Leibniz Universität Hannover

Well-Calibrated Regression Uncertainty in Medical Imaging with Deep Learning

Max-Heinrich Laves, Sontje Ihler, Jacob F. Fast, Lüder A. Kahrs, Tobias Ortmaier

Medical Imaging With Deep Learning (MIDL)

6–9 July 2020

Institute of Mechatronic Systems
Leibniz Universität Hannover

Motivation I

Regression in Medical Imaging

1. Age estimation from hand CT (Halabi et al., 2019)
2. Natural landmark localization (Payer et al., 2019)
3. Cell detection in histology (Xie et al., 2018)
4. Instrument pose estimation (Gessert et al., 2018)
5. Deformable registration (Dalca et al., 2019)

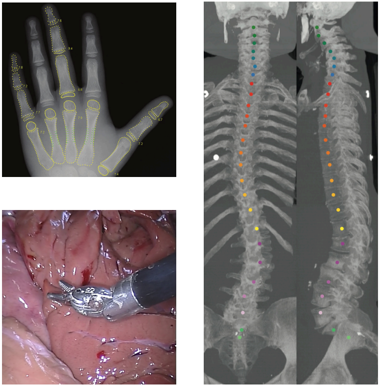


Figure: Medical regression tasks.

Motivation II

Predictive Uncertainty

- Reliable predictions are crucial
- Two types of uncertainty (Kendall et al., 2017)

Figure: Different types of uncertainty, note object boundaries (Kendall et al., 2017).

Institute of Mechatronic Systems
Leibniz Universität Hannover

Motivation II

Predictive Uncertainty

- Reliable predictions are crucial
- Two types of uncertainty (Kendall et al., 2017)
- Aleatoric
  - Arises from data directly (e. g. sensor noise)

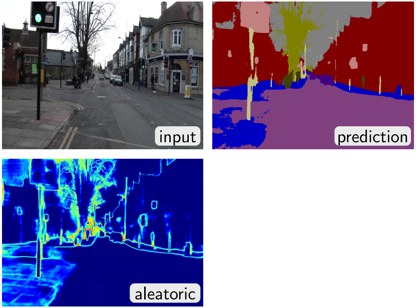


Figure: Different types of uncertainty, note object boundaries (Kendall et al., 2017).

Motivation II

Predictive Uncertainty

- Reliable predictions are crucial
- Two types of uncertainty (Kendall et al., 2017)
- Aleatoric
  - Arises from data directly (e. g. sensor noise)
- Epistemic
  - From limited training data

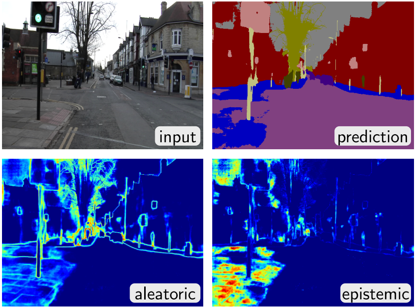


Figure: Different types of uncertainty, note object boundaries (Kendall et al., 2017).

Motivation II

Predictive Uncertainty

- Reliable predictions are crucial
- Two types of uncertainty (Kendall et al., 2017)
- Aleatoric
  - Arises from data directly (e. g. sensor noise)
- Epistemic
  - From limited training data
- Bayesian Neural Networks

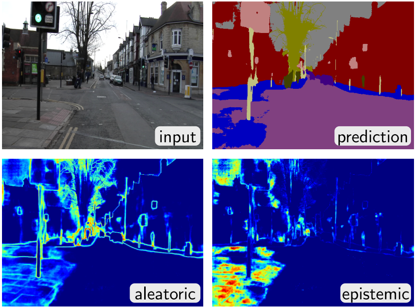


Figure: Different types of uncertainty, note object boundaries (Kendall et al., 2017).

Motivation II

Predictive Uncertainty

- Reliable predictions are crucial
- Two types of uncertainty (Kendall et al., 2017)
- Aleatoric
  - Arises from data directly (e. g. sensor noise)
- Epistemic
  - From limited training data
- Bayesian Neural Networks
- Uncertainty is miscalibrated

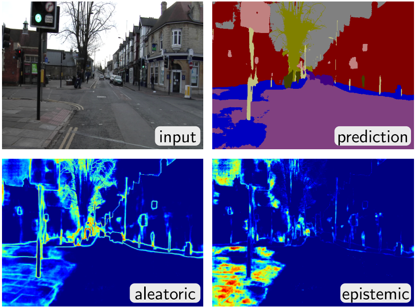


Figure: Different types of uncertainty, note object boundaries (Kendall et al., 2017).

Estimation of Aleatoric Uncertainty I

Leibniz Universität Hannover

Conditional Log-Likelihood for Regression

$$\boldsymbol{f_\theta}(\boldsymbol{x}) = \left[\hat{\boldsymbol{y}}(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x})\right], \;\; \hat{\boldsymbol{y}} \in \mathbb{R}^d$$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \frac{1}{\hat{\sigma}^2(\boldsymbol{x}_i)} \left\| \boldsymbol{y}_i - \hat{\boldsymbol{y}}(\boldsymbol{x}_i) \right\|^2 + \log \hat{\sigma}^2(\boldsymbol{x}_i)$$

Institute of Mechatronic Systems
Leibniz Universität Hannover
imes

Estimation of Aleatoric Uncertainty I

Conditional Log-Likelihood for Regression

$$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \left[\hat{\boldsymbol{y}}(\boldsymbol{x}), \hat{\sigma}^2(\boldsymbol{x})\right], \ \hat{\boldsymbol{y}} \in \mathbb{R}^d$$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \frac{1}{\hat{\sigma}^2(\boldsymbol{x}_i)} \left\| \boldsymbol{y}_i - \hat{\boldsymbol{y}}(\boldsymbol{x}_i) \right\|^2 + \log \hat{\sigma}^2(\boldsymbol{x}_i)$$

Problem Statement

Minimizing NLL w.r.t. $\hat{\sigma}^2(\boldsymbol{x}_i)$ yields

$$\hat{\sigma}^2(\boldsymbol{x}_i) = \underset{\hat{\sigma}^2(\boldsymbol{x}_i)}{\arg\min}\, \mathcal{L} = \left\| \boldsymbol{y}_i - \hat{\boldsymbol{y}}(\boldsymbol{x}_i) \right\|^2 \ \forall\, i \ .$$

Estimation of Aleatoric Uncertainty II

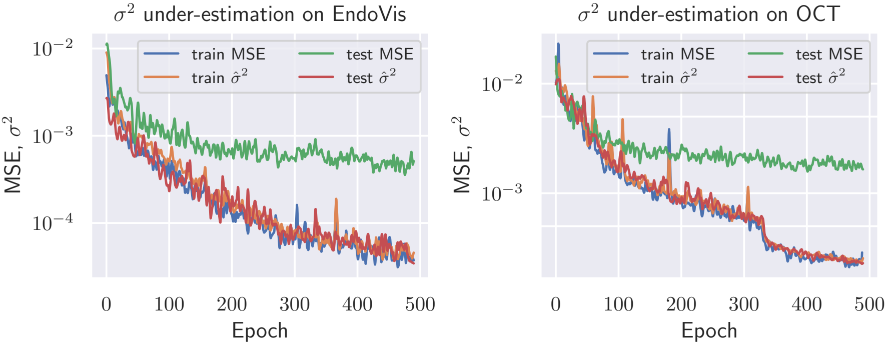


Figure: σ^2 is estimated relative to the MSE.

σ Scaling for Calibrated Regression Uncertainty

Leibniz Universität Hannover

Recalibration of Standard Deviation

$$p\left(\boldsymbol{y}|\boldsymbol{x}\right) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}(\boldsymbol{x}), (s \cdot \hat{\sigma})^2(\boldsymbol{x})\right)$$

Institute of Mechatronic Systems
Leibniz Universität Hannover

σ Scaling for Calibrated Regression Uncertainty

Leibniz Universität Hannover

Recalibration of Standard Deviation

$$p\left(\boldsymbol{y}|\boldsymbol{x}\right) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}(\boldsymbol{x}), (s \cdot \hat{\sigma})^2(\boldsymbol{x})\right)$$

$$\mathcal{L}(s) = m \log(s) + \frac{s^{-2}}{2} \sum_{i=1}^{m} \hat{\sigma}^2(\boldsymbol{x}_i) \left\| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{(i)} \right\|^2$$

Institute of Mechatronic Systems
Leibniz Universität Hannover

σ Scaling for Calibrated Regression Uncertainty

Recalibration of Standard Deviation

$$p\left(\boldsymbol{y}|\boldsymbol{x}\right) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}(\boldsymbol{x}), (s \cdot \hat{\sigma})^2(\boldsymbol{x})\right)$$

$$\mathcal{L}(s) = m \log(s) + \frac{s^{-2}}{2} \sum_{i=1}^{m} \hat{\sigma}^2(\boldsymbol{x}_i) \left\| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{(i)} \right\|^2$$

$$s = \pm \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(\hat{\sigma}_{\boldsymbol{\theta}}^{(i)}\right)^{-2} \left\| \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{(i)} \right\|^2} \,.$$

$\rightarrow$ We refer to this as σ scaling.

Well-Calibrated Estimation of Predictive Uncertainty

- So far: maximum posterior point estimate $\hat{\boldsymbol{\theta}}$
- Bayesian model with Monte Carlo dropout VI (Gal et al., 2016)

Well-Calibrated Estimation of Predictive Uncertainty

- So far: maximum posterior point estimate $\hat{\boldsymbol{\theta}}$
- Bayesian model with Monte Carlo dropout VI (Gal et al., 2016)

Predictive Uncertainty

Combines aleatoric (data) and epistemic (model) uncertainty (Kendall et al., 2017).

$$\hat{\Sigma}^2 = \underbrace{\frac{1}{N}\sum_{n=1}^{N}\left(\hat{\boldsymbol{y}}_n - \frac{1}{N}\sum_{n=1}^{N}\hat{\boldsymbol{y}}_n\right)^2}_{\text{epistemic}} + \underbrace{\frac{1}{N}\sum_{n=1}^{N}\hat{\sigma}_n^2}_{\text{aleatoric}}$$

VI under-estimates predictive variance.
$\rightarrow$ Apply σ scaling to calibrate predictive uncertainty $(s \cdot \hat{\Sigma}(\boldsymbol{x}))^2$.

Institute of Mechatronic Systems
Leibniz Universität Hannover

Quantification of Miscalibration

Leibniz Universität Hannover

Definition of Miscalibration

Difference in expectation between predictive error and uncertainty

$$\mathbb{E}_{\hat{\Sigma}^2}\left[\left|\left(\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2 \,\middle|\, \hat{\Sigma}^2=\Sigma^2\right)-\Sigma^2\right|\right] \quad \forall \left\{\Sigma^2 \in \mathbb{R} \,\middle|\, \Sigma^2 \geq 0\right\} \tag{1}$$

Institute of Mechatronic Systems
Leibniz Universität Hannover
imes

Quantification of Miscalibration

Definition of Miscalibration

Difference in expectation between predictive error and uncertainty

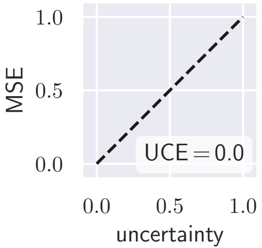
$$\mathbb{E}_{\hat{\Sigma}^2}\left[\left|\left(\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2 \,\middle|\, \hat{\Sigma}^2=\Sigma^2\right)-\Sigma^2\right|\right] \quad \forall\left\{\Sigma^2 \in \mathbb{R} \,\middle|\, \Sigma^2 \geq 0\right\} \tag{1}$$

Uncertainty Calibration Error

Partitioning into M bins (Guo et al., 2017)

$$\mathrm{UCE} := \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{err}(B_m) - \mathrm{uncert}(B_m)\right|$$

Calibration Diagram



Institute of Mechatronic Systems
Leibniz Universität Hannover

Experiments

- Four medical datasets with $\boldsymbol{y} \in \mathbb{R}^d$
  1. tumor cellularity in breast histology ($d = 1$) (Martel et al., 2019)
  2. RNSA bone age data set ($d = 1$) (Halabi et al., 2019)
  3. EndoVis surgical instrument tracking ($d = 2$) (EndoVis, 2015)
  4. needle pose estimation from 3D-OCT, own dataset ($d = 6$)[1]
- Uncertainty calibration
- Rejection of uncertain predictions
- Out-of-distribution detection (see paper)

[1] github.com/mlaves/3doct-pose-dataset

Intra-Training Calibration

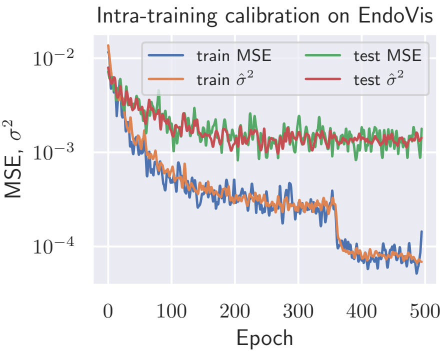


Figure: σ^2 is not under-estimated.

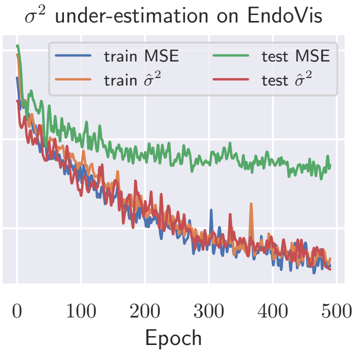
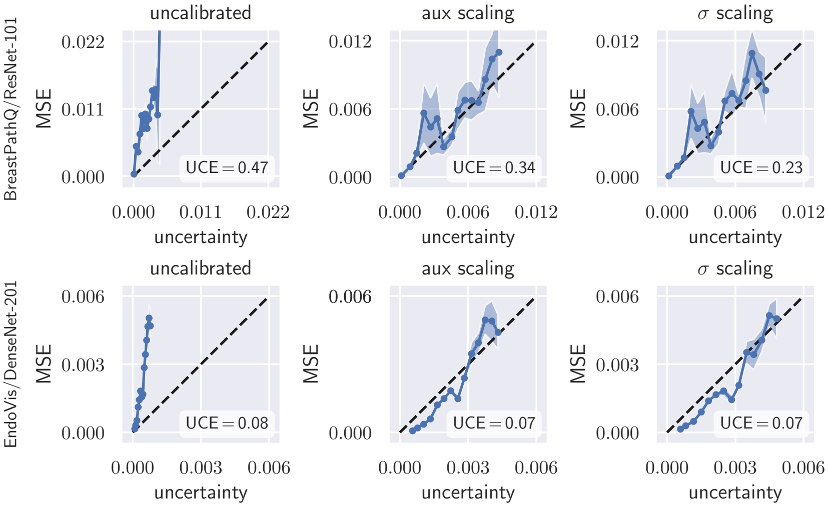


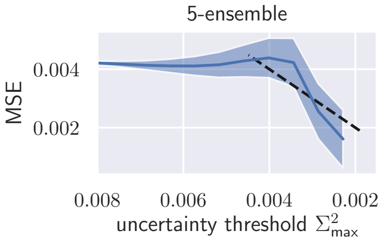
Figure: σ^2 is under-estimated.

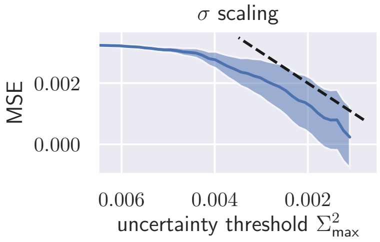
Calibration Diagrams



Rejection Experiments

- Uncertainty threshold Σ^2_{max}
- Reject, where $\hat{\Sigma}^2 > \Sigma^2_{\text{max}}$
- Reduce Σ^2_{max}, observe test MSE
- Compare to ensemble uncertainty
- σ scaling: monotonic decrease





EndoVis Example Result

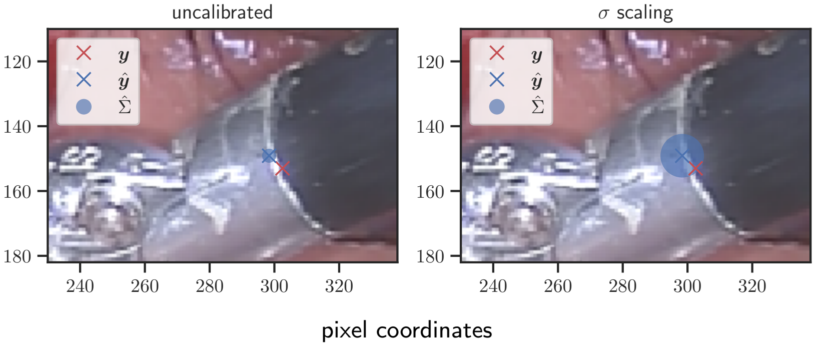


Figure: After σ scaling, the uncertainty better reflects the predictive error.

Conclusion

Leibniz Universität Hannover

- Well-calibrated predictive uncertainty for regression
- Miscalibration is considerably reduced
- If already calibrated: $s \to 1$
- Reliably detects distribution shift
- Ensemble outperformed on rejection task

Institute of Mechatronic Systems
Leibniz Universität Hannover
imes

Conclusion

- Well-calibrated predictive uncertainty for regression
- Miscalibration is considerably reduced
- If already calibrated: $s \to 1$
- Reliably detects distribution shift
- Ensemble outperformed on rejection task
- Simple to implement
- Does not affect accuracy
- Closes gap between test MSE and uncertainty
- Well-calibrated uncertainty should be considered in any medical imaging task with deep learning

Well-Calibrated Regression Uncertainty in Medical Imaging with Deep Learning

Max-Heinrich Laves, Sontje Ihler, Jacob F. Fast, Lüder A. Kahrs,
Tobias Ortmaier

Medical Imaging With Deep Learning (MIDL)

6–9 July 2020

Institute of
Mechatronic Systems
Leibniz Universität Hannover

References I

Dalca, Adrian V. et al. (2019). “Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces”. In: *Med Image Anal* 57, pp. 226–236. DOI: https://doi.org/10.1016/j.media.2019.07.006.

EndoVis (2015). *Instrument Subchallenge Dataset*. https://opencas.webarchiv.kit.edu/?q=node/30.

Gal, Yarin and Zoubin Ghahramani (2016). “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *ICML*, pp. 1050–1059.

Gessert, Nils, Matthias Schlüter, and Alexander Schlaefer (2018). “A deep learning approach for pose estimation from volumetric OCT data”. In: *Med Image Anal* 46, pp. 162–179. DOI: 10.1016/j.media.2018.03.002.

Guo, Chuan et al. (2017). “On Calibration of Modern Neural Networks”. In: *ICML*, pp. 1321–1330.

References II

Halabi, Safwan S. et al. (2019). "The RSNA Pediatric Bone Age Machine Learning Challenge". In: *Radiol* 290.2, pp. 498–503. DOI: 10.1148/radiol.2018180736.

Kendall, Alex and Yarin Gal (2017). "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *NeurIPS*, pp. 5574–5584.

Martel, A. L. et al. (2019). "Assessment of Residual Breast Cancer Cellularity after Neoadjuvant Chemotherapy using Digital Pathology [Data set]". In: *The Cancer Imaging Archive*. DOI: 10.7937/TCIA.2019.4YIBTJNO.

Payer, Christian et al. (2019). "Integrating spatial configuration into heatmap regression based CNNs for landmark localization". In: *Med Image Anal* 54, pp. 207–219. DOI: 10.1016/j.media.2019.03.007.

Xie, Yuanpu et al. (2018). "Efficient and robust cell detection: A structured regression approach". In: *Med Image Anal* 44, pp. 245–254. DOI: 10.1016/j.media.2017.07.003.

Institute of Mechatronic Systems
Leibniz Universität Hannover