MODEL CALIBRATION AND UNCERTAINTY QUANTIFICATION OF FINE-TUNED GEOSPATIAL FOUNDATION MODELS

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cnes CENTRE NATIONAL D'ÉTUDES SPATIALES

Christian Hümmer (christian.hummer@cnes.fr)[1], Paul Mauduit (paul.mauduit@thalesgroup.com)[2]
[1]Centre National d'Études Spatiales (CNES), [2] Thales Services Numeriques SAS

CONTEXT

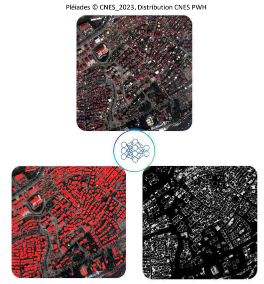
Why do we need uncertainty estimation?

- Models: over-/under confidence
- Reliability, interpretability & qualification of results
- Downstream-tasks benefit from reliability:
  - Natural disaster response (building damage)
  - Urban change detection (LU/LC monitoring, climate change adaptation & risk prevention)

Variety of uncertainty estimation (UE) methods for modern deep learning architectures [6], e.g.

- Approx. Bayesian Neural Networks (BNN)
- Ensemble Learning (EL)
- Test Time Augmentations (TTA)

Integration into Fine-tuning of geospatial Foundation Models?



MOTIVATION

- Previous work on uncertainty-aware change detection for natural disaster response has shown the benefits of integrating uncertainty estimation into our change detection pipeline while improving its reliability and providing a qualification of results
- The integration of dedicated UE methods helped in improving model calibration & uncertainty quantification when training CNN architectures from scratch, especially under distribution shift
- Aim: Integration of uncertainty estimation into transfer learning with pre-trained geospatial Foundation Models
- Built upon existing open-source frameworks: PANGAEA, PEFT, Lightning UQ-Box with additional metrics for uncertainty quantification
- Partially stochastic networks and PEFT methods for subspace Bayesian inference addressing the FM's high dimensional parameter-space
  - BNN-Decoder, Decoder-Sub-Ensembles, Checkpoint-Ensembles for frozen encoder training
  - Using reduced parameter subspaces of PEFT methods (LoRA) to obtain efficient uncertainty estimation

[9] Marsocci et al: "PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models", 2024, https://arxiv.org/abs/2412.04204.
[10] Mangrulkar et al: "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods", 2024, **https://github.com/huggingface/peft.**
[11] Lehmann et al: "Lightning UQ Box: Uncertainty Quantification for Neural Networks", 2025 http://jmlr.org/papers/v26/24-2110.html

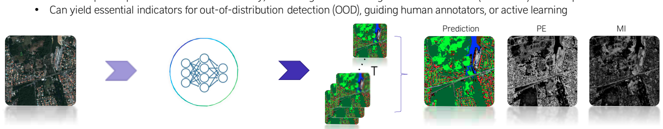
UNCERTAINTY

By adding baseline UE methods from the aforementioned categories (BNN Variational Inference - Bayes By Backprop [3], Monte Carlo Dropout [1,2], Deep Ensembles [4], TTA), we try to capture different types of uncertainty:

- *Epistemic* Uncertainty (model uncertainty): Due to insufficient training data (e.g. unseen o.o.d. samples)
- *Aleatoric* Uncertainty (data uncertainty): Due to ambiguity or noise inherent in our observations (data inherent randomness)
- Together: Predictive Uncertainty of the network

Different uncertainty quantification (UQ) measures exist to represent the model's uncertainty estimation [2]:

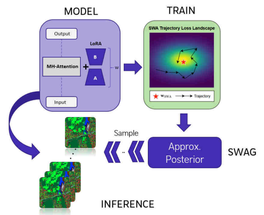
- Predictive Entropy (PE):
  - Represents the entropy of the predictive distribution
  - Captures predictive uncertainty, which combines both epistemic and aleatoric uncertainties.
- Mutual Information (MI), variance, mean class-wise standard deviation:
  - Rather capture epistemic or model uncertainty, stemming from the disagreement between ***T*** (stochastic) forward passes
  - Can yield essential indicators for out-of-distribution detection (OOD), guiding human annotators, or active learning



METHODOLOGY

- Stochastic Weight Averaging Gaussian (SWAG) [7] with LoRA has shown promising fine-tuning results in LLMs and depth estimation
- Builds on Stochastic Weight averaging (avg. of model weights over trajectory of SGD)
  - ➢ generalization, robustness
- Treats SGD iterates as samples from a Gaussian distribution
  - ➢ information in trajectory approximates posterior distribution over weights
- Fits a Gaussian distribution to the first two moments of SGD iterates

➢ LoRA-SWAG performs SWAG on the parameter-efficient subspaces constructed by the low-rank approximations of linear layers in Attention-Blocks



SWAG-LoRA. Illustration adapted from https://arxiv.org/abs/2405.03425

EVALUATION

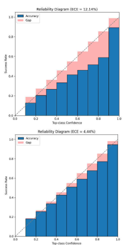
CAL

- Model confidence should match the segmentation accuracy - **calibration** quality (CAL):
- Brier-Score (Br): MSE between predicted probabilities and labels over all samples/pixels *n* for each class *k*

$$Br = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}[p(\hat{y}_i = y_k|x_i,\theta) - (\hat{y}_i = y_k)]^2$$

- Especially in binary classification/segmentation cases with severe class-imbalance and a dominating majority class, it can make sense to use the **Stratified-Brier-Score** instead
- Expected Calibration Error (ECE) based on reliability diagrams (accuracy as function over confidence):
  - Pixel-wise predictions are partitioned into m equally-sized bins based on confidence value
  - ECE: summing up the weighted average of differences between acc. and confidence / bin

$$ECE = \sum_{m=1}^{M}\frac{|B_m|}{n}|acc(B_m) - conf(B_m)|$$

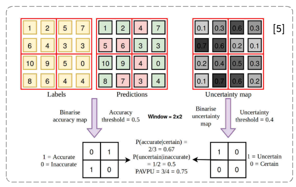


UQ

- Measuring DNN's performance by considering their **uncertainty quantification (UQ)** capabilities
- The Patch Accuracy vs Patch Uncertainty (PAvPU) [5] metric aims to capture these properties by two main conditional probabilities on a patch-level:

$$p(\mathbf{uncertain}|\mathbf{inaccurate}) = \frac{n_{iu}}{(n_{ic} + n_{iu})}, \quad p(\mathbf{accurate}|\mathbf{certain}) = \frac{n_{ac}}{(n_{ac} + n_{ic})} \quad \Rightarrow \quad \text{PAvPU} = \frac{(n_{ac} + n_{iu})}{(n_{ac} + n_{au} + n_{ic} + n_{iu})}$$

- Derived from confusion matrix of [in]accurate and [un]certain patches
- Results strongly depend on choice of uncertainty threshold

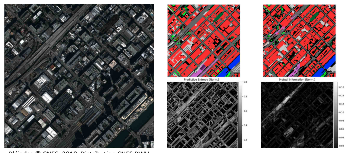


SUMMARY & FIRST RESULTS

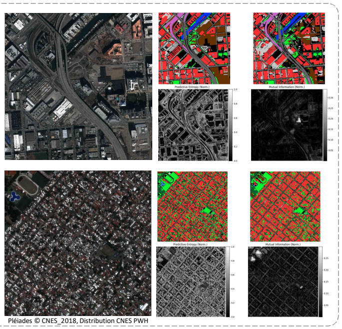
QUALITATIVE EXAMPLES

Example Digitanie (train France, test San-Francisco and Buenos Aires):

Ground Truth (left) compared to prediction (right) and associated uncertainty estimation from a PEFT SWAG-LoRA model (Scale-MAE encoder). Generally high PE (left) under geographic shift combines both types of uncertainty, whereas MI (right) rather captures unusual OOD samples, e.g triangle-shaped pool, sandy soccer field & untypical building structures.



Pléiades © CNES_2018, Distribution CNES PWH



Pléiades © CNES_2018, Distribution CNES PWH

QUANTITATIVE EXAMPLES

Some beneficial examples where dedicated UE methods helped in slightly improving the fine-tuning results under distribution shift.

The evaluation considers modified datasets that introduce a geographic shift between the train & test set, e.g.

Digitanie:

- Train (France) - Strasbourg, Arcachon, Biarritz, Montpellier, Toulouse, Paris
- Test – Cairo, San-Francisco, Can-Tho, Buenos Aires

HLS Burn-Scars (modified): Train (west coast), test (east coast)

HLS (MOD)

	CLASSIC	BNN	ENSEMBLE	SWAG
IoU ↑	60.79	62.41	62.09	61.35
Brier (Strat.) ↓	0.53	0.47	0.46	0.50

DIGITANIE (MOD)

	IoU ↑	Brier ↓	ECE ↓	A|C ↑	U|I ↑
LoRA-Classic	47.92	0.47	0.13	0.84	0.77
LoRA-SWAG	**49.66**	**0.44**	**0.10**	**0.87**	**0.80**

WORK IN PROGRESS

- *First comparison of different methods using available datasets for natural disaster response, urban planning & change detection*
  - *Change detection & natural disaster response: SpaceNet, HLS BurnScars, xBD/xView2*
  - *HR & VHR multiclass semantic segmentation (urban): SegMunich (Sentinel-2), Digitanie (Pléiades, manually annotated)*
- *Dedicated test scenarios to compare network calibration and uncertainty quantification under different data constraints:*
  - *Data sparsity: 50 %, 10 % subsampled datasets (stratified / random)*
  - *Domain shift: geographically divided subsets*
- Initial observation: Partially stochastic networks and PEFT methods for subspace Bayesian inference constitute a baseline for parameter-efficient uncertainty estimation in the foundation model fine-tuning context, but:
  - In the case of fine-tuning geospatial FM's, the improvement in uncertainty quantification seems to be **less impactful** as for the integration of dedicated UE approaches when training basic model architectures (e.g. Unet) from scratch
  - Both, simple approaches like Sub-/Checkpoint-Ensembles or MCDO and BNN/(LoRA)-SWAG can **slightly** improve **reliability, model calibration and predictive performance** for **certain cases**
  - However: The first results **do not** indicate **general, consistent and significant improvements** in model calibration and uncertainty quantification over multiple test configurations -> the scenario-dependent results require **more benchmarking**
- **Dedicated metrics for model reliability can help in quantifying model calibration and uncertainty estimation capabilities and in picking a suitable uncertainty estimation method for specific use-cases**

[1] Y. Gal, Z. Ghahramani, "Dropout as a Bayesian approximation: representing model uncertainty in deep learning" in Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, pp 1050–1059, 2016
[2] Y. Gal, "Uncertainty in Deep Learning", Dissertation, 2016, https://api.semanticscholar.org/CorpusID:86522127
[3] C. Blundell et al., "Weight uncertainty in neural networks" in Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, pp 1613–1622, 2015
[4] B. Lakshminarayanan, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles" in Advances in Neural Information Processing System, 2017
[5] J. Mukhoti, Y. Gal, "Evaluating Bayesian Deep Learning Methods for Semantic Segmentation", arXiv e-prints, pp.arXiv-1811, 2018
[6] J. Gawlikowski et al. "A survey of uncertainty in deep neural networks", in Artificial Intelligence Review, 56, pp.1513-1589, 2023
[7] W.J. Maddox et al. "A simple baseline for Bayesian uncertainty in deep learning", in Advances in Neural Information Processing Systems, pp. 13153-13164, 2019