linear) model? Can I actually give under some assumptions **a confidence interval** for their coefficient in the linear model and so on.*

*In statistics, we think of these as like two very big areas of research. And maybe historically, actually **inference** has been even bigger than **prediction**. And in machine learning, it's exactly the opposite: So prediction dominates, inference is tiny.*

UNCERTAINTIES: A PARADIGM SHIFT

Let's say I have data and I fit a linear model to predict y from x. …So that would be the predictive take. The inferential take would be: can I say something about which features are significant in this (linear) model? Can I actually give under some assumptions ***a confidence interval*** *for their coefficient in the linear model and so on.*

In statistics, we think of these as like two very big areas of research. And maybe historically, actually ***inference*** *has been even bigger than* ***prediction****. And in machine learning, it's exactly the opposite: So prediction dominates, inference is tiny.*

And historically in machine learning, inference may have been very, very small. And now I think it's grown in a way that people do talk about inference, they use the word ***uncertainty quantification****. They don't think about inference and typically in the traditional way statistically. But I think it has somehow emerged as maybe more of a focus.*

Ryan Tibshirani in "The Gradient Podcast", see here for the full interview.

ORIGINS OF UNCERTAINTY IN ML 1/3

Supervised Machine Learning

- *dataset* $\mathcal{D} = \{(\vec{x}_0, y_0), (\vec{x}_1, y_1), ..., (\vec{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$
  $\mathcal{X}$ ...instance space
  $\mathcal{Y}$ ...outcomes associated with instance

ORIGINS OF UNCERTAINTY IN ML 1/3

Supervised Machine Learning

- *dataset* $\mathcal{D} = \{(\vec{x}_0, y_0), (\vec{x}_1, y_1), ..., (\vec{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$
  $\mathcal{X}$ ...instance space
  $\mathcal{Y}$ ...outcomes associated with instance
- each sample of $\mathcal{D}$ is considered an i.i.d. sample

ORIGINS OF UNCERTAINTY IN ML 1/3

Supervised Machine Learning

- *dataset* $\mathcal{D} = \{(\vec{x}_0, y_0), (\vec{x}_1, y_1), ..., (\vec{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$
  $\mathcal{X}$ ...instance space
  $\mathcal{Y}$ ...outcomes associated with instance
- each sample of $\mathcal{D}$ is considered an i.i.d. sample
- given a *hypothesis* space $\mathcal{H}$ (with $h : \mathcal{X} \rightarrow \mathcal{Y}$) and

ORIGINS OF UNCERTAINTY IN ML 1/3

Supervised Machine Learning

- *dataset* $\mathcal{D} = \{(\vec{x}_0, y_0), (\vec{x}_1, y_1), ..., (\vec{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}$
  $\mathcal{X}$ ...instance space
  $\mathcal{Y}$ ...outcomes associated with instance
- each sample of $\mathcal{D}$ is considered an i.i.d. sample
- given a *hypothesis* space $\mathcal{H}$ (with $h : \mathcal{X} \to \mathcal{Y}$) and
- given a *loss function* $l : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

ORIGINS OF UNCERTAINTY IN ML 2/3

i.i.d. dataset $\mathcal{D}$	hypothesis space	loss function
$\{(\vec{x}_0, y_0), ..., (\vec{x}_N, y_N)\}$	$\mathcal{H}$, $h : \mathcal{X} \rightarrow \mathcal{Y}$	$l = f(h(x), y))$

ORIGINS OF UNCERTAINTY IN ML 2/3

i.i.d. dataset $\mathcal{D}$

$\{(\vec{x}_0, y_0), ..., (\vec{x}_N, y_N)\}$

hypothesis space

$\mathcal{H}, \ h : \mathcal{X} \to \mathcal{Y}$

loss function

$l = f(h(x), y))$

Learner Goal

To induce a hypothesis $h^* \in \mathcal{H}$ with low risk $R(h)$:

$$R(h) := \int_{\mathcal{X}\times\mathcal{Y}} l(h(x), y) dP(x, y)$$

ORIGINS OF UNCERTAINTY IN ML 2/3

i.i.d. dataset $\mathcal{D}$

$$\{(\vec{x}_0, y_0), ..., (\vec{x}_N, y_N)\}$$

hypothesis space

$$\mathcal{H},\ h : \mathcal{X} \rightarrow \mathcal{Y}$$

loss function

$$l = f(h(x), y))$$

Learner Goal

To induce a hypothesis $h^* \in \mathcal{H}$ with low risk $R(h)$:

$$R(h) := \int_{\mathcal{X}\times\mathcal{Y}} l(h(x), y) dP(x, y)$$

Learner Guesses

a good hypothesis h guided by the empirical risk $R_{emp}(h)$:

$$R_{emp} := \frac{1}{N} \sum_{i=1}^{N} l(h(x_i), y_i)$$

ORIGINS OF UNCERTAINTY IN ML 2/3

i.i.d. dataset $\mathcal{D}$

$\{(\vec{x}_0, y_0), ..., (\vec{x}_N, y_N)\}$

hypothesis space

$\mathcal{H},\ h : \mathcal{X} \rightarrow \mathcal{Y}$

loss function

$l = f(h(x), y))$

Learner Goal

To induce a hypothesis $h^* \in \mathcal{H}$ with low risk $R(h)$:

$$R(h) := \int_{\mathcal{X}\times\mathcal{Y}} l(h(x), y) dP(x, y)$$

Learner Guesses

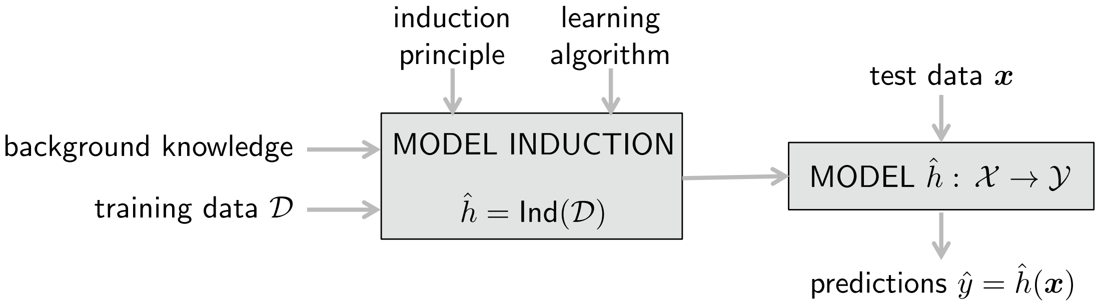
a good hypothesis h guided by the empirical risk $R_{emp}(h)$:

$$R_{emp} := \frac{1}{N}\sum_{i=1}^{N} l(h(x_i), y_i)$$

R_{emp} only estimates R!

- $h^* := \arg\min_{h\in\mathcal{H}} R(h)$ will not coincide with $\hat{h} := \arg\min_{h\in\mathcal{H}} R_{emp}(h)$
- **uncertainy created!** what is h^*? How close is $\hat{h}$ to h^*? What is $R(\hat{h})$?

WRAP-UP ORIGINS OF UNCERTAINTY



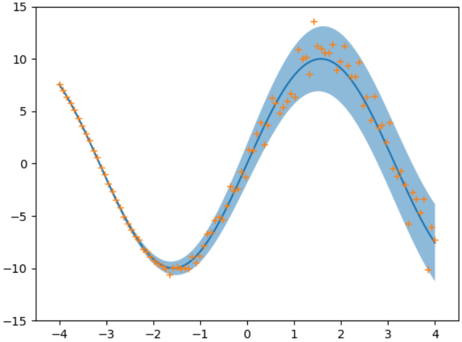
Central Point of Interest: **Predictive Uncertainties**

(uncertainty for $y_q = \hat{h}(x_q)$ for a concrete instance $x_q \in \mathcal{X}$)

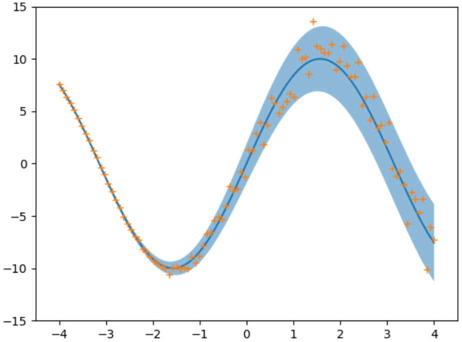
[Hüllermeier and Waegeman 2021]

UQ FOR REGRESSION

A HISTORIC VIEW ON UNCERTAINTY SOURCES



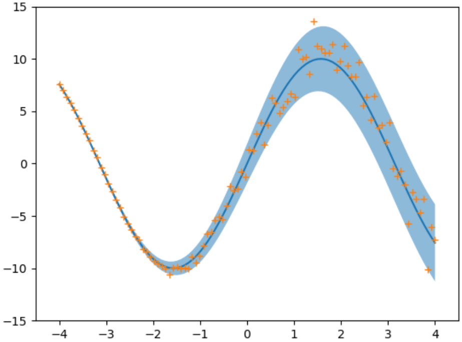
A HISTORIC VIEW ON UNCERTAINTY SOURCES



- **Aleatoric** or
  Data related **uncertainty**

  (uncertainties from noise in the data, does not decrease with more data)

A HISTORIC VIEW ON UNCERTAINTY SOURCES



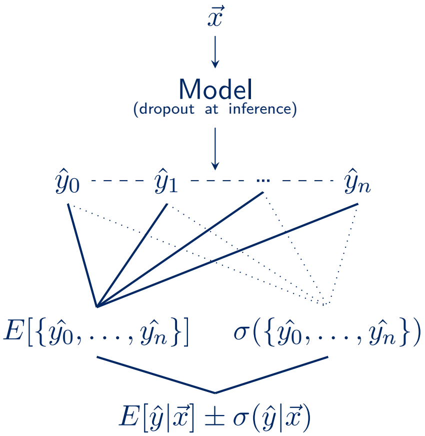
- **Aleatoric** or
  Data related **uncertainty**

  (uncertainties from noise in the data, does not decrease with more data)
- **Epistemic** or
  Model related **uncertainty**

  (uncertainties related to finding the best hypothesis, can be reduced with more data)

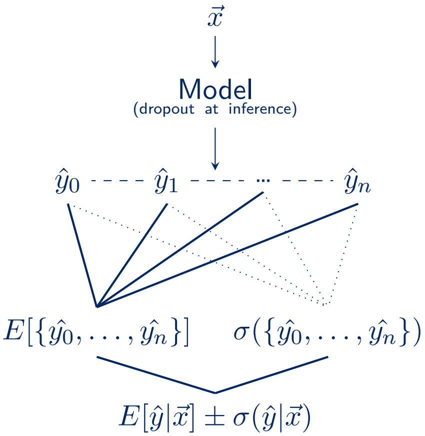
OBTAINING (EPISTEMIC) UNCERTAINTIES

Monte Carlo Dropout



OBTAINING (EPISTEMIC) UNCERTAINTIES

Monte Carlo Dropout

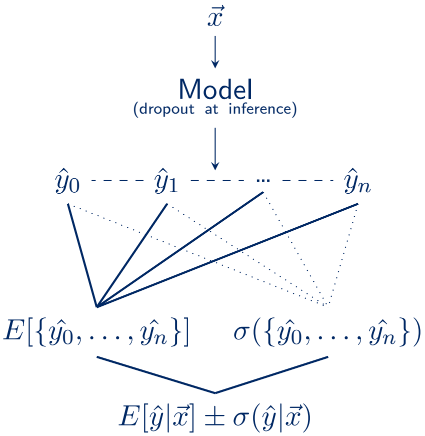


[Gal and Ghahramani 2016]

- model set up including dropout layers
- dropout layers set random portions of weights to 0.
  (implicit regularisation during training)

OBTAINING (EPISTEMIC) UNCERTAINTIES

Monte Carlo Dropout

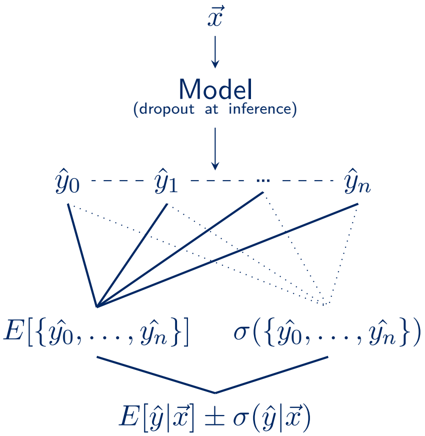


[Gal and Ghahramani 2016]

- model set up including dropout layers
- dropout layers set random portions of weights to 0.
  (implicit regularisation during training)
- **trick:** keep dropout enabled during prediction and call `model.predict` multiple times

OBTAINING (EPISTEMIC) UNCERTAINTIES

Monte Carlo Dropout

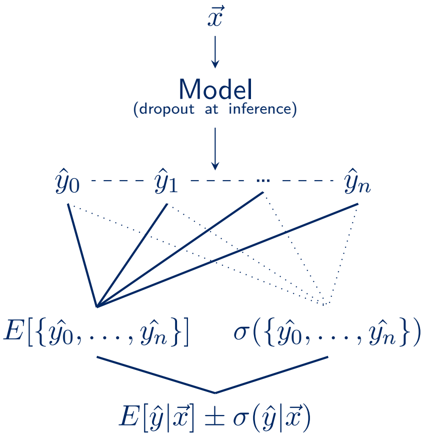


[Gal and Ghahramani 2016]

- model set up including dropout layers
- dropout layers set random portions of weights to 0.
  (implicit regularisation during training)
- **trick:** keep dropout enabled during prediction and call `model.predict` multiple times
- allows to sample the PDF describing $\hat{y}$
  (weights of model disabled randomly)

OBTAINING (EPISTEMIC) UNCERTAINTIES

Monte Carlo Dropout



[Gal and Ghahramani 2016]

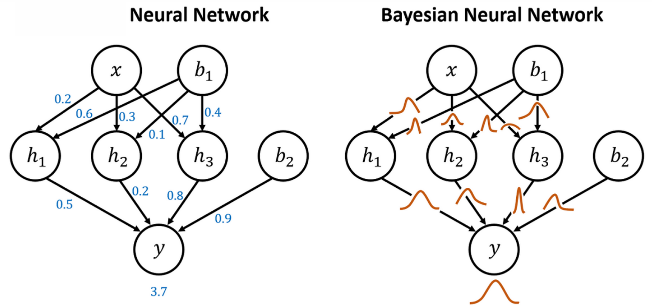
- model set up including dropout layers
- dropout layers set random portions of weights to 0.
  (implicit regularisation during training)
- **trick:** keep dropout enabled during prediction and call `model.predict` multiple times
- allows to sample the PDF describing $\hat{y}$ (weights of model disabled randomly)
- obtain mean prediction $E[\hat{y}]$ and prediction deviation $\sigma(\hat{y})$ for each input

TRY MCDROPOUT FOR YOURSELF

Please open `01_mcdropout_1D_regression.ipynb`!

Go through the notebook.

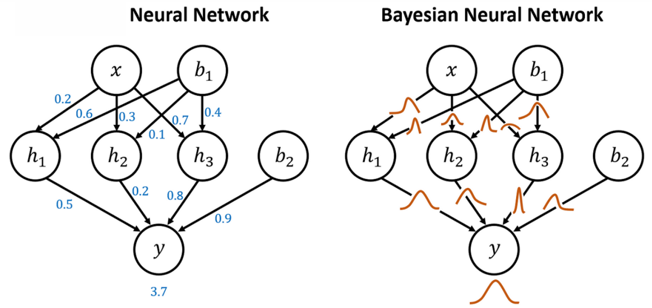
DROPOUT RECAP



from jonascleveland.com

- MCDropout approximates a Bayesian Neural Network (BNNs are computational intensive)

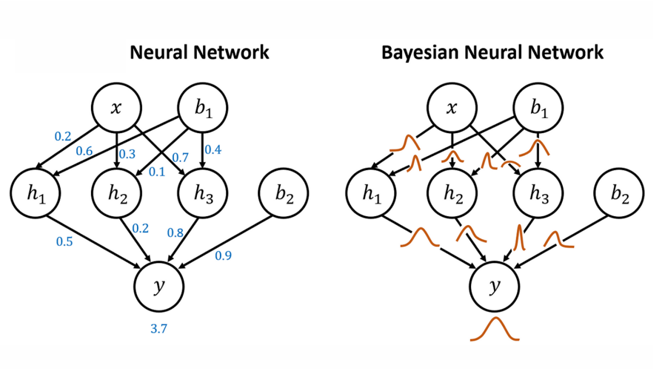
DROPOUT RECAP



from jonascleveland.com

- MCDropout approximates a Bayesian Neural Network (BNNs are computational intensive)
- average across possible network configurations (enable Dropout layers during inference)

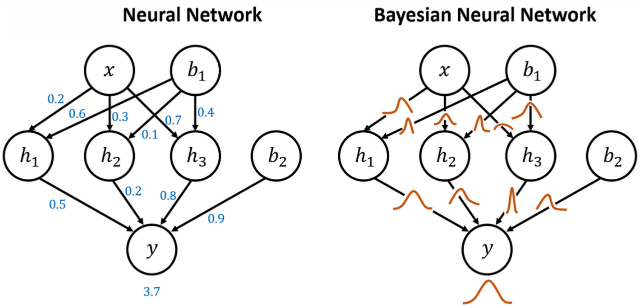
DROPOUT RECAP



from jonascleveland.com

- MCDropout approximates a Bayesian Neural Network
  (BNNs are computational intensive)
- average across possible network configurations
  (enable Dropout layers during inference)
- perform Bayesian Model Averaging
  see [Gawlikowski et al. 2022] for details

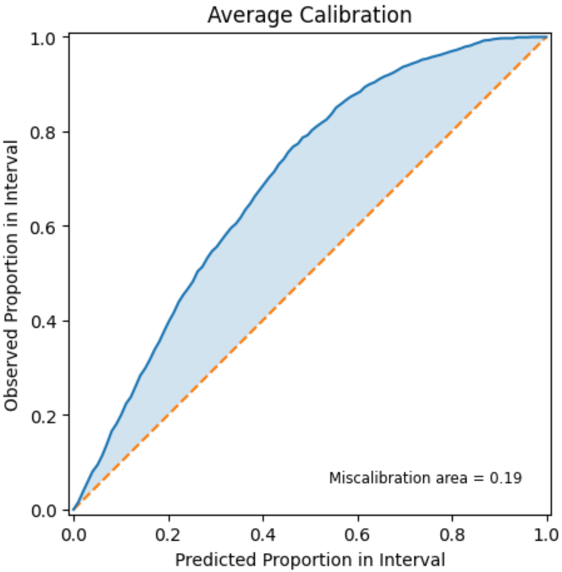
DROPOUT RECAP



from jonascleveland.com

- MCDropout approximates a Bayesian Neural Network (BNNs are computational intensive)
- average across possible network configurations (enable Dropout layers during inference)
- perform Bayesian Model Averaging see [Gawlikowski et al. 2022] for details
- **epistemic uncertainties!**

DIAGNOSING A UQ METHOD: CALIBRATION CURVES



HOW TO JUDGE PREDICTIVE UNCERTAINTIES?

At Inference, we have

- a label y_{test}
- a prediction $E[y_{test}]$ (by our model)
- an uncertainty for that prediction σy_{test} (from the UQ method)

HOW TO JUDGE PREDICTIVE UNCERTAINTIES?

At Inference, we have

- a label y_{test}
- a prediction $E[y_{test}]$ (by our model)
- an uncertainty for that prediction σy_{test} (from the UQ method)

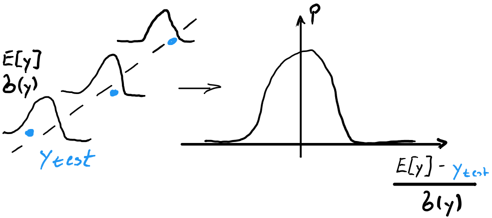
Assumption

- $E[y_{test}]$ and σy_{test} model a gaussian distribution around y_{test}

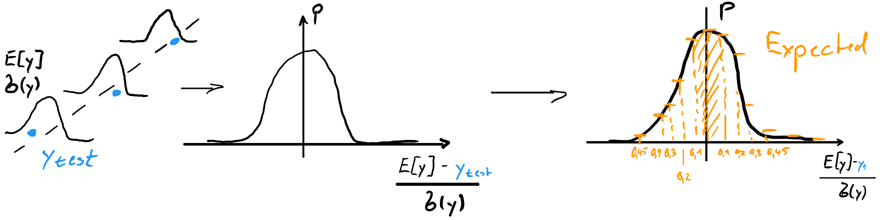
[Kuleshov, Fenner, and Ermon 2018]

CALIBRATION IN A NUTSHELL [KULESHOV, FENNER, AND ERMON 2018]

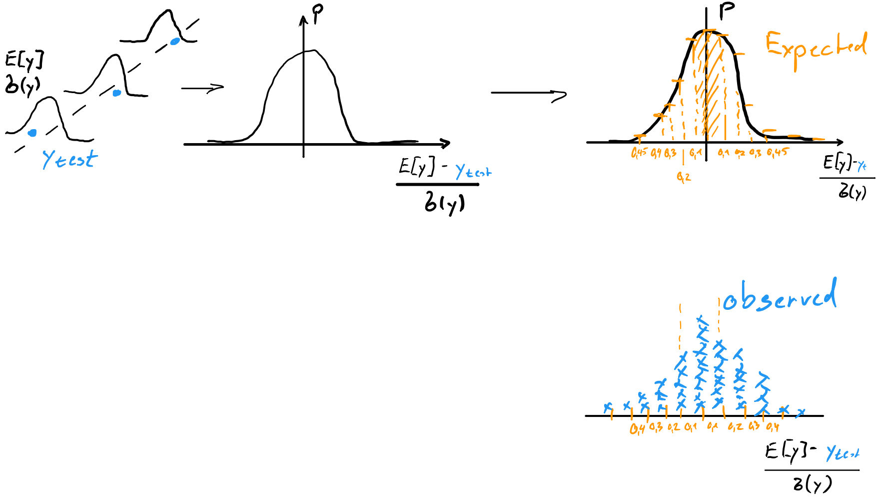
CALIBRATION IN A NUTSHELL [KULESHOV, FENNER, AND ERMON 2018]



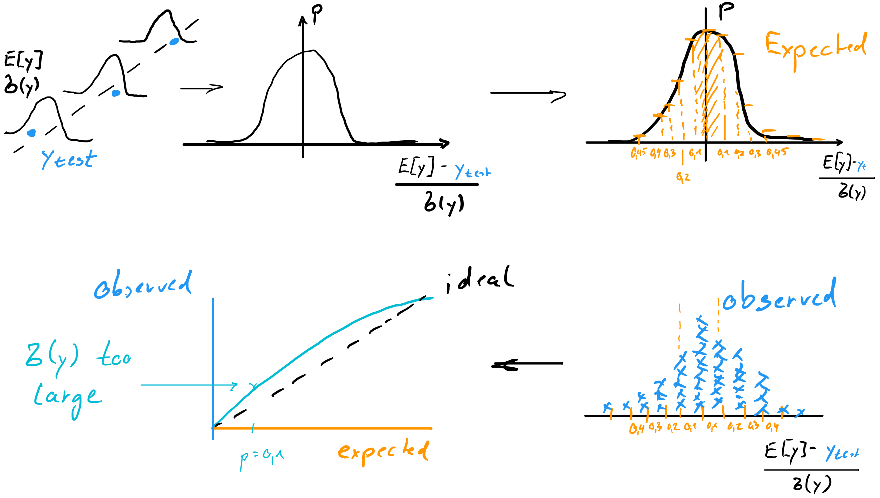
CALIBRATION IN A NUTSHELL [KULESHOV, FENNER, AND ERMON 2018]



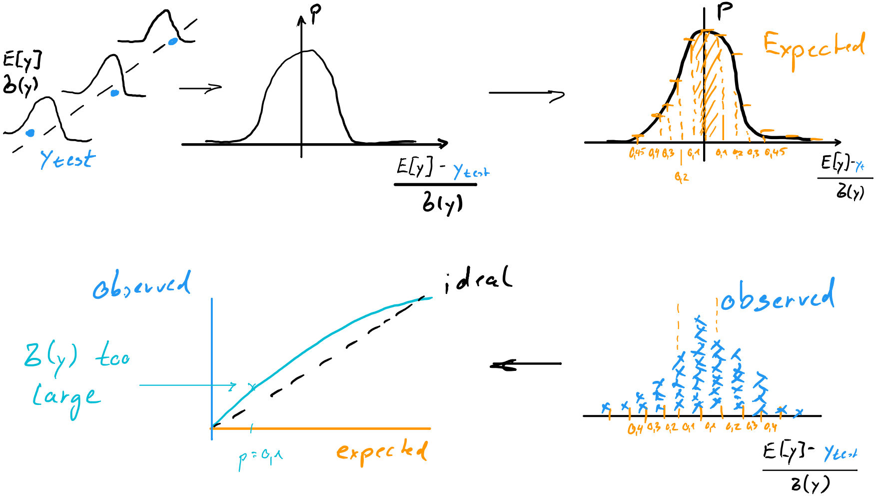
CALIBRATION IN A NUTSHELL [KULESHOV, FENNER, AND ERMON 2018]



CALIBRATION IN A NUTSHELL [KULESHOV, FENNER, AND ERMON 2018]



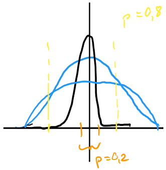
CALIBRATION IN A NUTSHELL [KULESHOV, FENNER, AND ERMON 2018]



Note: **only sufficient criteria** for uncertainty quality [Levi et al. 2020]!

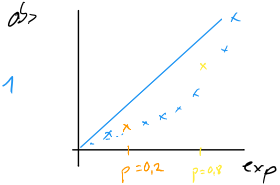
CALIBRATION PATHOLOGIES [KULESHOV, FENNER, AND ERMON 2018]

Overconfident: $\sigma(y)$ is too small

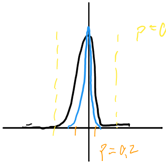


obs: $z_{eff} \gg 1$

$\frac{y - E(y)}{\sigma(y)}$

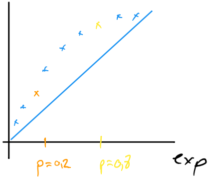


Underconfident: $\sigma(y)$ is too large



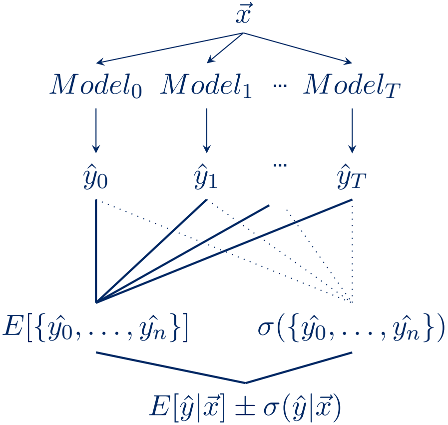
obs: $z_{eff} \ll 1$

$\frac{y - E(y)}{\sigma(y)}$



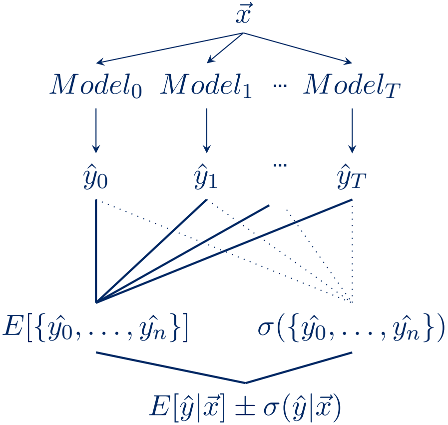
OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble?



OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble?

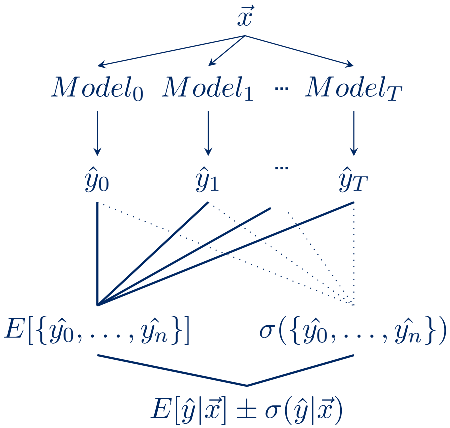


[Lakshminarayanan, Pritzel, and Blundell 2017]

- use existing model and training loop
- train T models (different seeds)

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble?

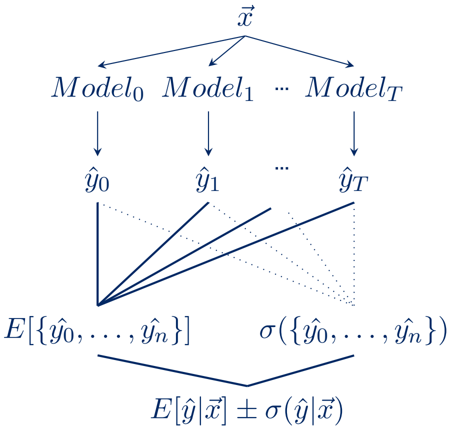


[Lakshminarayanan, Pritzel, and Blundell 2017]

- use existing model and training loop
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction from each model

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble?

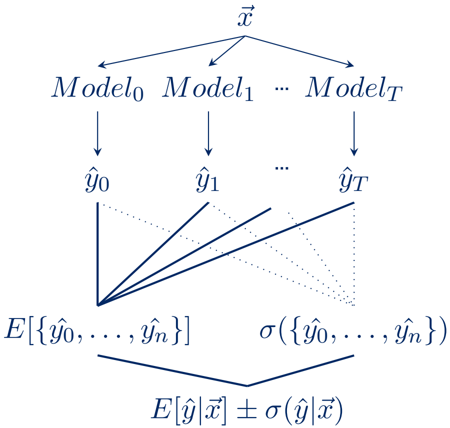


[Lakshminarayanan, Pritzel, and Blundell 2017]

- use existing model and training loop
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction from each model
- obtain mean prediction $E[\hat{y}]$ and prediction deviation $\sigma(\hat{y})$

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble?

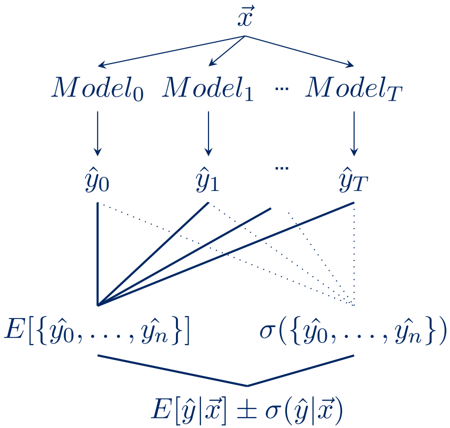


[Lakshminarayanan, Pritzel, and Blundell 2017]

- use existing model and training loop
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction from each model
- obtain mean prediction $E[\hat{y}]$ and prediction deviation $\sigma(\hat{y})$
- **epistemic uncertainty!**

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble?

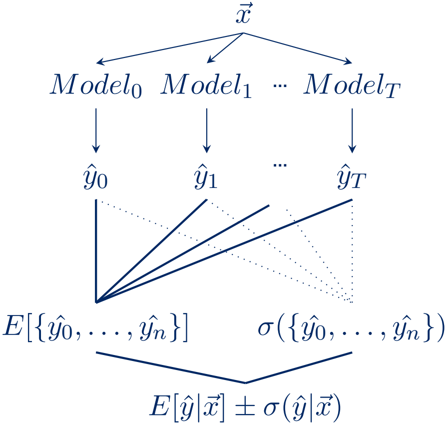


[Lakshminarayanan, Pritzel, and Blundell 2017]

- use existing model and training loop
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction from each model
- obtain mean prediction $E[\hat{y}]$ and prediction deviation $\sigma(\hat{y})$
- **epistemic uncertainty!**

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble?



[Lakshminarayanan, Pritzel, and Blundell 2017]

- use existing model and training loop
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction from each model
- obtain mean prediction $E[\hat{y}]$ and prediction deviation $\sigma(\hat{y})$
- **epistemic uncertainty!**

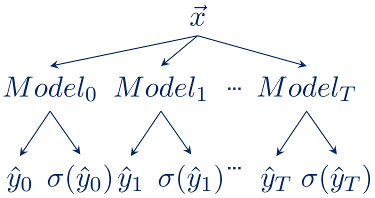
Our nickname: “Simple Ensembles”!

TRY SIMPLE ENSEMBLES FOR YOURSELF

Please open `02_simpleensembles_1D_regression.ipynb`!
Go through the notebook.

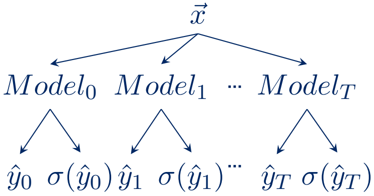
OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble!



OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble!



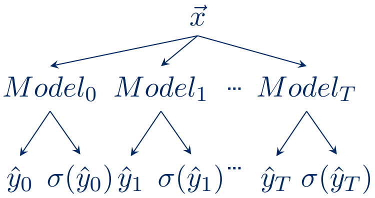
$$l_{nll} = -\log p = \frac{\log \sigma(\hat{y})^2}{2} + \frac{1}{2}\frac{(y-\hat{y})^2}{\sigma(\hat{y})^2}$$

$$E_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} \hat{y}_t$$

$$\sigma_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} (\sigma(\hat{y}_t)^2 + \hat{y}_t^2) - E_{ens}[y_n]^2$$

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble!



- use expanded model to predict expectation and std deviation (mean variance estimation, MVE)
- train T models (different seeds)

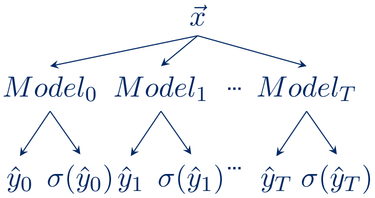
$$l_{nll} = -\log p = \frac{\log \sigma(\hat{y})^2}{2} + \frac{1}{2}\frac{(y-\hat{y})^2}{\sigma(\hat{y})^2}$$

$$E_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T}\hat{y}_t$$

$$\sigma_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T}(\sigma(\hat{y}_t)^2 + \hat{y}_t^2) - E_{ens}[y_n]^2$$

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble!



$$l_{nll} = -\log p = \frac{\log \sigma(\hat{y})^2}{2} + \frac{1}{2}\frac{(y-\hat{y})^2}{\sigma(\hat{y})^2}$$

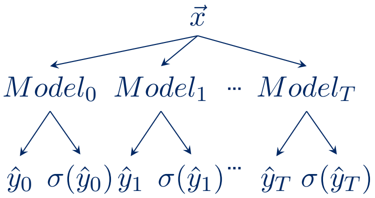
$$E_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} \hat{y}_t$$

$$\sigma_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T}(\sigma(\hat{y}_t)^2 + \hat{y}_t^2) - E_{ens}[y_n]^2$$

- use expanded model to predict expectation and std deviation (mean variance estimation, MVE)
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction and one sigma from each model

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble!



$$l_{nll} = -\log p = \frac{\log \sigma(\hat{y})^2}{2} + \frac{1}{2}\frac{(y-\hat{y})^2}{\sigma(\hat{y})^2}$$

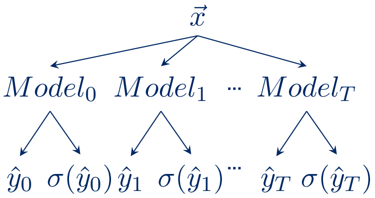
$$E_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} \hat{y}_t$$

$$\sigma_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T}(\sigma(\hat{y}_t)^2 + \hat{y}_t^2) - E_{ens}[y_n]^2$$

- use expanded model to predict expectation and std deviation (mean variance estimation, MVE)
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction and one sigma from each model
- allows to learn model of **aleatoric uncertainty** (MVE)

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble!



$$l_{nll} = -\log p = \frac{\log \sigma(\hat{y})^2}{2} + \frac{1}{2}\frac{(y-\hat{y})^2}{\sigma(\hat{y})^2}$$

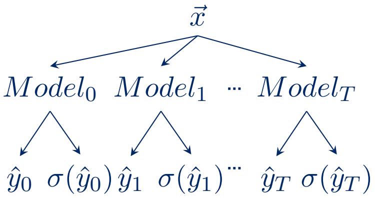
$$E_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} \hat{y}_t$$

$$\sigma_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T}(\sigma(\hat{y}_t)^2 + \hat{y}_t^2) - E_{ens}[y_n]^2$$

- use expanded model to predict expectation and std deviation (mean variance estimation, MVE)
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction and one sigma from each model
- allows to learn model of **aleatoric uncertainty** (MVE)
- ensemble allows to assess **epistemic uncertainty**

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble!



$$l_{nll} = -\log p = \frac{\log \sigma(\hat{y})^2}{2} + \frac{1}{2}\frac{(y-\hat{y})^2}{\sigma(\hat{y})^2}$$

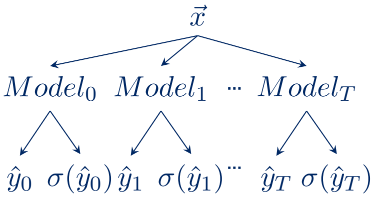
$$E_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} \hat{y}_t$$

$$\sigma_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} (\sigma(\hat{y}_t)^2 + \hat{y}_t^2) - E_{ens}[y_n]^2$$

- use expanded model to predict expectation and std deviation (mean variance estimation, MVE)
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction and one sigma from each model
- allows to learn model of **aleatoric uncertainty** (MVE)
- ensemble allows to assess **epistemic uncertainty**
- but: different loss, different model

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble!



$$l_{nll} = -\log p = \frac{\log \sigma(\hat{y})^2}{2} + \frac{1}{2}\frac{(y-\hat{y})^2}{\sigma(\hat{y})^2}$$

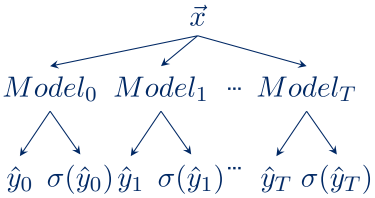
$$E_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} \hat{y}_t$$

$$\sigma_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} (\sigma(\hat{y}_t)^2 + \hat{y}_t^2) - E_{ens}[y_n]^2$$

- use expanded model to predict expectation and std deviation (mean variance estimation, MVE)
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction and one sigma from each model
- allows to learn model of **aleatoric uncertainty** (MVE)
- ensemble allows to assess **epistemic uncertainty**
- but: different loss, different model

OBTAINING UNCERTAINTIES IN ENSEMBLES

Deep Ensemble!



$$l_{nll} = -\log p = \frac{\log \sigma(\hat{y})^2}{2} + \frac{1}{2}\frac{(y-\hat{y})^2}{\sigma(\hat{y})^2}$$

$$E_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} \hat{y}_t$$

$$\sigma_{ens}[y_n] = \frac{1}{T}\sum_{t=0}^{T} (\sigma(\hat{y}_t)^2 + \hat{y}_t^2) - E_{ens}[y_n]^2$$

- use expanded model to predict expectation and std deviation (mean variance estimation, MVE)
- train T models (different seeds)
- **trick:** for each $\vec{x}$ get one prediction and one sigma from each model
- allows to learn model of **aleatoric uncertainty** (MVE)
- ensemble allows to assess **epistemic uncertainty**
- but: different loss, different model

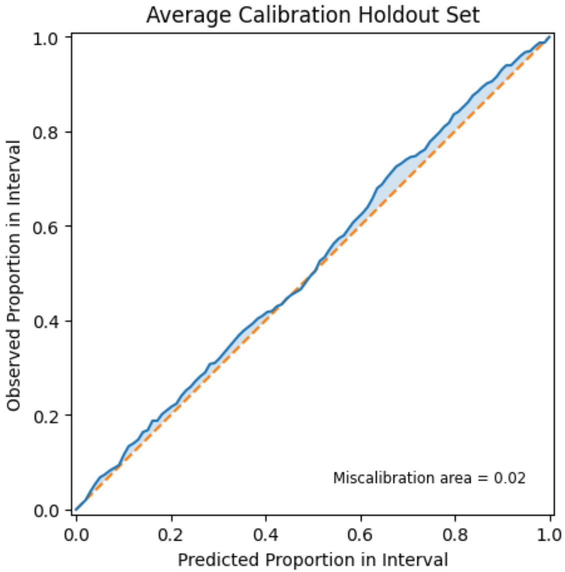
These are “Deep Ensembles”!

[Lakshminarayanan, Pritzel, and Blundell 2017]

TRY DEEP ENSEMBLES FOR YOURSELF

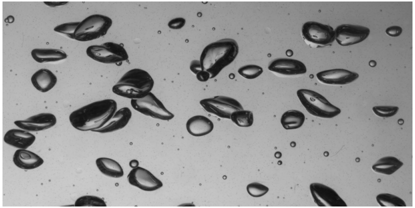
Please open
`03_deepensembles_1D_regression.ipynb`!
Go through the notebook.

WELL CALIBRATED UNCERTAINTIES WITH DEEPENSEMBLES



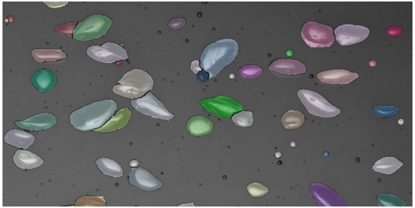
INSTANCE SEGMENTATION TASK

Input



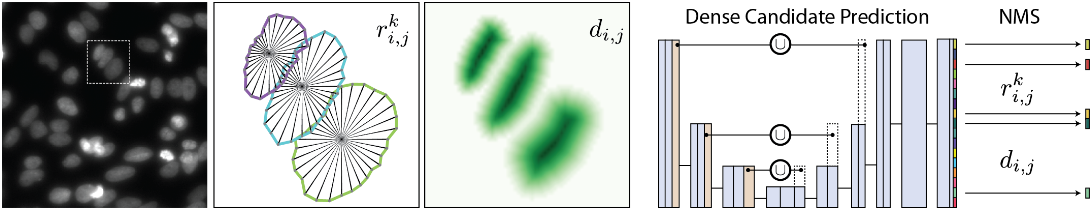
[Hessenkemper et al. 2022]

Labels



- goal: accurate spatial prediction
- adding uncertainty = reliable and robust prediction

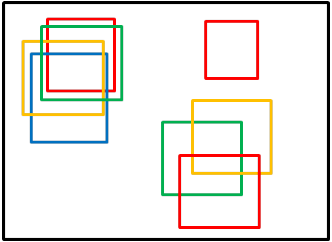
INSTANCE SEGMENTATION TOOLING: STARDIST



UQ FOR STARDIST [SIDDIQUI, STARKE, AND STEINBACH 2023]

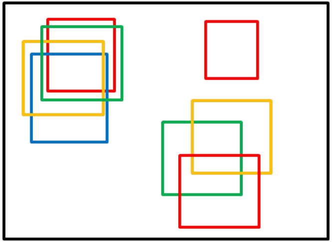
ensemble predictions

provide multitude of labels

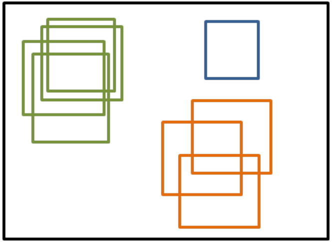


UQ FOR STARDIST [SIDDIQUI, STARKE, AND STEINBACH 2023]

ensemble predictions provide multitude of labels

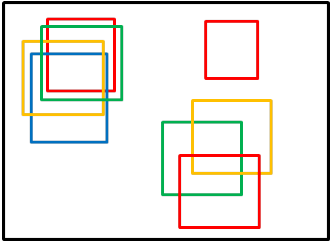


clustering for homogenous instance labels

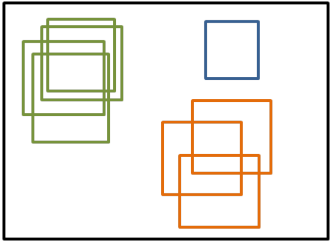


UQ FOR STARDIST [SIDDIQUI, STARKE, AND STEINBACH 2023]

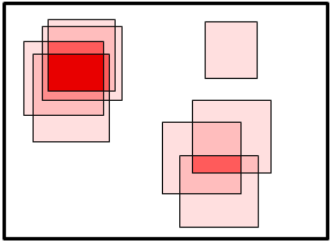
ensemble predictions provide multitude of labels



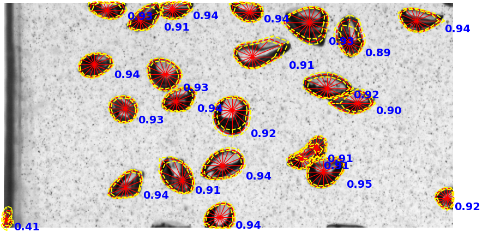
clustering for homogenous instance labels



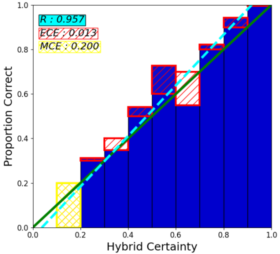
calibrated **certainty scores** by region of most overlap



INFORMED PREDICTIONS WITH UNCERTAINTIES AND CALIBRATION PLOTS

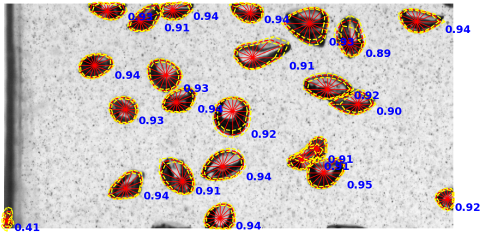


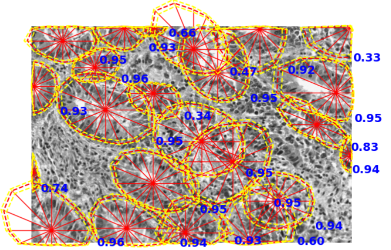
benign segmentation
(bubble segmentation)



[Siddiqui, Starke, and Steinbach 2023]

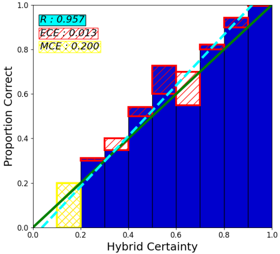
INFORMED PREDICTIONS WITH UNCERTAINTIES AND CALIBRATION PLOTS

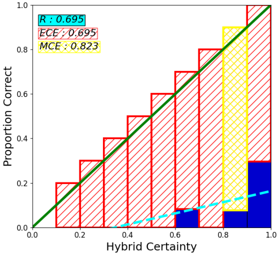




benign segmentation
(bubble segmentation)

malignant segmentation
(gland tumor cell segmentation)

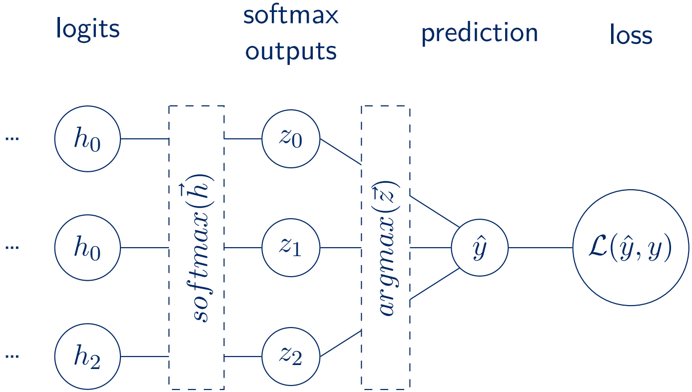




[Siddiqui, Starke, and Steinbach 2023]

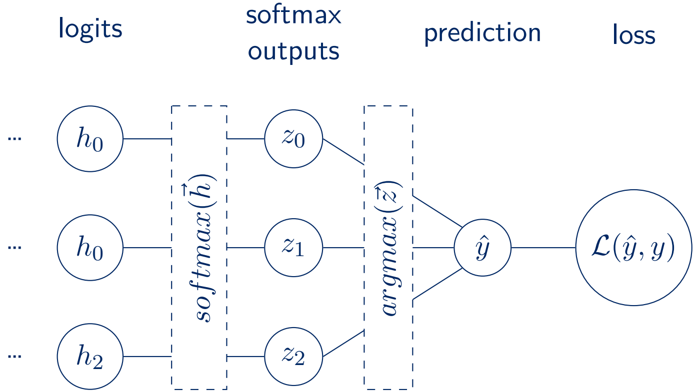
UQ FOR CLASSIFICATION

RECAP: NEURAL NETWORK CLASSIFICATION



- classifaction prediction obtained from $argmax$
- how to express uncertainty for a class label?

RECAP: NEURAL NETWORK CLASSIFICATION



- classifaction prediction obtained from $argmax$
- how to express uncertainty for a class label?

- classification = often used as goto benchmark for UQ methods
- distinct differences to regression

CLASSIFICATION RATIONAL FOR UQ

Prediction by NN

Given inputs $\vec{x}$, a fitted model $\hat{h}$ provides two outputs:

$$\hat{h}(\vec{x}) = (\hat{y}, \hat{p})$$

class prediction $\hat{y}$ and confidence $\hat{p}$.

CLASSIFICATION RATIONAL FOR UQ

Prediction by NN

Given inputs $\vec{x}$, a fitted model $\hat{h}$ provides two outputs:

$$\hat{h}(\vec{x}) = (\hat{y}, \hat{p})$$

class prediction $\hat{y}$ and confidence $\hat{p}$.

Goal

Confidence p (probability of correctness) should be calibrated!
In other words, p should be a true probability.

CLASSIFICATION RATIONAL FOR UQ

Prediction by NN

Given inputs $\vec{x}$, a fitted model $\hat{h}$ provides two outputs:

$$\hat{h}(\vec{x}) = (\hat{y}, \hat{p})$$

class prediction $\hat{y}$ and confidence $\hat{p}$.

Goal

Confidence p (probability of correctness) should be calibrated!

In other words, p should be a true probability.

Intuition

Given 100 predictions, each with confidence $p = 0.8$. We expect $80/100$ to be correctly classified.

For more details, see [Guo et al. 2017]

RELIABILITY DIAGRAMS [GUO ET AL. 2017]

Plot expected sample accuracy versus (observed) confidence

1. run model on (validation) data and obtain predictions $(\hat{y}, \hat{p})$
2. bin predictions into M interval bins (bin width of $1/M$) let $\mathcal{B}_m$ be all sample indices that fall into bin m
3. calculate accuracy for samples in each bin

$$acc(\mathcal{B}_m) = \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} \mathbf{1}(\hat{y}_i = y_i)$$

4. calculate average confidence in each bin

$$conf(\mathcal{B}_m) = \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} \hat{p}_i$$

RELIABILITY DIAGRAMS [GUO ET AL. 2017]

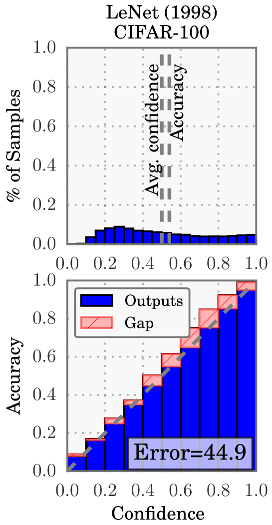
Plot expected sample accuracy versus (observed) confidence

1. run model on (validation) data and obtain predictions $(\hat{y}, \hat{p})$
2. bin predictions into M interval bins (bin width of $1/M$) let $\mathcal{B}_m$ be all sample indices that fall into bin m
3. calculate accuracy for samples in each bin

$$acc(\mathcal{B}_m) = \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} \mathbf{1}(\hat{y}_i = y_i)$$

4. calculate average confidence in each bin

$$conf(\mathcal{B}_m) = \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} \hat{p}_i$$



CONCLUSIONS

SUMMARY

- **uncertainties express variability** in Machine Learning predictions (induction, limits of dataset)

SUMMARY

- **uncertainties express variability** in Machine Learning predictions (induction, limits of dataset)
- uncertainties should become **first class citizens** in Machine Learning

SUMMARY

- **uncertainties express variability** in Machine Learning predictions (induction, limits of dataset)
- uncertainties should become **first class citizens** in Machine Learning
- established methods for obtaining **predictive uncertainties** (MCDropout, SimpleEnsembles, DeepEnsembles)

SUMMARY

- **uncertainties express variability** in Machine Learning predictions (induction, limits of dataset)
- uncertainties should become **first class citizens** in Machine Learning
- established methods for obtaining **predictive uncertainties** (MCDropout, SimpleEnsembles, DeepEnsembles)
- depending on UQ method, uncertainties refer to different things (epistemic vs aleatoric uncertainties for example)

SUMMARY

- **uncertainties express variability** in Machine Learning predictions (induction, limits of dataset)
- uncertainties should become **first class citizens** in Machine Learning
- established methods for obtaining **predictive uncertainties** (MCDropout, SimpleEnsembles, DeepEnsembles)
- depending on UQ method, uncertainties refer to different things (epistemic vs aleatoric uncertainties for example)

SUMMARY

- **uncertainties express variability** in Machine Learning predictions (induction, limits of dataset)
- uncertainties should become **first class citizens** in Machine Learning
- established methods for obtaining **predictive uncertainties** (MCDropout, SimpleEnsembles, DeepEnsembles)
- depending on UQ method, uncertainties refer to different things (epistemic vs aleatoric uncertainties for example)

Thank you for your attention!
Looking forward to questions, feedback and comments.

BIBLIOGRAPHY I

- Bouthillier, Xavier et al. (2021). "Accounting for Variance in Machine Learning Benchmarks". In: *CoRR* abs/2103.03098. arXiv: 2103.03098. URL: https://arxiv.org/abs/2103.03098 (cit. on p. 100).
- Gal, Yarin and Zoubin Ghahramani (2016). "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR, pp. 1050–1059 (cit. on pp. 24–28).
- Gawlikowski, Jakob et al. (Jan. 18, 2022). *A Survey of Uncertainty in Deep Neural Networks*. DOI: 10.48550/arXiv.2107.03342. arXiv: 2107.03342[cs,stat]. URL: http://arxiv.org/abs/2107.03342 (visited on 06/15/2023) (cit. on pp. 30–33).
- Guo, Chuan et al. (2017). "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html (cit. on pp. 73–77).
- Hessenkemper, Hendrik et al. (2022). "Bubble identification from images with machine learning methods". In: *International Journal of Multiphase Flow* 155, p. 104169 (cit. on p. 63).

BIBLIOGRAPHY II

Howard, Jeremy et al. (2022). *Imagenette*. Commit at time of writing. URL: https://github.com/fastai/imagenette (cit. on pp. 94, 95).

Hüllermeier, Eyke and Willem Waegeman (Mar. 1, 2021). "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods". In: *Machine Learning* 110.3, pp. 457–506. ISSN: 1573-0565. DOI: 10.1007/s10994-021-05946-3. URL: https://doi.org/10.1007/s10994-021-05946-3 (visited on 11/08/2023) (cit. on p. 19).

Kuleshov, Volodymyr, Nathan Fenner, and Stefano Ermon (2018). "Accurate Uncertainties for Deep Learning Using Calibrated Regression". In: *CoRR* abs/1807.00263. arXiv: 1807.00263. URL: http://arxiv.org/abs/1807.00263 (cit. on pp. 35–43).

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (2017). "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30 (cit. on pp. 44–50, 52–60).

BIBLIOGRAPHY III

- Levi, Dan et al. (2020). *Evaluating and Calibrating Uncertainty Prediction in Regression Tasks*. arXiv: 1905.11659 [cs.LG]. URL: https://arxiv.org/abs/1905.11659 (cit. on p. 42).
- Liu, Zhuang et al. (2022). "A ConvNet for the 2020s". In: *CoRR* abs/2201.03545. arXiv: 2201.03545. URL: https://arxiv.org/abs/2201.03545 (cit. on p. 93).
- Park, Namuk and Songkuk Kim (2022). *How Do Vision Transformers Work?* arXiv: 2202.06709. URL: https://arxiv.org/abs/2202.06709 (cit. on pp. 98, 99).
- Raschka, Sebastian (2018). "Model evaluation, model selection, and algorithm selection in machine learning". In: arXiv: 1811.12808. URL: http://arxiv.org/abs/2103.03098 (cit. on p. 97).
- Rohatgi, Ankit (2021). *Webplotdigitizer: Version 4.5*. URL: https://automeris.io/WebPlotDigitizer (cit. on pp. 98, 99).
- Russell, Stuart J and Peter Norvig (2010). *Artificial intelligence a modern approach*. London (cit. on pp. 4–6).

BIBLIOGRAPHY IV

Siddiqui, Qasim M. K., Sebastian Starke, and Peter Steinbach (2023). *Uncertainty Estimation in Instance Segmentation with Star-convex Shapes*. in review at WACV'24. arXiv: 2309.10513 [cs.CV] (cit. on pp. 65–69).

Steinbach, Peter et al. (2022). "Machine learning state-of-the-art with uncertainties". In: *ICML* ML Evaluation Standards workshop. URL: https://ml-eval.github.io/accepted-papers/#11 (cit. on pp. 94–99, 101–105).

Tan, Aik Rui et al. (Dec. 16, 2023). "Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles". In: *npj Computational Materials* 9.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–11. ISSN: 2057-3960. DOI: 10.1038/s41524-023-01180-8. URL: https://www.nature.com/articles/s41524-023-01180-8 (visited on 01/18/2024) (cit. on p. 7).

BIBLIOGRAPHY V

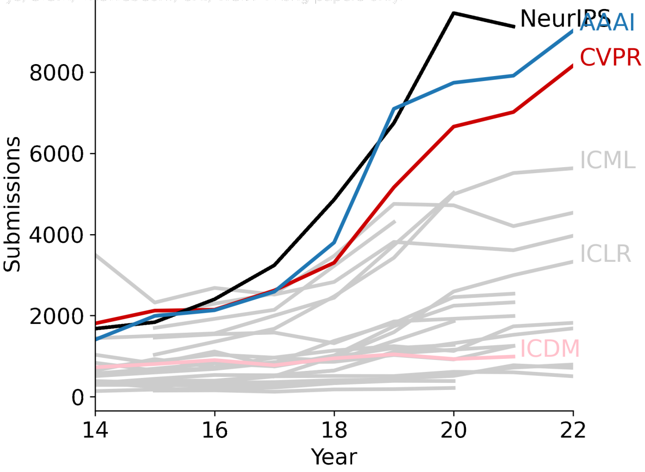
Tomasev, Nenad et al. (2022). "Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?" In: *CoRR* abs/2201.05119. arXiv: 2201.05119. URL: https://arxiv.org/abs/2201.05119 (cit. on p. 93).

Xin, Li (2022). *Statistics of acceptance rate for the main AI conferences*. Commit at time of writing. URL: https://github.com/lixin4ever/Conference-Acceptance-Rate (cit. on p. 92).

APPENDIX

DERIVED UNCERTAINTIES

HIGH DEMAND FOR REVIEWING

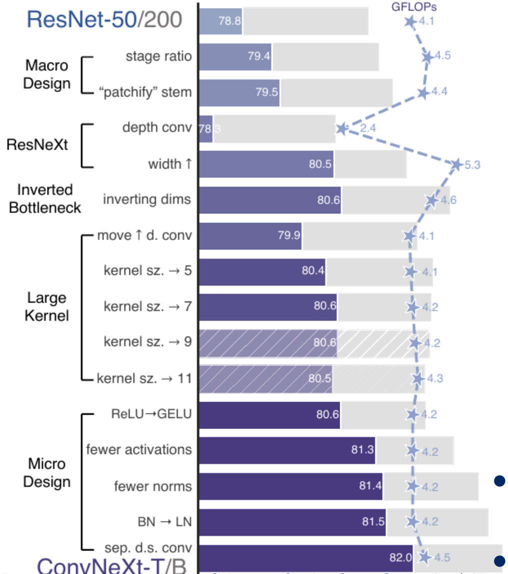


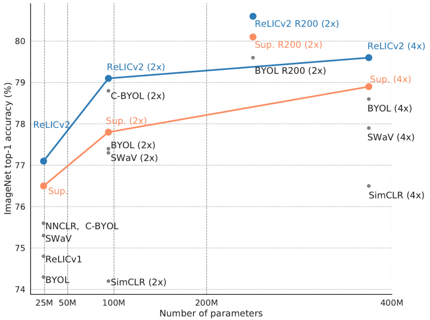
adapted from [Xin 2022]

International Conference on Learning Representations 2022 (2023)

- accepted/submitted $1095/3328$ $(1574/4956)$
- acceptance rate 32.9% (32.0%)
- 54 (91) orals
- 176 (280) spotlights
- 865 (1203) posters

STATE-OF-THE-ART, SOTA





from [Tomasev et al. 2022]

- SOTA = (uncertified) reference to check for progress
- accuracy [redacted] figure of merit

A CLASSIFICATION SOTA FOR DEMONSTRATION

image classification on imagenette [Howard et al. 2022]

...

A CLASSIFICATION SOTA FOR DEMONSTRATION

image classification on imagenette [Howard et al. 2022]

...

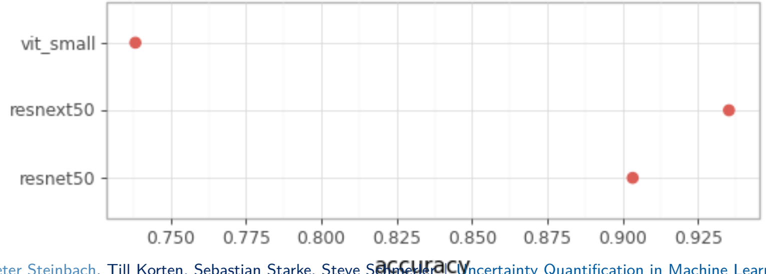


Figure 1 (a): Accuracy estimates on 10-class image classification for three different ML architectures. Taken from [Steinbach et al. 2022]

ACCURACIES WITH UNCERTAINTIES FROM CROSS-VALIDATION

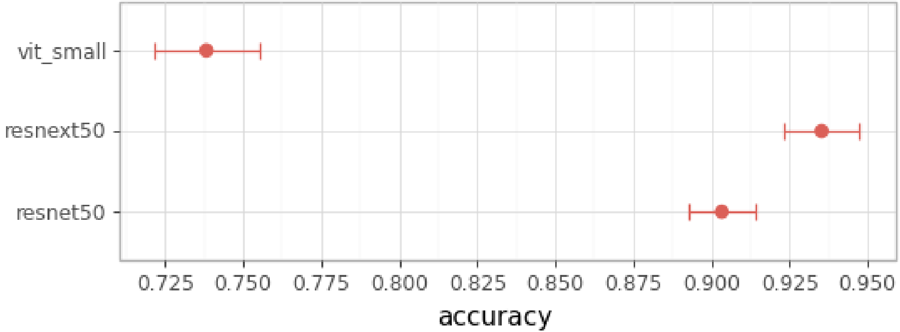


Figure 1 (b): Accuracy estimates on 10-class image classification for three different ML architectures. Point estimates and confidence intervals obtained from 20-fold cross validation is shown. Taken from [Steinbach et al. 2022]

APPROXIMATED UNCERTAINTIES $\hat{\sigma}$

Approximate Accuracy as a Bernoulli probability

$$\mu_{\mathrm{ACC}} \pm \hat{\sigma}_{\mathrm{ACC}} = \mu_{\mathrm{ACC}} \pm z \sqrt{\frac{1}{n_{holdout}} \mathrm{ACC}_{holdout} \left(1 - \mathrm{ACC}_{holdout}\right)}$$

In the limit of large numbers, this converges to a normal distribution. Use z to construct confidence interval assuming normality.

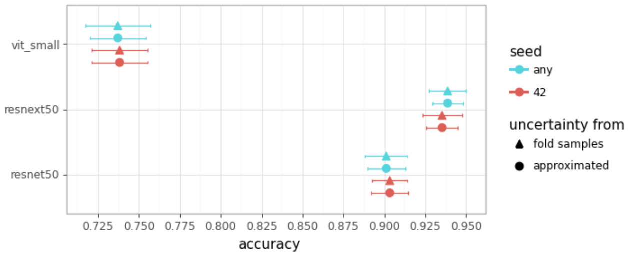
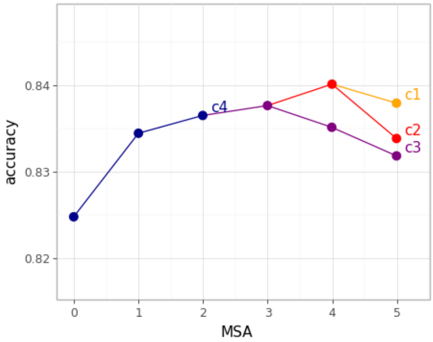


Figure 5: Comparison of fold sample based uncertainty with approximated uncertainty using eq. (1) [Raschka 2018]. Each estimate was obtained for one seed (42) or any seed available (total 6 seeds). The uncertainty plotted for seed 42 was obtained using the approximation in eq. (1). The uncertainty plotted for all seeds was obtained using the sample standard deviation. Taken from [Steinbach et al. 2022]

HOW DO VISION TRANSFORMERS WORK? [PARK AND KIM 2022]



HOW DO VISION TRANSFORMERS WORK? [PARK AND KIM 2022]

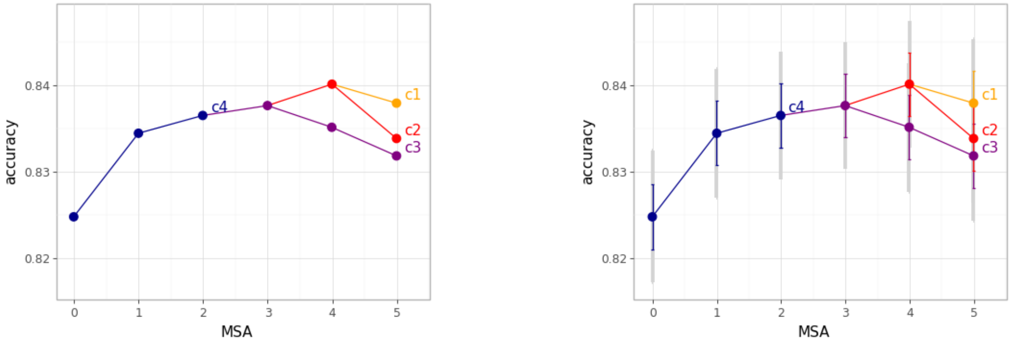


Figure 2: Reproduction of figure 12a from [Park and Kim 2022] (left). Augmentation of the same figure with estimated accuracy calculated using eq. (1) using a one-sigma 68.2% (colored) and two-sigma 95% (grey) confidence interval (right). Data to reproduce these figures was obtained by using [Rohatgi 2021] on the figures from the preprint PDF. Taken from [Steinbach et al. 2022]

MORE SOURCES OF VARIANCE

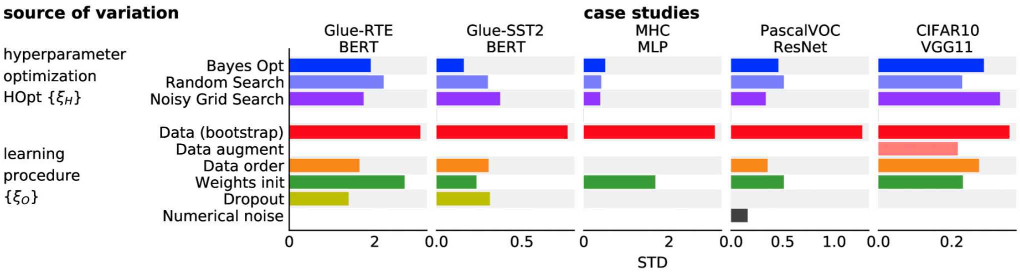


Figure 1 of [Bouthillier et al. 2021]: Different sources of variation of the measured performance: across our different case studies, as a fraction of the variance induced by bootstrapping the data. For hyperparameter optimization, we studied several algorithms.

TAKEAWAYS: LET'S ``INCREASE THE QUALITY OF EVIDENCE"[1]

- **uncertainties are essential**

  (strong hint for communicating and reviewing academic results)

[1] G. Varoquaux at ICLR's ML Eval workshop 2022

TAKEAWAYS: LET'S "INCREASE THE QUALITY OF EVIDENCE"[1]

- **uncertainties are essential**

  (strong hint for communicating and reviewing academic results)
- **uncertainties can be laborious**

  (cross-validation, running training multipe times)

[1] G. Varoquaux at ICLR's ML Eval workshop 2022

TAKEAWAYS: LET'S ``INCREASE THE QUALITY OF EVIDENCE"[1]

- **uncertainties are essential**

  (strong hint for communicating and reviewing academic results)
- **uncertainties can be laborious**

  (cross-validation, running training multipe times)
- **approximations for uncertainties** provide a solution **with minimal runtime cost**

[1] G. Varoquaux at ICLR's ML Eval workshop 2022

TAKEAWAYS: LET'S ``INCREASE THE QUALITY OF EVIDENCE"[1]

- **uncertainties are essential**
  (strong hint for communicating and reviewing academic results)
- **uncertainties can be laborious**
  (cross-validation, running training multipe times)
- **approximations for uncertainties** provide a solution **with minimal runtime cost**

[1] G. Varoquaux at ICLR's ML Eval workshop 2022

TAKEAWAYS: LET'S ``INCREASE THE QUALITY OF EVIDENCE"[1]

- **uncertainties are essential**
  (strong hint for communicating and reviewing academic results)
- **uncertainties can be laborious**
  (cross-validation, running training multipe times)
- **approximations for uncertainties** provide a solution **with minimal runtime cost**

See [Steinbach et al. 2022] for more details!

[1] G. Varoquaux at ICLR's ML Eval workshop 2022