What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?

Alex Kendall
University of Cambridge
agk34@cam.ac.uk

Yarin Gal
University of Cambridge
yg279@cam.ac.uk

NIPS 2017

Outline

- Introduction
- Uncertainty in Bayesian Neural Network
- Method
  - Regression
  - Classification
- Experiment

Outline

- Introduction
- Uncertainty in Bayesian Neural Network
- Method
  - Regression
  - Classification
- Experiment

Importance of Uncertainty

Two Examples:

- An assisted driving system confuses the white side of a trailer for bright sky
- An image classification system erroneously identified two African Americans as gorillas

If assign a high level of uncertainty to their erroneous predictions, then the system may have been able to avoid disaster. However,

- Deep learning does not allow for uncertainty representation in regression settings
- Deep learning classification models just give normalized score vectors, which do not necessarily capture uncertainty.

Uncertainty

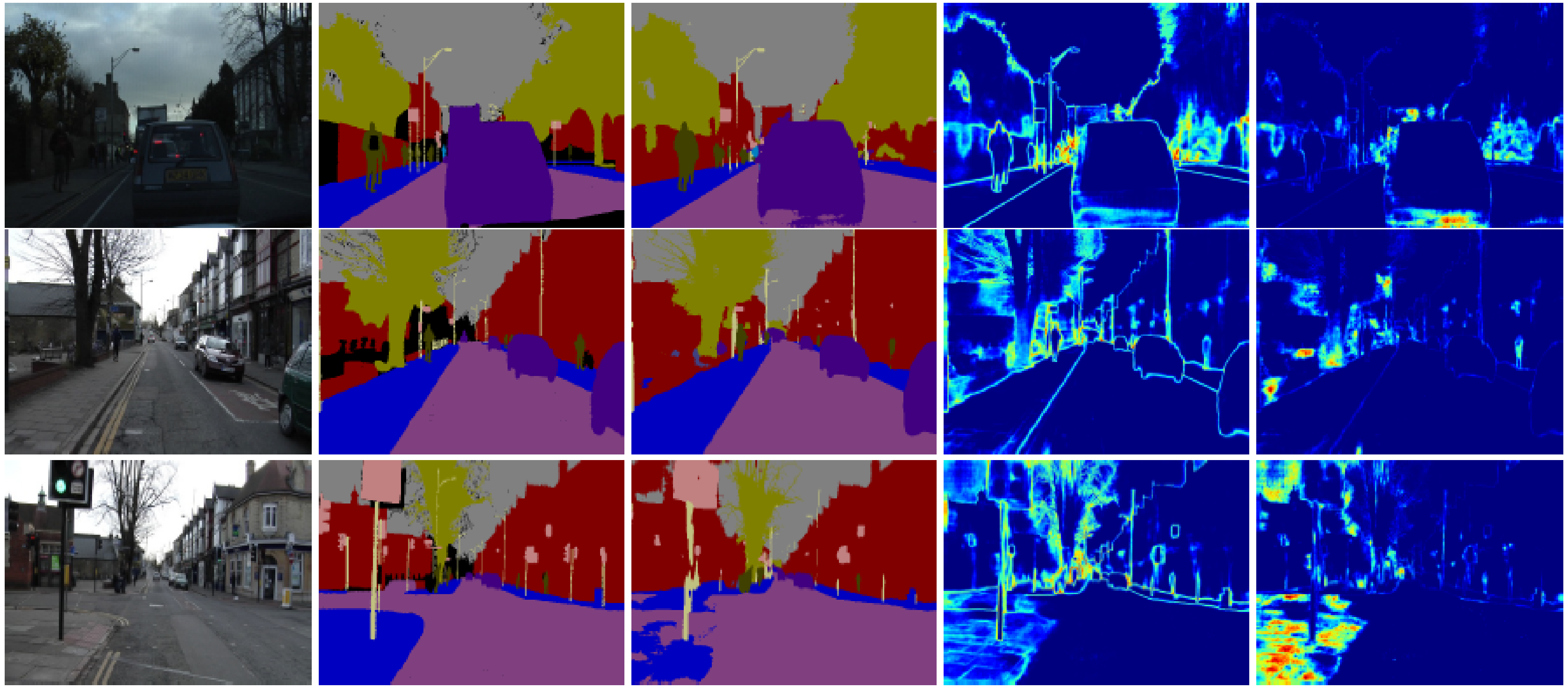
Aleatoric uncertainty

- captures noise inherent in the observations
- cannot be reduced

Epistemic uncertainty

- captures model's lack of knowledge
- can be explained away with the large amounts of data

Example in Segmenation



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Figure 1: **Illustrating the difference between aleatoric and epistemic uncertainty** for semantic segmentation on the CamVid dataset [8]. *Aleatoric* uncertainty captures noise inherent in the observations. In (d) our model exhibits increased aleatoric uncertainty on object boundaries and for objects far from the camera. *Epistemic* uncertainty accounts for our ignorance about which model generated our collected data. This is a notably different measure of uncertainty and in (e) our model exhibits increased epistemic uncertainty for semantically and visually challenging pixels. The bottom row shows a failure case of the segmentation model when the model fails to segment the footpath due to increased epistemic uncertainty, but not aleatoric uncertainty.

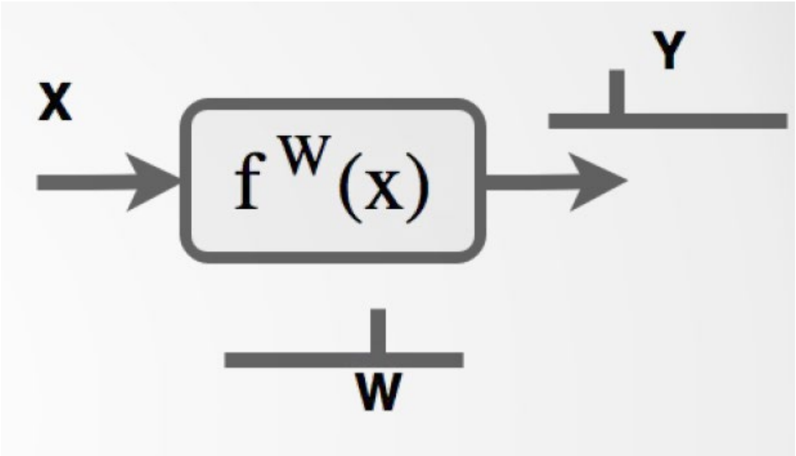
Outline

- Introduction
- Uncertainty in Bayesian Neural Network
- Method
  - Regression
  - Classification
- Experiment

NN v.s. BNN

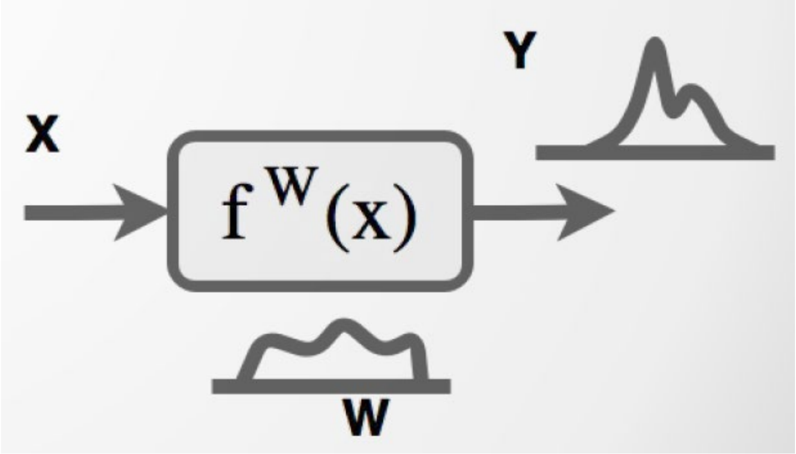
☐ Neural Network

- Finds a function y = f(x)
- One best model



☐ Bayesian Neural Network

- Distribution over weights
- Output is a distribution
- Many models within a network



Dropout Variational Inference

```
y = []
for _ in xrange(10):
    y.append(model.output(x, dropout=True))
y_mean = numpy.mean(y)
y_var = numpy.var(y)
```

$$\mathrm{KL}(q_\theta^*(\mathbf{W}), p(\mathbf{W}|\mathbf{X}, \mathbf{Y}))$$

[Yarin Gal, Zoubin Ghahramani. Bayesian convolutional neural networks with Bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158, 2015.]

Epistemic Uncertainty

- Classification:

$$p(y=c|\mathbf{x},\mathbf{X},\mathbf{Y}) \approx \frac{1}{T}\sum_{t=1}^{T} \mathrm{Softmax}(\mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x}))$$

$$H(\mathbf{p}) = -\sum_{c=1}^{C} p_c \log p_c$$

- Regression:

$$\mathrm{Var}(\mathbf{y}) \approx \boxed{\sigma^2} + \boxed{\frac{1}{T}\sum_{t=1}^{T} \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x})^T \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x}\ ) - E(\mathbf{y})^T E(\mathbf{y})}$$

$$E(\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}^{\widehat{\mathbf{W}}_t}(\mathbf{x})$$

$$\frac{1}{T}\sum_{t=1}^{T}\left(f^{\widehat{W}_t}(x) - E(y)\right)^2$$

Aleatoric Uncertainty

- Minimization objective $\widehat{\mathbf{W}}_i \sim q_\theta^*(\mathbf{W})$

$$\mathcal{L}(\theta, p) = -\frac{1}{N}\sum_{i=1}^{N} \boxed{\log p(\mathbf{y}_i | \mathbf{f}^{\widehat{\mathbf{W}}_i}(\mathbf{x}_i))} + \boxed{\frac{1-p}{2N}||\theta||^2}$$

Regularization term

Minimize negative log likelihood

- For a Gaussian likelihood

$$-\log p(\mathbf{y}_i | \mathbf{f}^{\widehat{\mathbf{W}}_i}(\mathbf{x}_i)) \propto \frac{1}{2\sigma^2}||\mathbf{y}_i - \mathbf{f}^{\widehat{\mathbf{W}}_i}(\mathbf{x}_i)||^2 + \frac{1}{2}\log\sigma^2$$

- Learn aleatoric uncertainty as a function of the data:

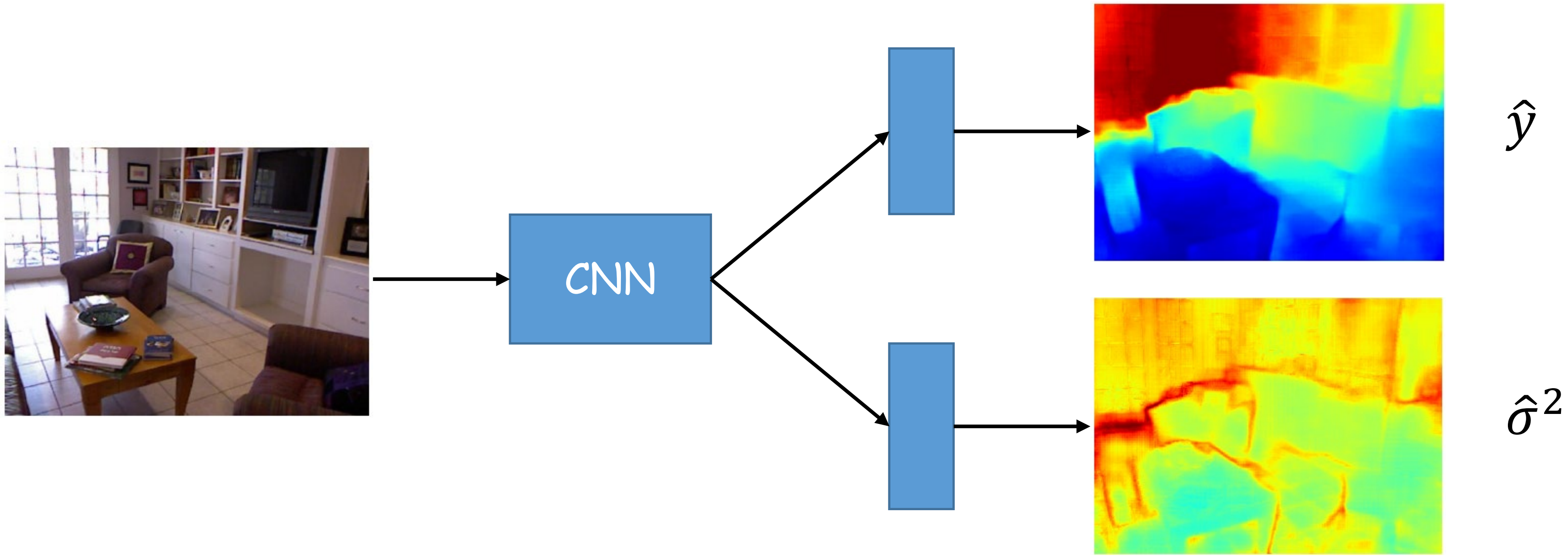
$$\mathcal{L}_{\text{NN}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2\sigma(\mathbf{x}_i)^2}||\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i)||^2 + \frac{1}{2}\log\sigma(\mathbf{x}_i)^2$$

- Does not capture epistemic model uncertainty which is a property of the model and not of the data.

Outline

- Introduction
- Uncertainty in Bayesian Neural Network
- Method
- Experiment

Model



Regression

- Model output: $[\hat{\mathbf{y}}, \hat{\sigma}^2] = \mathbf{f}^{\widehat{\mathbf{W}}}(\mathbf{x})$
- Minimization objective: $\mathcal{L}_{BNN}(\theta) = \frac{1}{D}\sum_i \frac{1}{2}\hat{\sigma}_i^{-2}||\mathbf{y}_i - \hat{\mathbf{y}}_i||^2 + \boxed{\frac{1}{2}\log \hat{\sigma}_i^2}$

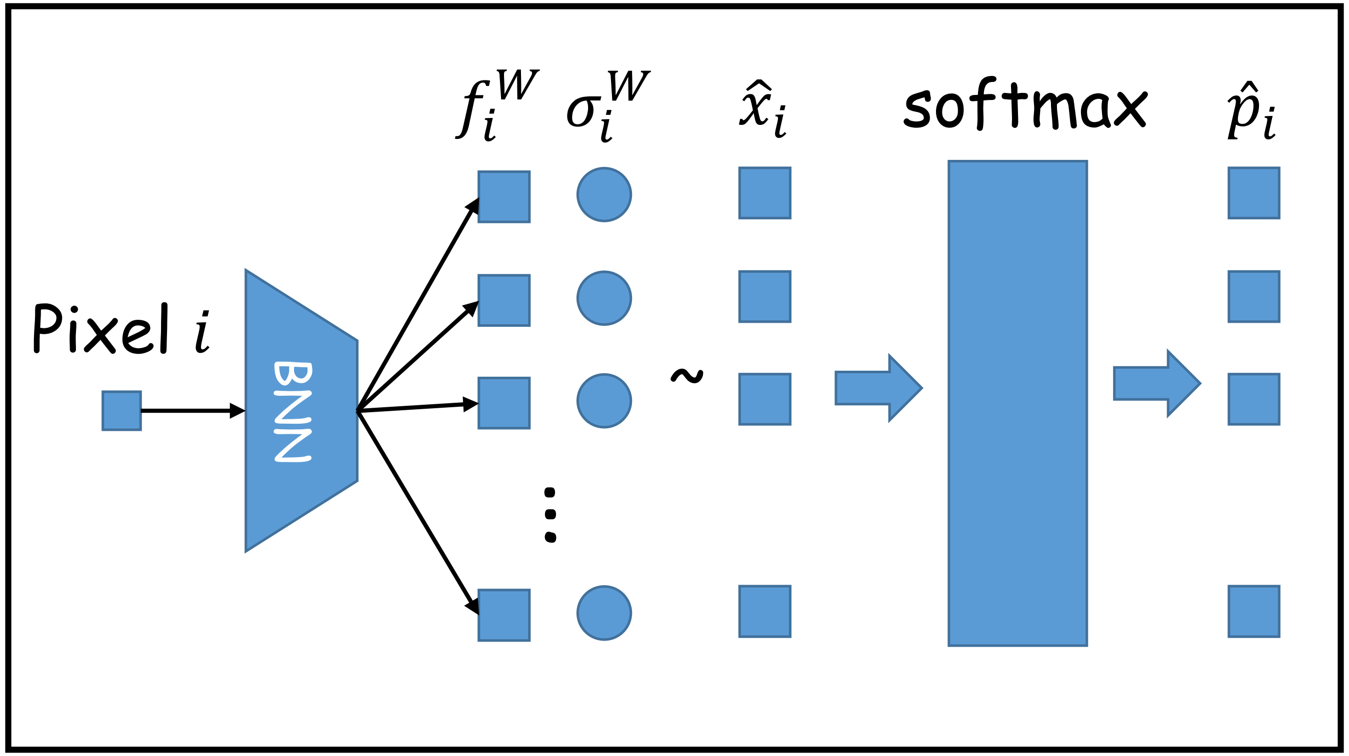
prevents the network from predicting infinite uncertainty

- Do not need 'uncertainty labels' to learn uncertainty, only need to supervise the learning of the regression task.

- The predictive uncertainty:

$$\mathrm{Var}(\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t\right)^2 + \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_t^2$$

Classification



Inference:

$$\hat{\mathbf{x}}_i|\mathbf{W} \sim \mathcal{N}\boxed{(\mathbf{f}_i^{\mathbf{W}}, (\sigma_i^{\mathbf{W}})^2)}$$ Model Output

$$\hat{\mathbf{p}}_i = \mathrm{Softmax}(\hat{\mathbf{x}}_i).$$

Training:

$$\hat{\mathbf{x}}_{i,t} = \mathbf{f}_i^{\mathbf{W}} + \sigma_i^{\mathbf{W}} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I)$$

$$\mathcal{L}_x = \sum_i \log \frac{1}{T} \sum_t \exp(\hat{x}_{i,t,c} - \log \sum_{c'} \exp \hat{x}_{i,t,c'})$$

$$-\mathbf{y}_i \log \hat{\mathbf{p}}_i = -y_{i,c} \log \hat{p}_{i,c} = -y_{i,c} \log \frac{\exp(\hat{x}_{i,c})}{\sum_{c'} \exp(\hat{x}_{i,c'})}$$

Outline

- Introduction
- Uncertainty in Bayesian Neural Network
- Method
- Experiment

Experiments

Semantic Segmentation

CamVid: road scene segmentation dataset

NYUv2: indoor segmentation dataset

CamVid	IoU
SegNet [28]	46.4
FCN-8 [29]	57.0
DeepLab-LFOV [24]	61.6
Bayesian SegNet [22]	63.1
Dilation8 [30]	65.3
Dilation8 + FSO [31]	66.1
DenseNet [20]	66.9
This work:	
DenseNet (Our Implementation)	67.1
+ Aleatoric Uncertainty	67.4
+ Epistemic Uncertainty	67.2
+ Aleatoric & Epistemic	**67.5**

(a) CamVid dataset for road scene segmentation.

NYUv2 40-class	Accuracy	IoU
SegNet [28]	66.1	23.6
FCN-8 [29]	61.8	31.6
Bayesian SegNet [22]	68.0	32.4
Eigen and Fergus [32]	65.6	34.1
This work:		
DeepLabLargeFOV	70.1	36.5
+ Aleatoric Uncertainty	70.4	37.1
+ Epistemic Uncertainty	70.2	36.7
+ Aleatoric & Epistemic	**70.6**	**37.3**

(b) NYUv2 40-class dataset for indoor scenes.

Table 1: **Semantic segmentation performance.** Modeling both aleatoric and epistemic uncertainty gives a notable improvement in segmentation accuracy over state of the art baselines.

Experiments

Pixel-wise Depth Regression

Make3D	rel	rms	$\log_{10}$
Karsch et al. [33]	0.355	9.20	0.127
Liu et al. [34]	0.335	9.49	0.137
Li et al. [35]	0.278	7.19	0.092
Laina et al. [26]	0.176	4.46	0.072
This work:			
DenseNet Baseline	0.167	3.92	0.064
+ Aleatoric Uncertainty	**0.149**	3.93	**0.061**
+ Epistemic Uncertainty	0.162	**3.87**	0.064
+ Aleatoric & Epistemic	**0.149**	4.08	0.063

(a) Make3D depth dataset [25].

NYU v2 Depth	rel	rms	$\log_{10}$	δ_1	δ_2	δ_3
Karsch et al. [33]	0.374	1.12	0.134	-	-	-
Ladicky et al. [36]	-	-	-	54.2%	82.9%	91.4%
Liu et al. [34]	0.335	1.06	0.127	-	-	-
Li et al. [35]	0.232	0.821	0.094	62.1%	88.6%	96.8%
Eigen et al. [27]	0.215	0.907	-	61.1%	88.7%	97.1%
Eigen and Fergus [32]	0.158	0.641	-	76.9%	95.0%	98.8%
Laina et al. [26]	0.127	0.573	0.055	81.1%	95.3%	98.8%
This work:						
DenseNet Baseline	0.117	0.517	0.051	80.2%	95.1%	98.8%
+ Aleatoric Uncertainty	0.112	0.508	0.046	81.6%	95.8%	98.8%
+ Epistemic Uncertainty	0.114	0.512	0.049	81.1%	95.4%	98.8%
+ Aleatoric & Epistemic	**0.110**	**0.506**	**0.045**	**81.7%**	**95.9%**	**98.9%**

(b) NYUv2 depth dataset [23].

Table 2: **Monocular depth regression performance.** Comparison to previous approaches on depth regression dataset NYUv2 Depth. Modeling the combination of uncertainties improves accuracy.

Analysis

Train dataset	Test dataset	RMS	Aleatoric variance	Epistemic variance
Make3D / 4	Make3D	5.76	0.506	7.73
Make3D / 2	Make3D	4.62	0.521	4.38
Make3D	Make3D	3.87	0.485	2.78
Make3D / 4	NYUv2	-	0.388	15.0
Make3D	NYUv2	-	0.461	4.87

(a) Regression

Train dataset	Test dataset	IoU	Aleatoric entropy	Epistemic logit variance ($\times 10^{-3}$)
CamVid / 4	CamVid	57.2	0.106	1.96
CamVid / 2	CamVid	62.9	0.156	1.66
CamVid	CamVid	67.5	0.111	1.36
CamVid / 4	NYUv2	-	0.247	10.9
CamVid	NYUv2	-	0.264	11.8

(b) Classification

1. Aleatoric uncertainty cannot be explained away with more data,
2. Aleatoric uncertainty does not increase for out-of-data examples (situations different from training set), whereas epistemic uncertainty does.