



university of
groningen

artificial
intelligence

Uncertainty in Machine Learning LLMs/VLMs and Computer Vision

Dr. Matias Valdenegro

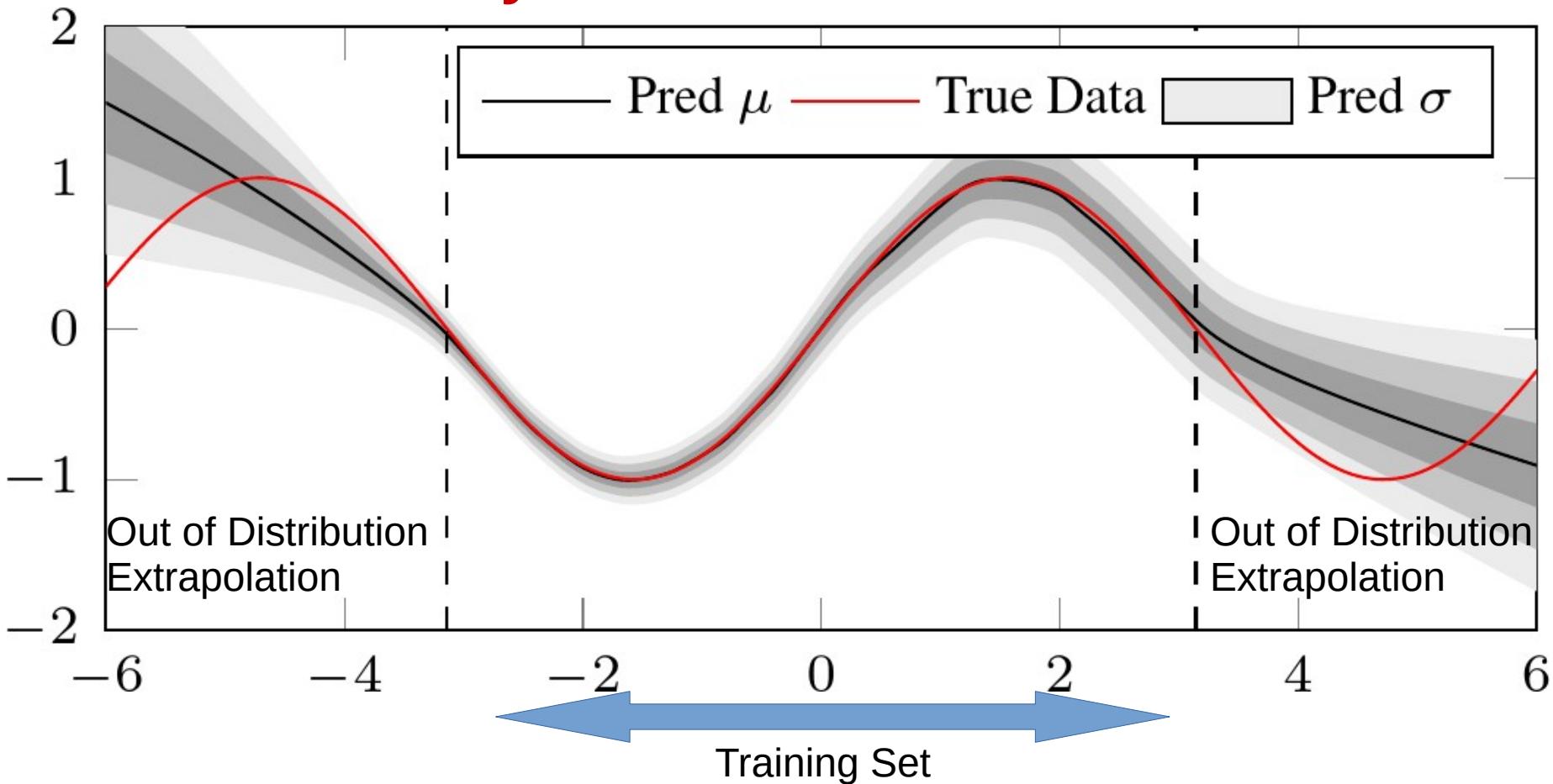
Assistant Professor of Machine Learning (Tenured)

m.a.valdenegro.toro@rug.nl

CWI Workshop UQ in High-Dimensional Problems

With contributions of my Bachelor, Master, and PhD Students.

Uncertainty Estimation



Uncertainty Estimation

- › Most ML models do not have proper uncertainty estimation built-in.
- › Additional methods can be used: MC-Dropout, Ensembling, Bayesian NNs, etc.
- › The point is that output confidence/uncertainty reflects correctness of prediction.

Uncertainty in Machine Learning

- › Methods in a nutshell.
 - MC-Dropout / MC-DropConnect. Enable at inference.
 - Ensembles. Train N models on same data.
 - Bayesian Neural Networks via Flipout, using Variational Inference. Weights are Gaussian distributions.
- › Predictive posterior distribution reconstructed via sampling/ensembling, computing Gaussian mean/std of samples/ensemble results.



university of
groningen

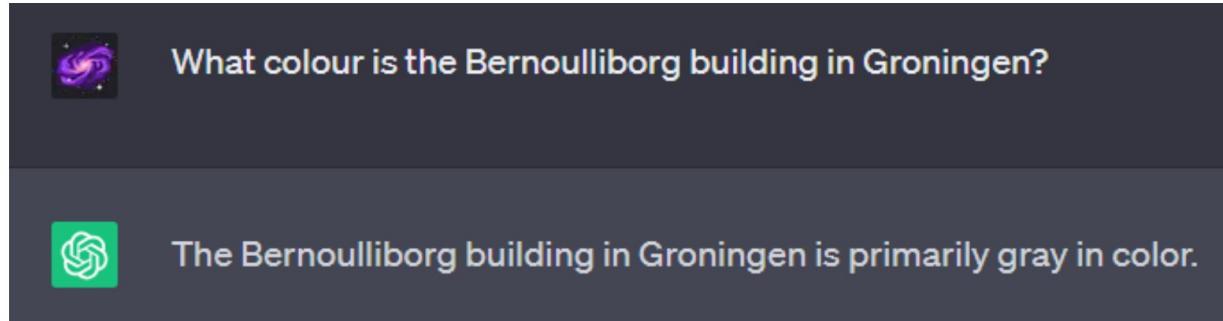
artificial
intelligence

Large Language and Vision- Language Models

Verbalized Uncertainty

Introduction

This is work of Martino Pelucchi, Bachelor Thesis
Available at <https://arxiv.org/abs/2311.06427>



(a) ChatGPT Prompt and Answer



(b) Bernoulliborg

Fig. 1. ChatGPT answering with an incorrect fact rather than admitting its lack of knowledge on the topic. The building in question is primarily blue as shown on the right.

LLMs would give you a hallucination (lie) instead of admitting not knowing.

Verbalized Uncertainty - Prompt



Hello, I would like you to perform sentiment analysis on 3 english sentences. Please classify them as positive, negative, or neutral. Moreover, please give your confidence in the answer you gave between 0 and 100%. The answer should be a list with the format “Sentiment (confidence%)”. Say “understood” if you understand. The list of sentences will follow.



Understood, please provide the three sentences for sentiment analysis.



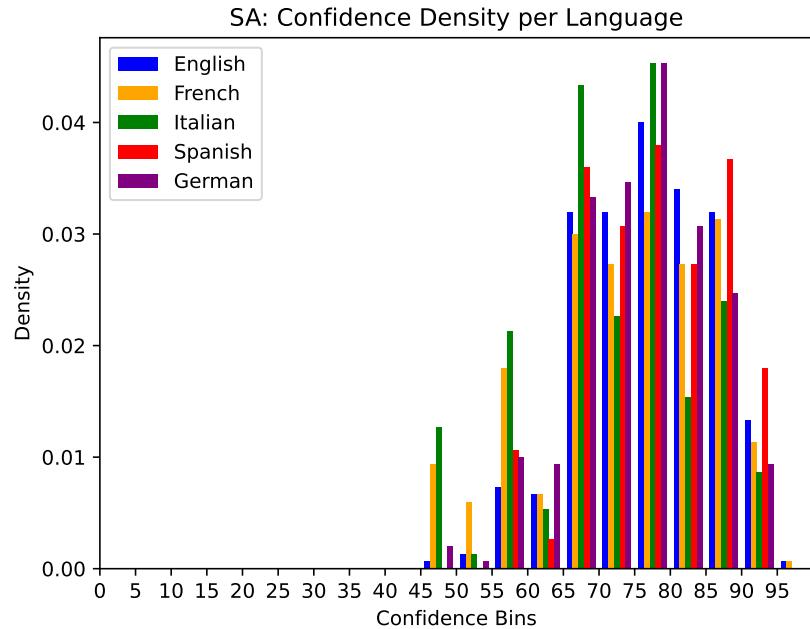
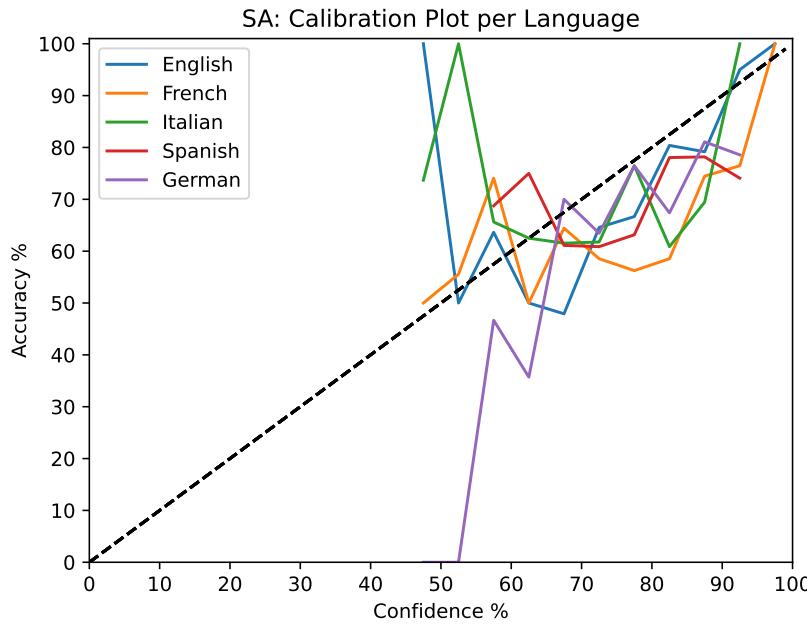
1. Can I just fast forward to the Ed Sheeran concert Saturday
2. Love and Books and Valentines
3. "I liked a @user video John Cena lectures Nikki Bella about the laundry hampers: Total Divas, December 1,"



1. Positive (85%)
2. Neutral (70%)
3. Neutral (55%)

Calibration Evaluation

Calibration and Confidence Estimation is **bad**.



No confidence lower than 40%.

Vision-Language Model Uncertainty

This is work of Tobias Groot, Bachelor Thesis
Available at <https://arxiv.org/abs/2405.02917>



How many consumable sushi pieces are in this photo?
Moreover, please rate your confidence in your answer
between 0 and 100%. The answer should be in the format:
"Answer (confidence%)".



ChatGPT

Answer: 39 (90%) [\[–\]](#)



Verbalized Uncertainty can be
extended to VLMs.
Japanese objects dataset
captured by Matias.

Confidence Intervals and
mean/standard deviation also
possible.

Japanese Objects Dataset



1. **Prompt:** How many food items are shown in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



2. **Prompt:** How many desserts are shown in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



3. **Prompt:** How many sushi pieces are displayed here? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



7. **Prompt:** Who is depicted in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



8. **Prompt:** Where was this photo taken? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



9. **Prompt:** How many octopuses are shown in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".

Correct Answer: Unknown, there are around 50 meals/plates, but a lot more food items. The ideal answer of the model would be to have 0% confidence and not give a prediction.

Correct Answer: 20-30

Correct Answer: 201



4. **Prompt:** What is shown in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



5. **Prompt:** How many sushi pieces are in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



6. **Prompt:** What kind of food is presented in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



10. **Prompt:** How many objects are shown in this photo, what are they? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



11. **Prompt:** How many babies are shown in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".



12. **Prompt:** What is depicted in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".

Correct Answer: Okonomiyaki

Answer: **Correct Answer:** Right answer would notice that these are sushi drawings.

Correct Answer: Egg fried rice, fried chicken, and gyoza. Japanese food is also correct.

Correct Answer: 30-35 Ema plaques.

Correct Answer: 0, because these are not babies.

Correct Answer: A Japanese graveyard or cemetery.

38 Images challenging Japanese objects and settings.

VLM Verbalized Uncertainty



Prompt: How many birds are shown in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)"



Prompt: Where was this photo taken? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)"



Prompt: How many octopuses are shown in this photo? Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)"

GPT4V

It is not feasible to count the exact number of birds in the photo due to their movement, the potential for overlap, and the sheer number (confidence 0%). The image shows a flock of birds in flight, which makes it particularly challenging to provide an accurate count from a single static image.

Hiroshima, Japan, during a lantern floating ceremony (confidence 90%)

1 octopus (confidence 100%)

Gemini Pro Vision

100 (80%)

Hiroshima, Japan (80%)

2 (100%)

Correct Answer

Definitely over 100 birds, hard to count the exact number. Ideally no prediction and 0% confidence.

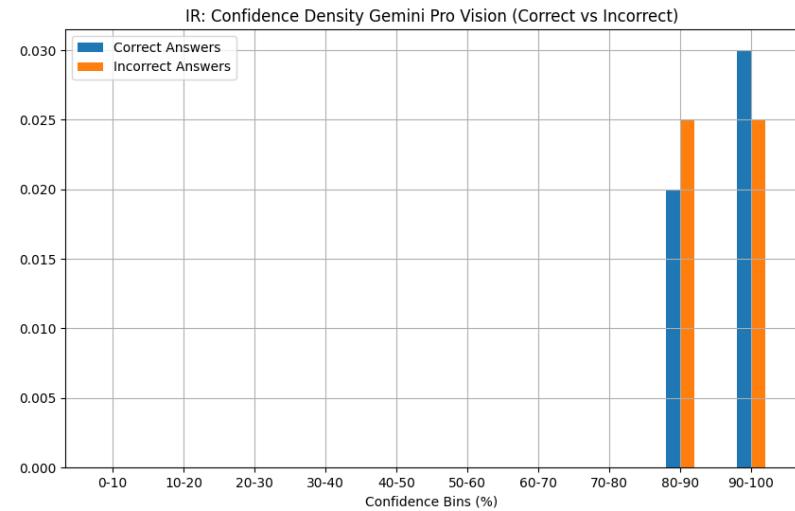
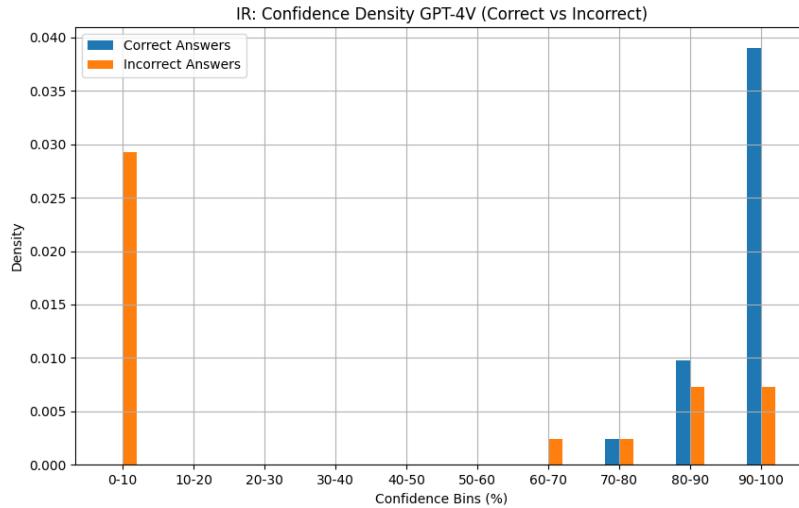
Japan, Hiroshima, Atomic Bomb Dome

5.

(Photo taken in Osaka, Dotonbori Area)

Not promising!

VLM Uncertainty Evaluation



Overall VLM verbalized uncertainty is not reliable, does not separate right from wrong answers.

But GPT4V seems to be better at UQ.

VLM Confidence Intervals and Mean/Std

Mean/standard deviation Prompts.

Moreover, please express your answer by giving a mean and a standard deviation to reflect the uncertainty in your answer. The answer should be in the format: "Mean = [mean value], SD = [standard deviation value]"

Confidence Interval Prompts.

Moreover, please express your estimate as a 95% confidence interval. This means you should provide a range within which you are 95% confident the true value lies. Format your answer as: '[Lower Bound, Upper Bound]', where the lower bound is the start of the range and the upper bound is the end of the range. Ensure that this interval reflects a 95% confidence level based on your estimation.

VLM Confidence Intervals and Mean/Std



Prompt: How many sushi pieces are displayed here?

GPT4V

Mean = 230,
Std = 10

CI. [155,159]

Gemini Pro Vision

Mean = 96.3,
Std = 1.5

201



Prompt: How many octopuses are shown in this photo?

Mean = 1,
Std = 1

CI. [1,1]

Mean = 1,
Std = 0.0

5



Prompt: How many lamps are shown in this photo?

Mean = 24,
Std = 3

CI. [12,22]

Mean = 15,
Std = 12

23

VLMs cannot do regression uncertainty.

VLM Uncertainty Robustness

This is work of Mirko Borszukovszki, Bachelor Thesis.

Verbalized Uncertainty + Image Corruptions at different severities.

Expectation is that uncertainty is not affected.



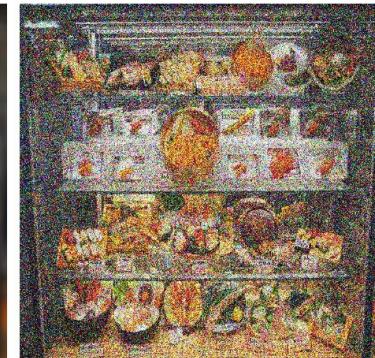
Original image



JPEG compression

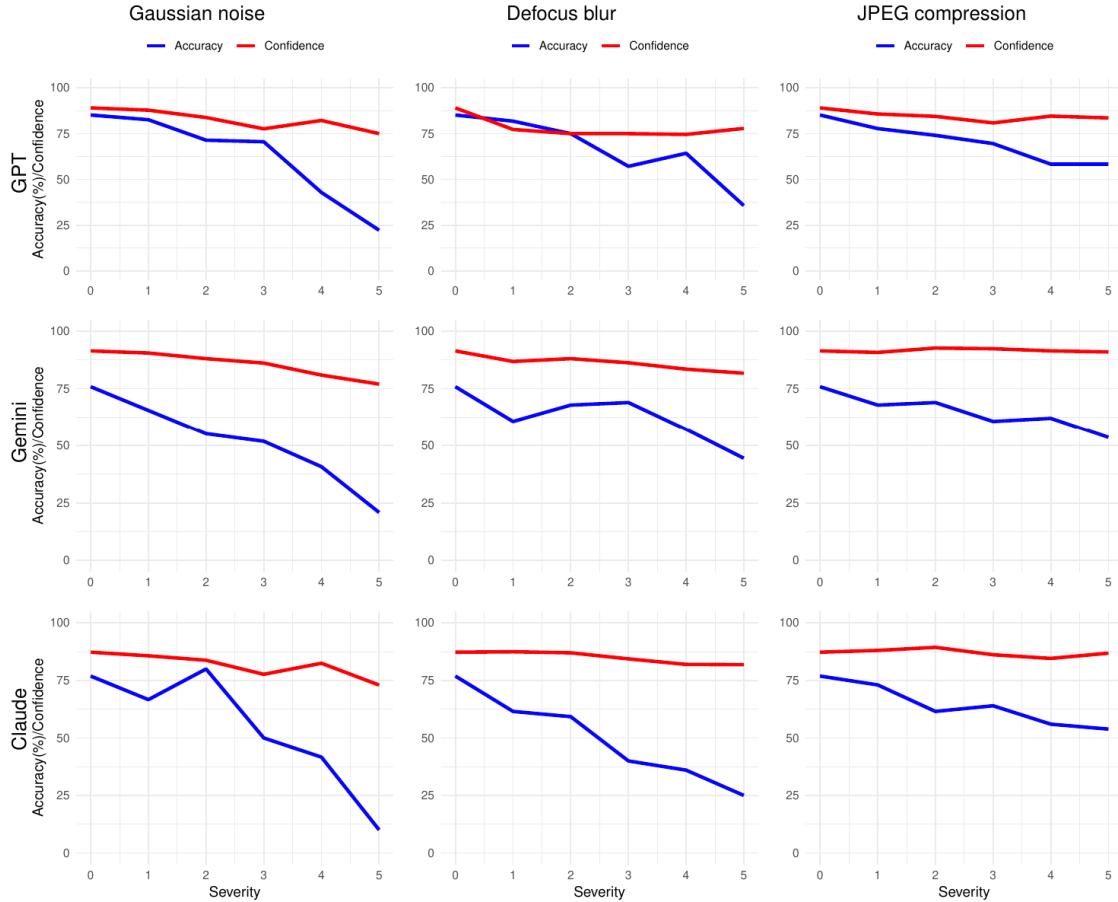


Defocus blur



Gaussian noise

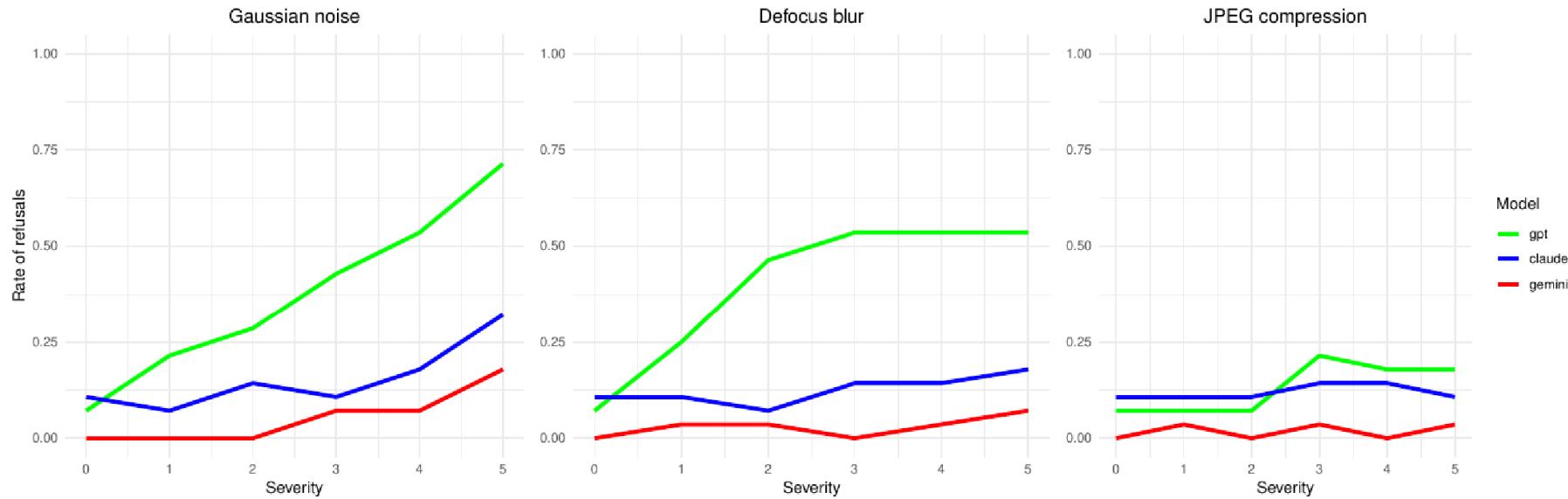
VLM Uncertainty Robustness



Results on JUS.

Gap between confidence and accuracy shows overconfidence.

VLM Uncertainty Robustness



GPT4 is trained to detect Gaussian Noise corruptions, but not other corruptions.

Discussion LLM/VLM Uncertainty

- › Definitely not reliable!
- › Would you trust a LLM/VLM now?
- › GPT4V seems to work a bit better than Gemini Pro Vision.
- › Further research required to properly imbue LLMs/VLMs with uncertainty estimation.
- › To detect hallucinations and incorrect answers.



university of
groningen

artificial
intelligence

Explainable Machine Learning

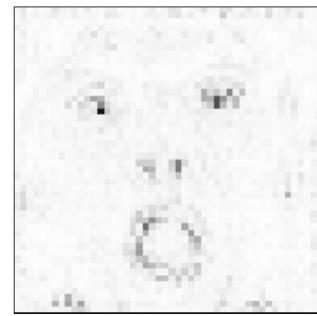
Explanation Uncertainty and Computer Vision

Introduction

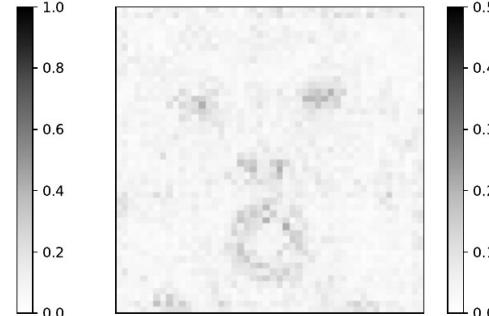
This is work of Mihir Mulye, Master Thesis.



(a) Input Image



(b) Explanation Mean

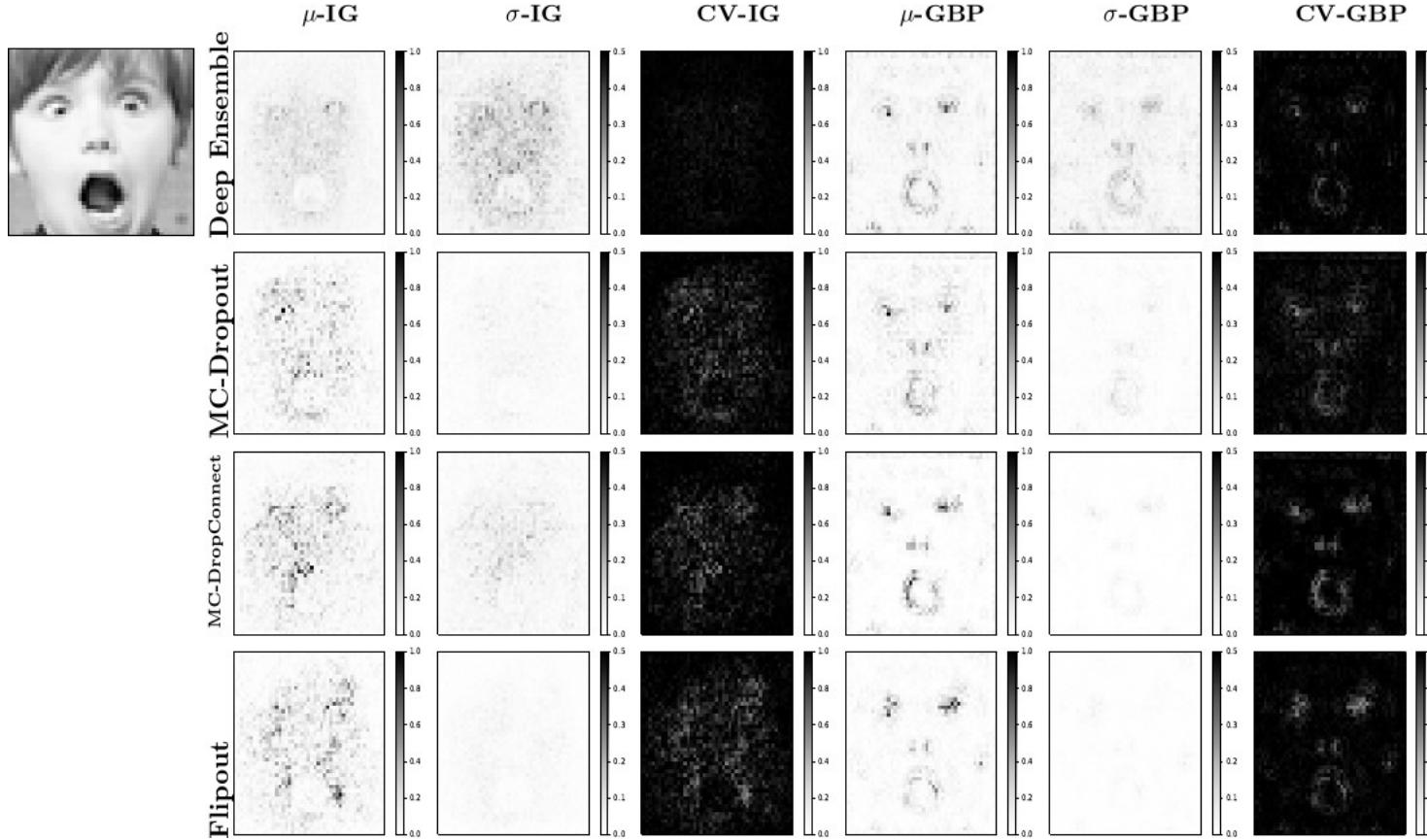


(c) Explanation Uncertainty

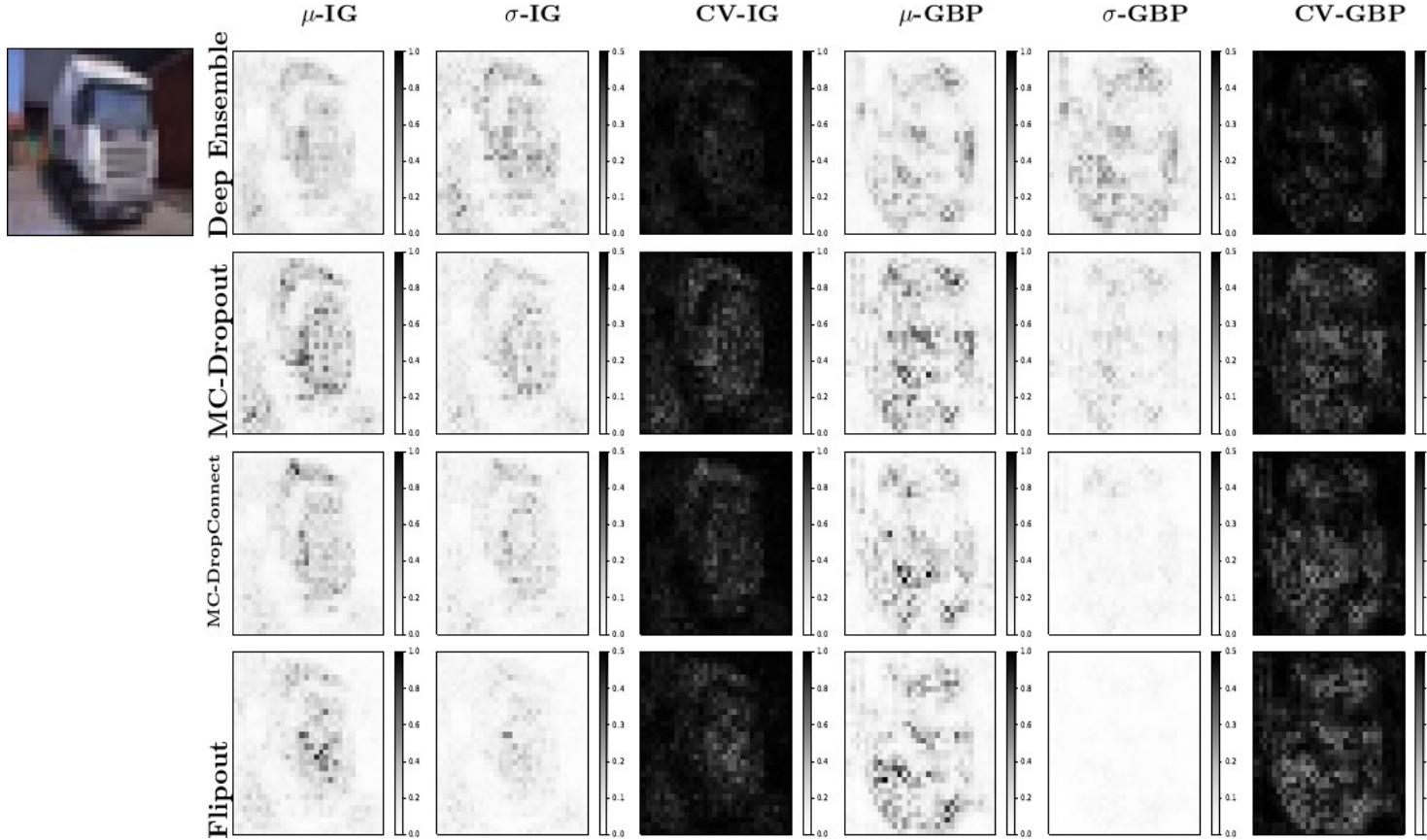
Fig. 1: Example explanation with uncertainty on a FER dataset image using Guided Backpropagation and Ensembles.

UQ Methods can be combined with XAI method to produce explanation uncertainty.

Explanation Uncertainty - FER+



Explanation Uncertainty - CIFAR10



Evaluation ExpUQ – Random Weights

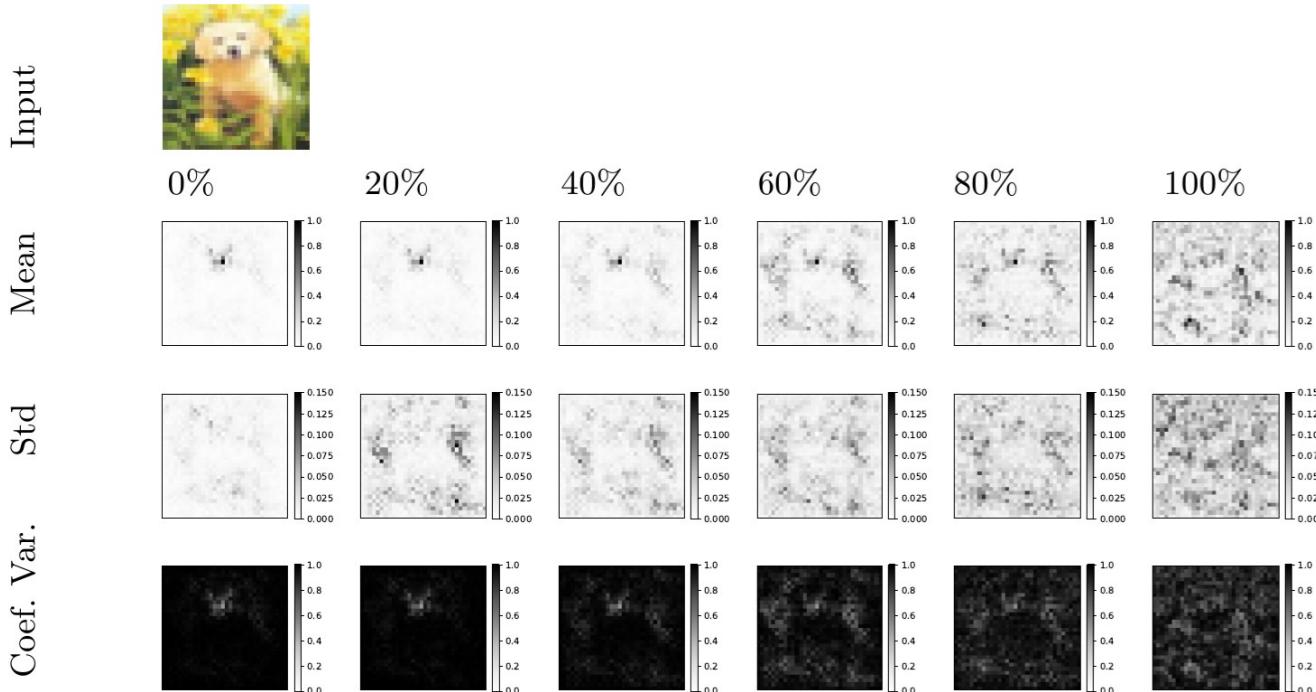


Fig. 4: Visualization of weight randomization effect on the explanations obtained using Dropout and Guided Backpropagation on a CIFAR10 sample. Note how the mean and standard deviation explanation become noisy with increasing random weights.

Evaluation ExpUQ – Random Labels



(a) Input Image

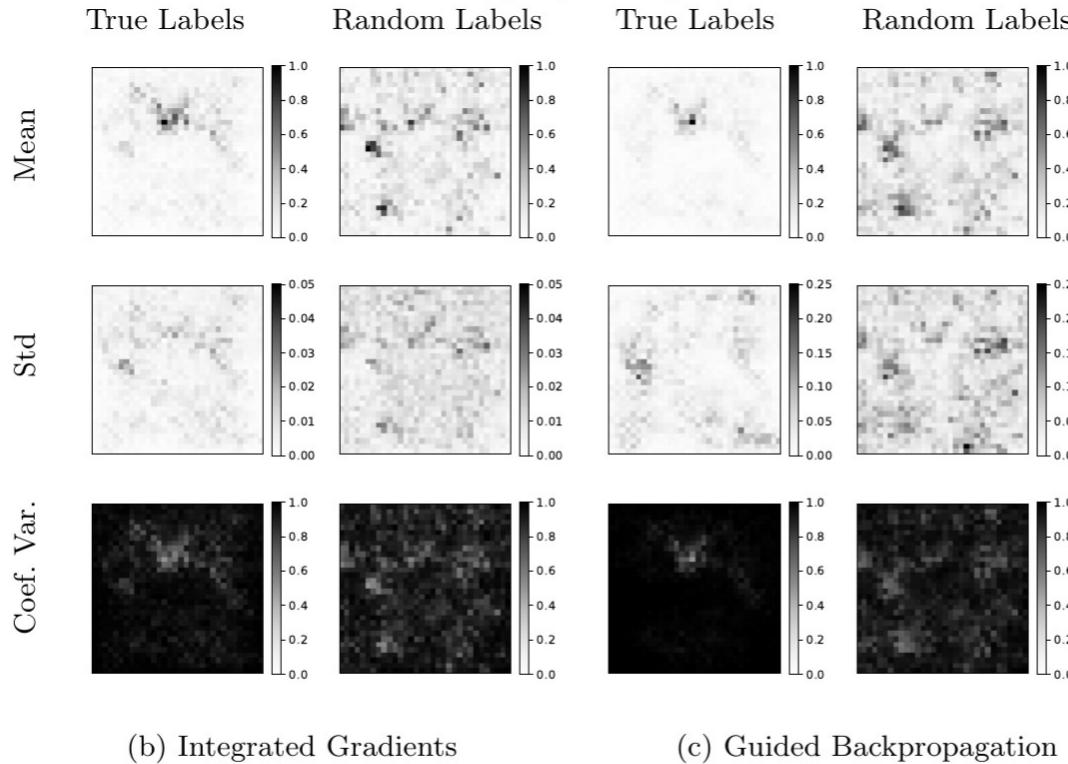


Fig. 5: Visualization of data randomization effect on explanations obtained using Dropout and GBP and IG. Note how the explanation for the model with random labels is considerably more noisy than explanation from a model with true labels.

Super Resolution Uncertainty

This is work of Sai Adapa, Master Thesis.
SRGAN Ensemble.



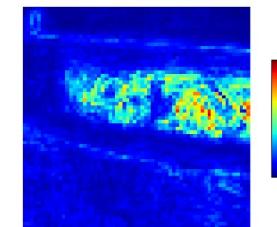
(a) Areas of interest



(b) Ground Truth



(c) Mean output

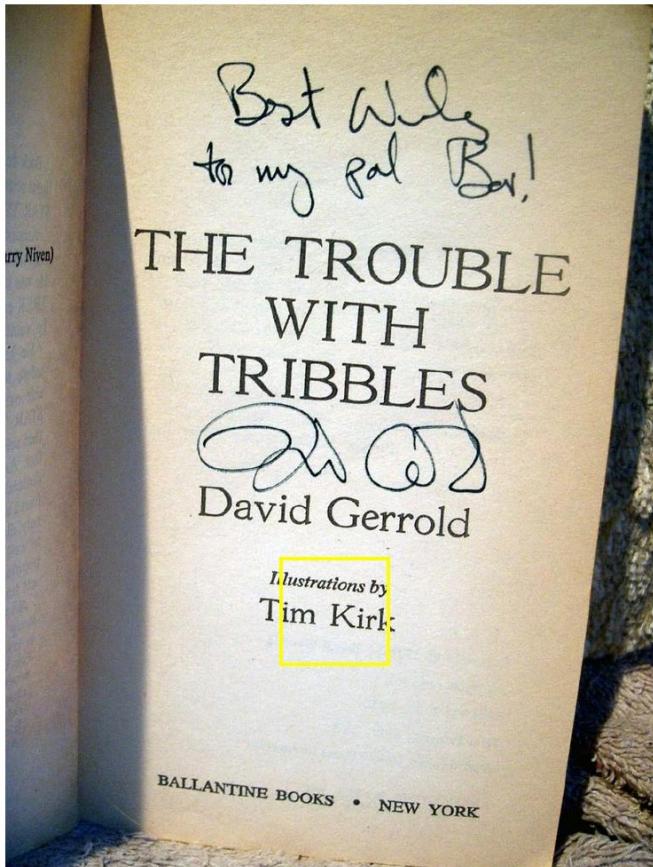


(d) Uncertainty map

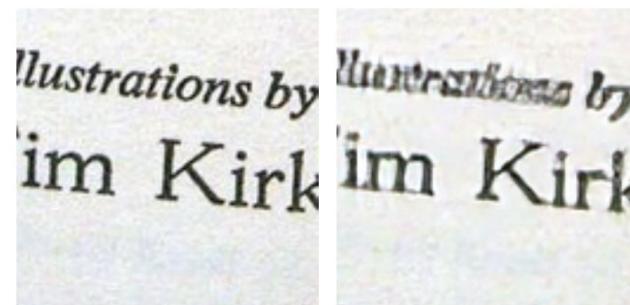
Be aware
where to
trust
predictions

UQ Methods can be applied to Super Resolution
for increased “explainability” of predictions.

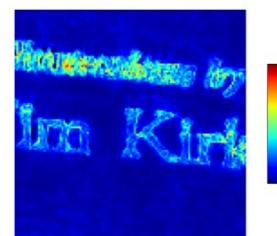
Super Resolution Uncertainty



(a) Areas of interest



(b) Ground Truth (c) Mean output



(d) Uncertainty map

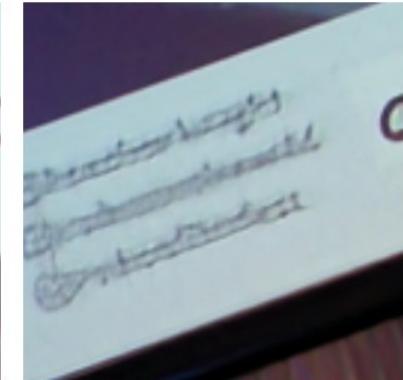
Super Resolution Uncertainty



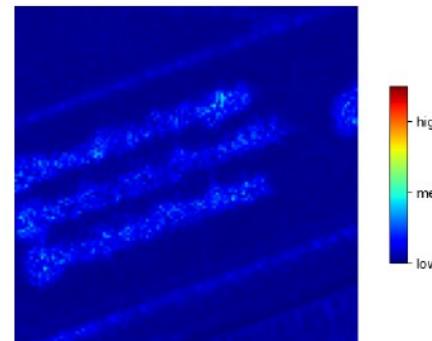
(a) Areas of interest



(b) Ground Truth



(c) Mean output



(d) Uncertainty map



university of
groningen

artificial
intelligence

Input Uncertainty

Aleatoric + Epistemic + Input Uncertainties

<https://arxiv.org/abs/2406.18787>

Input Uncertainty Propagation

IU MC Sampling. An alternative way to propagate IU through a model f_θ is via Monte Carlo sampling:

$$f_\theta(\mu^i \pm \sigma^{2^i}) \approx \mathbb{E}[f_\theta(\tilde{x})] \pm \text{Var}[f_\theta(\tilde{x})] \quad \tilde{x} \sim \mathcal{N}(\mu^i, \sigma^{2^i}) \quad (1)$$

The input is modelled as Gaussian random variable \tilde{x} , from which samples are drawn that are passed through the model f_θ . The result is an approximate Gaussian $\mathcal{N}(f_\theta(\tilde{x}), \text{Var}[f_\theta(\tilde{x})])$. The monte carlo approximation is computed using N samples via the sample mean and variance.

IU Propagation. Given an input with uncertainty $x = \{\mu^i, \sigma^i\}$, it is possible to propagate the uncertainties through a model f_θ by using a first order Taylor approximation:

$$f_\theta(\mu^i \pm \sigma^{2^i}) \approx f_\theta(\mu^i) \pm \mathbf{J} \sigma^{2^i} \mathbf{J}^T \quad (2)$$

Where $\mu^o = f_\theta(\mu^i)$ and $\sigma^{2^o} = \mathbf{J} \sigma^{2^i} \mathbf{J}^T$ and \mathbf{J} is the Jacobian matrix.

Formulation

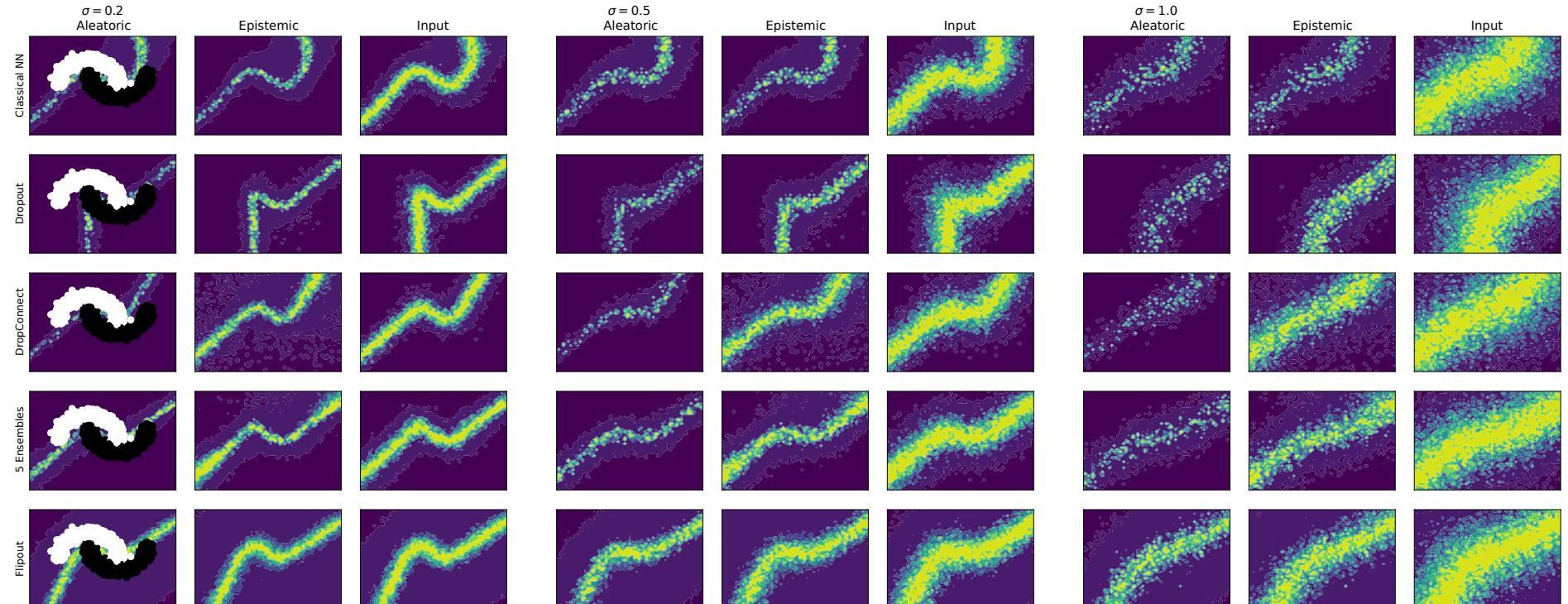
To build the formulation, we combine an input uncertainty estimation (propagate) with an epistemic uncertainty estimation method (epistemic) like MC-Dropout, Ensembles or Flipout, which produces samples $S_{\text{epi}}(\mu, \sigma^2)$ that consider input and epistemic uncertainty.

$$\begin{aligned} S_{\text{epi}}(\mu, \sigma^2) &= \text{propagate}(\text{epistemic}(\mu), \sigma^2) \\ \mu^s &= \mathbb{E}[S_{\text{epi}}(\mu, \sigma^2)] \\ \sigma^{2s} &= \text{Var}[S_{\text{epi}}(\mu, \sigma^2)] \end{aligned}$$

The general formulation then outputs three values:

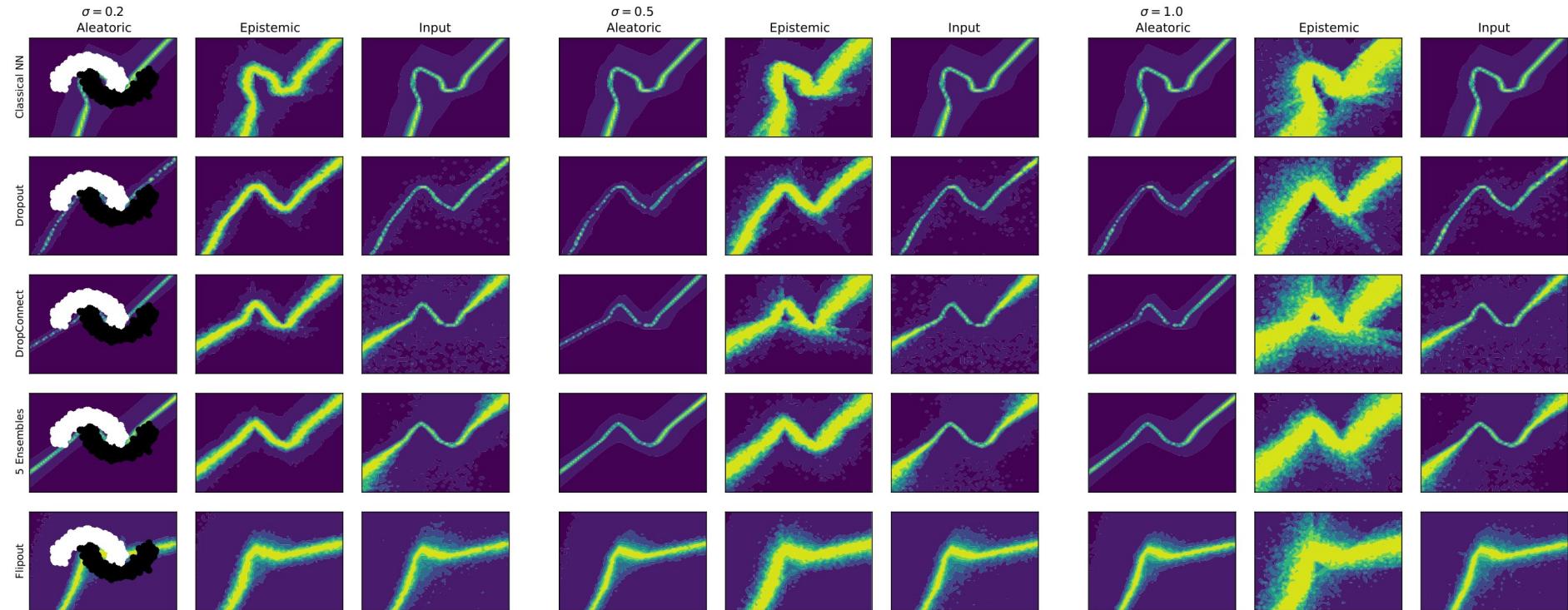
$$\begin{aligned} \mu^o &= \mathbb{E}[\mu^s] && \text{(Prediction)} \\ \sigma_{\text{inp}}^o &= \text{Var}[\mu^s] && \text{(IU)} \\ \sigma_{\text{epi}}^o &= \mathbb{E}[\sigma^{2s}] && \text{(EU)} \end{aligned}$$

Results on Two Moons – MC Sampling



Train with $\sigma = 0.1$ and test with $\sigma = 0.2, 0.5, 1.0$.
Uncertainty due to input should increase (it does).

Results on Two Moons – Propagation



Train with $\sigma = 0.1$ and test with $\sigma = 0.2, 0.5, 1.0$.
Uncertainty due to input should increase (it does not).

Take Home Messages

- › VLMs and LLMs cannot estimate their own uncertainty, further research is needed.
- › UQ + XAI methods can produce explanation uncertainty, for users to be aware which explanation regions to trust.
- › UQ can also be applied to Super Resolution, for additional information to the end user, where to trust.
- › Never blindly trust predictions from a model!

Questions?

