N-version machine learning models for safety critical systems

Fumio Machida

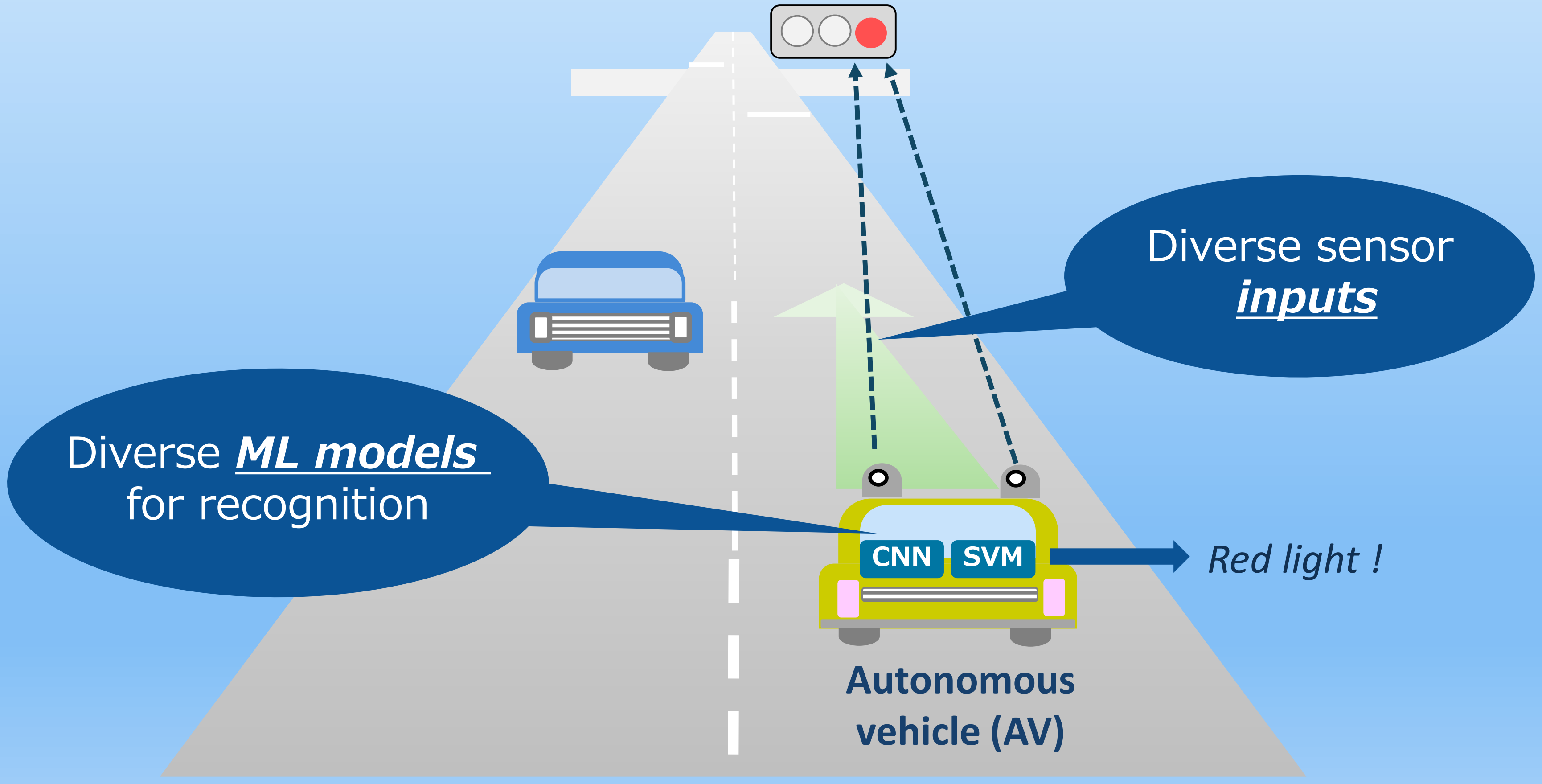
University of Tsukuba

June 24, 2019

In Dependable and Secure Machine Learning 2019

Machine learning (ML) in AV

For safe driving, a red light on the road ahead should be recognized accurately



Outline

1. Background
2. N-version machine learning architecture
3. Reliability model
4. Numerical example
5. Conclusion

Quality assurance of ML systems

Quality control becomes an emergent challenge for ML system providers

- ML systems
  - Information systems increasingly employ ML module as a core of intelligent function
    - Prediction, classification, decision making, etc.
- Threats to dependability
  - Outputs of ML models are generally uncertain and very sensitive to input data
  - ML models can be fooled easily (e.g. by adversarial examples)

Related studies

- Improving the robustness of ML models
  - Adversarial learning [Goodfellow et al. 2014]
  - Safety verification [Huang et al. 2017]
  - Robust optimization method [Mądry et al. 2017]
  - …

- White-box testing method for ML system
  - DeepXplore [Pai et al. 2017]

- Falsifying the execution of ML models
  - Falsification framework for CPS [Dreossi. 2017]

Our approach: N-version architecture

Different versions of ML models are used in a system to improve the output reliability

- Focus
  - Not on training a robust model
  - But on reliable system processing with multiple ML models whose outputs are probably inaccurate
- Approach
  - Taking a multi-version system architecture
  - Exploiting the diversity of ML models and input data
    - Even if a ML model fails to recognize a red light, another model can recognize it accurately

Contributions

- Our study formally first defines two types of diversity (*model diversity* and *input diversity*) that should be considered in N-version ML architecture

- We present a reliability model for N-version architecture with the diversity metrics

- Our numerical results on the reliability model shows that the *combination of two diversities* can achieve the best system reliability

Outline

1. Background
2. **N-version machine learning architecture**
3. Reliability model
4. Numerical example
5. Conclusion

N-version ML models

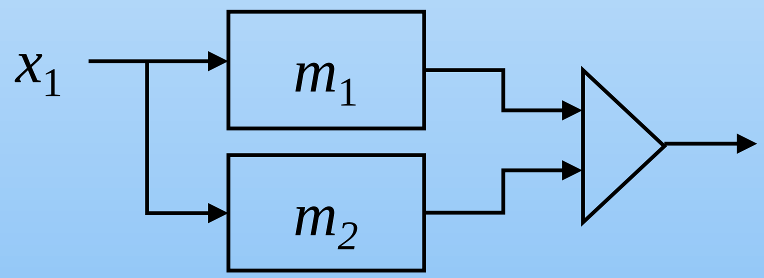
Motivated from N-version programming

	N-version programming	N-version ML
Target	Software program (generated from specification)	ML module (constructed from data)
Mitigation for	Software faults	Prediction errors
Components to use	More than two functionally equivalent programs from the same specification	More than two ML models for the same task
Sources of diversity	Development teams, programming languages, libraries and tools, etc.	ML algorithms, hyper parameters and input data

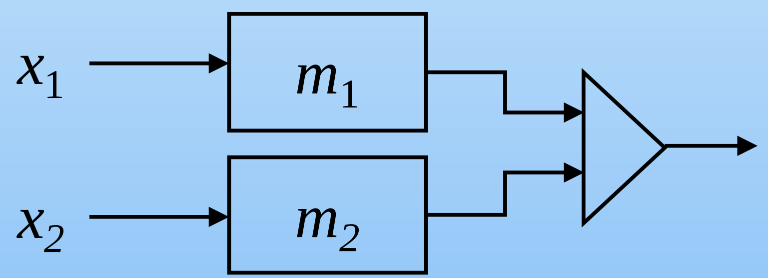
Two-version architecture

Use two independent versions of ML models

Double model with single input (DMSI)



Double model with double input (DMDI)

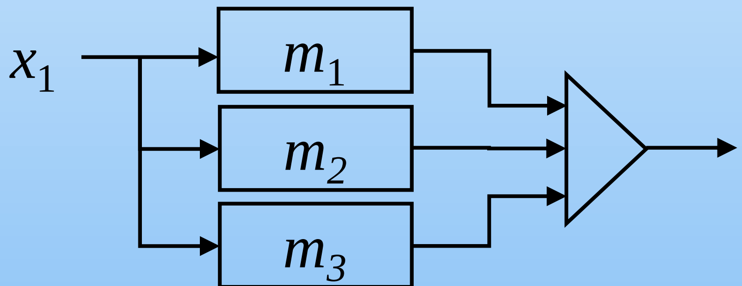


- The system fails when either module do not output expected answer (e.g., red signal)

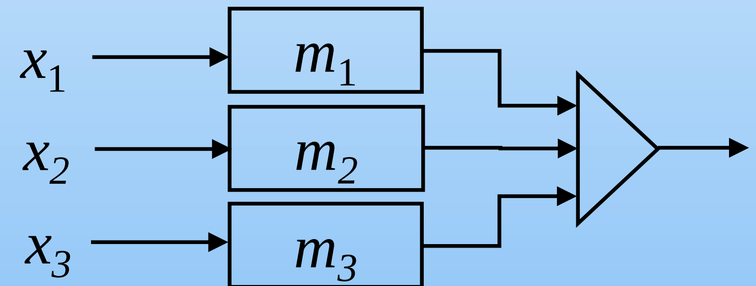
Three-version architecture

Use three versions with majority voting

Triple model with single input (TMSI)



Triple model with triple input (TMTI)

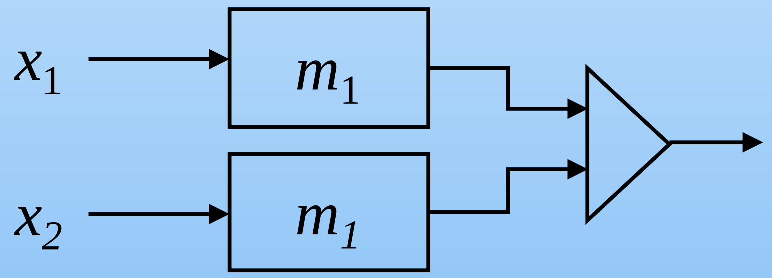


- The system fails when more than two modules output errors (by majority voting)

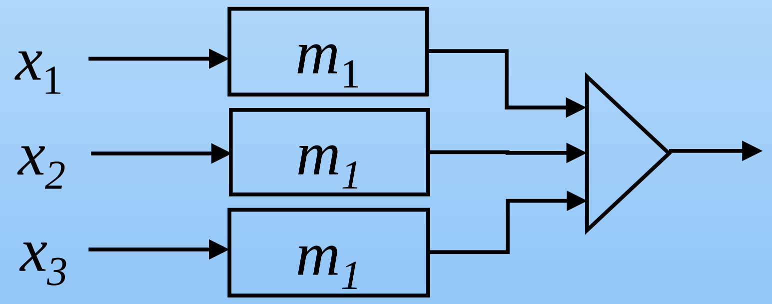
Single model architecture

Use the same model in parallel with different inputs

Single model with double input (SMDI)



Single model with triple input (SMTI)



- SMDI fails when both outputs are errors
- SMTI fails when more than two modules output errors

Outline

1. Background
2. N-version machine learning architecture
3. **Reliability model**
4. Numerical example
5. Conclusion

Notations

- System reliability
  - Probability that the output of the system is correct
  - $R_{i,j}$: Reliability of ML system with i versions and j diverse inputs

- Probability of error output
  - f_k : Probability that the ML model m_k outputs error

$$f_k = \frac{|E_k|}{|S|}$$

The set of input data that leads to output error by m_k

Total sample space of inputs in a given context

Definition of diversity

Intersection of errors (model diversity)

Let E_1 and E_2 be the subsets of input space S that make models m_1 and m_2 output errors, respectively. Define the intersection of errors $\alpha_{1,2} \in [0,1]$ as the ratio of the intersection over the smaller the size of E_1 and E_2.

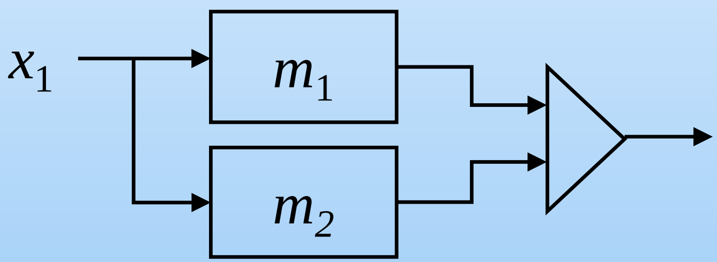
$$\alpha_{1,2} = \frac{|E_1 \bigcap E_2|}{\min\{|E_1|, |E_2|\}}.$$

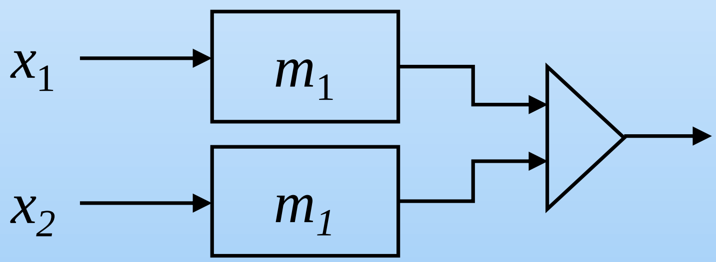
Conjunction of errors (input diversity)

Let x_1 and x_2 be the inputs from the same sample space S to model m_1. Define the conjunction of errors $\beta_1 \in [0,1]$ as the probability that m_1 outputs error by x_2 provided that m_1 outputs error by x_1.

$$\beta_1 = \Pr[x_2 \in E_1 | x_1 \in E_1].$$

Reliabilities of DMSI and SMDI

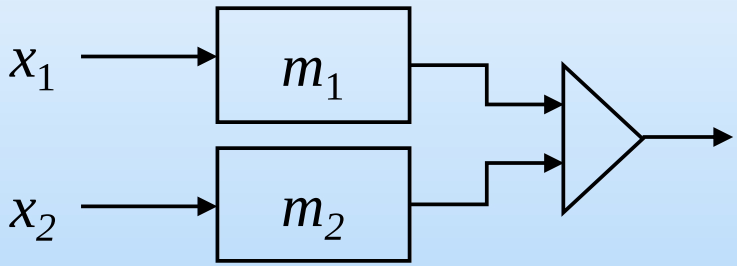
<table>
<tr><th></th><th>DMSI</th><th>SMDI</th></tr>
<tr><td></td><td>

Model diversity</td><td>

Input diversity</td></tr>
<tr><td>Failure probability</td><td>$f_{DMSI}(m_1, m_2)$
$= \frac{|E_1 \cap E_2|}{|S|}$
$= \alpha_{1,2} \cdot \frac{min\{|E_1|, |E_2|\}}{|S|}$</td><td>$f_{SMDI}(m_1)$
$= Pr[x_1 \in E_1, x_2 \in E_1]$
$= \Pr[x_2 \in E_1 | x_1 \in E_1]$
$\cdot Pr[x_1 \in E_1]$
$= \beta_1 \cdot f_1$</td></tr>
<tr><td>Reliability</td><td>$R_{2,1}(m_1, m_2) = 1 - \alpha_{1,2} \cdot f_1$</td><td>$R_{1,2}(m_1) = 1 - \beta_1 \cdot f_1$</td></tr>
</table>

*) we assume $|E_1| \leq |E_2|$

Reliability of DMDI



Model diversity & input diversity

Failure probability

$$f_{DMDI}(m_1, m_2) = \Pr[x_1 \in E_1, x_2 \in E_2]$$
$$= \Pr[x_2 \in E_2 | x_1 \in E_1] \cdot \Pr[x_1 \in E_1]$$

- When x_2 has conjunction with x_1

$$\Pr[x_2 \in E_1 | x_1 \in E_1] \cdot \Pr[x_2 \in E_2 | x_2 \in E_1] = \beta_1 \cdot \alpha_{1,2} \cdot \min\{f_1, f_2\} / f_1$$

- When x_2 has no conjunction with x_1

$$\Pr[x_2 \in \overline{E_1} | x_1 \in E_1] \cdot \Pr[x_2 \in E_2 | x_2 \in \overline{E_1}] = (1 - \beta_1) \cdot \frac{f_2 - \alpha_{1,2} \cdot \min\{f_1, f_2\}}{1 - f_1}$$

$$\therefore\ f_{DMDI}(m_1, m_2) = \left[\beta_1 \cdot \alpha_{1,2} + (1 - \beta_1) \cdot \frac{f_2 - \alpha_{1,2} \cdot f_1}{1 - f_1}\right] \cdot f_1$$

Reliability

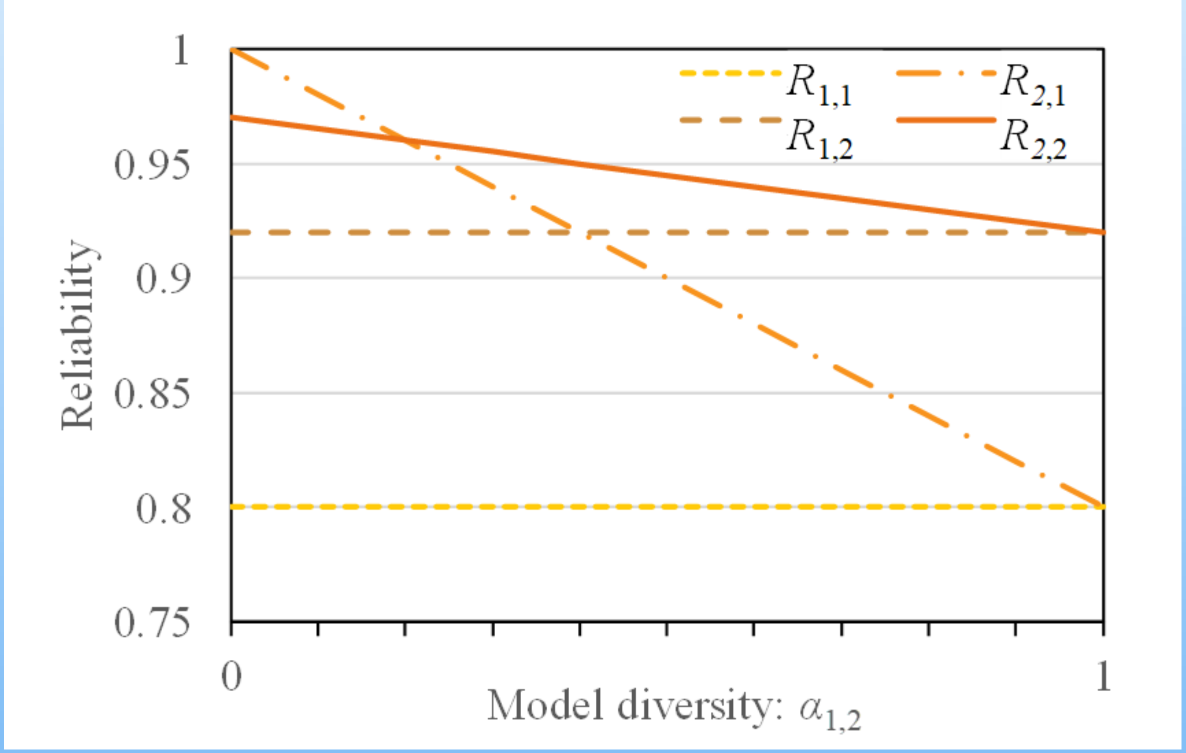
$$R_{2,2}(m_1, m_2) = 1 - \left[\left(\beta_1 - f_1\right) \cdot \alpha_{1,2} + f_2\right] \cdot f_1$$

Outline

1. Background
2. N-version machine learning architecture
3. Reliability model
4. **Numerical example**
5. Conclusion

Reliability impacts of model diversity

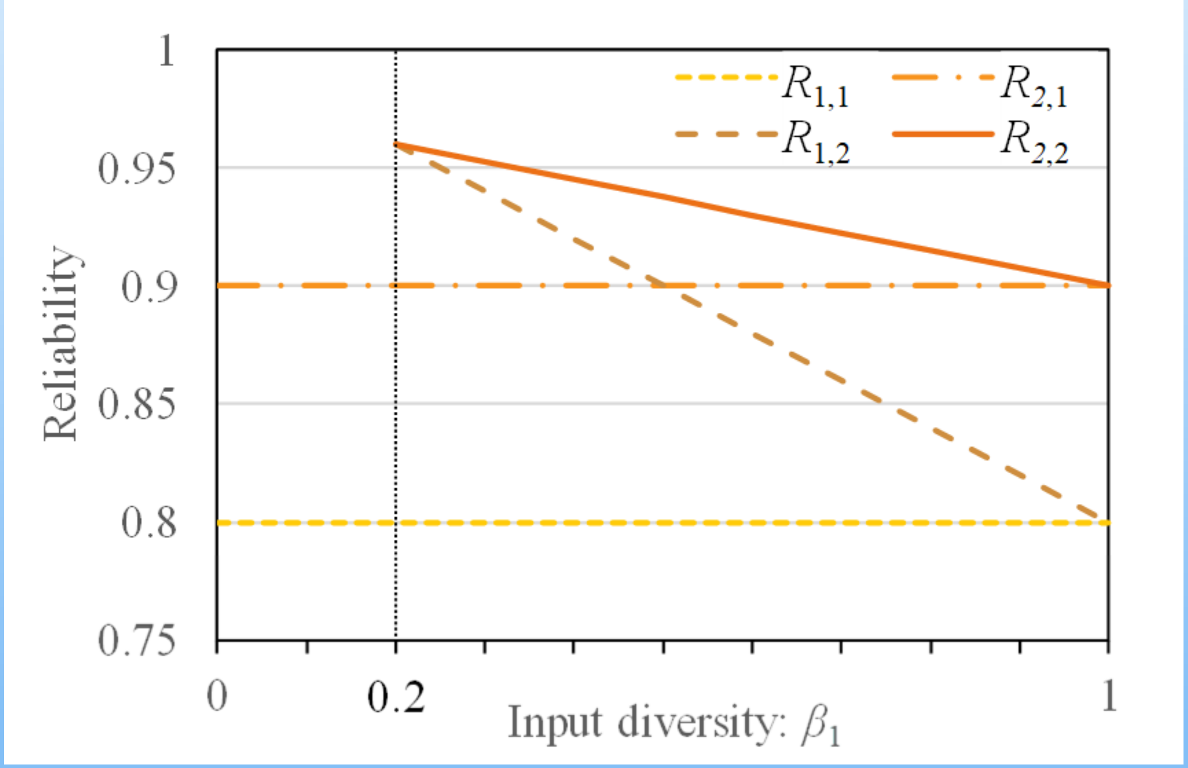
- Varying $\alpha_{1,2}$ with $f_1 = f_2 = 0.2$, and $\beta_1 = 0.4$



- $R_{2,1}$ achieves complete reliability when two models do not have intersection (i.e., $\alpha_{1,2}=0$)
- $R_{2,2}$ generally achieves better reliability

Reliability impacts of input diversity

- Varying β_1 with $f_1 = f_2 = 0.2$, and $a_{1,2} = 0.5$



- When $\beta_1 = 0.2$ $(=f_1)$, there is no conjunction and two modules output errors independently
- As β_1 increases, both $R_{1,2}$ and $R_{2,2}$ decrease

Conclusion

- For N-version machine learning architecture, two types of diversity are formally presented
- Numerical example on the proposed reliability model show that both diversities contribute to improve two-version architecture
- Future work will address the empirical study to show the reliability improvement by N-version architecture

Q & A

Thank you!