Challenges of Multimodal Fusion and Fact-checking

Laure Berti-Equille

IRD, ESPACE-DEV
Montpellier, France

https://laureberti.github.io/website/
laure.berti@ird.fr

Institut de Recherche pour le Développement
FRANCE

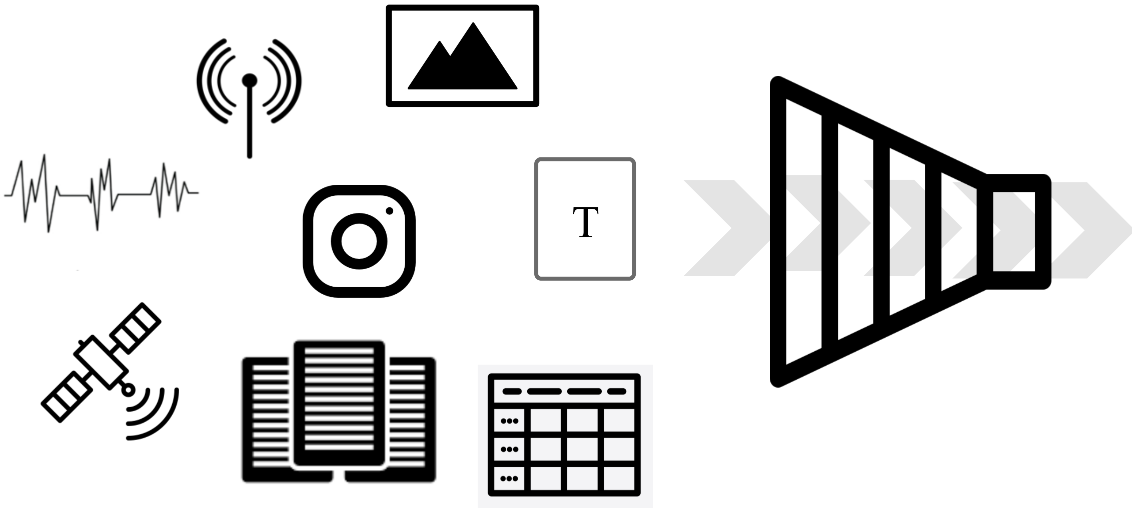
Intelligent Information Systems Seminar @FOI on June 4, 2024

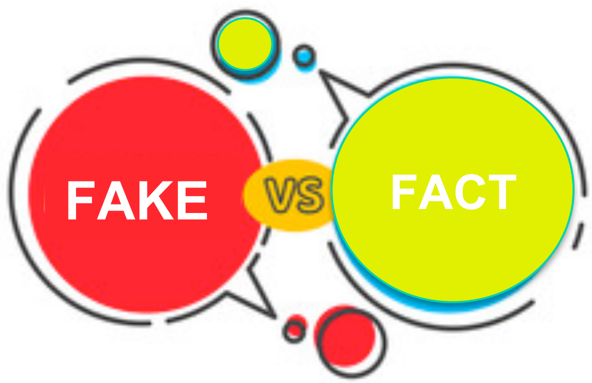
Outline

1. Motivations
2. Challenges in MML & LLM
3. Methods & Contributions
4. Conclusions

Motivations (1/4)

- Handling many information sources with various modalities
- Explain fact-checking output without overconfident decisions



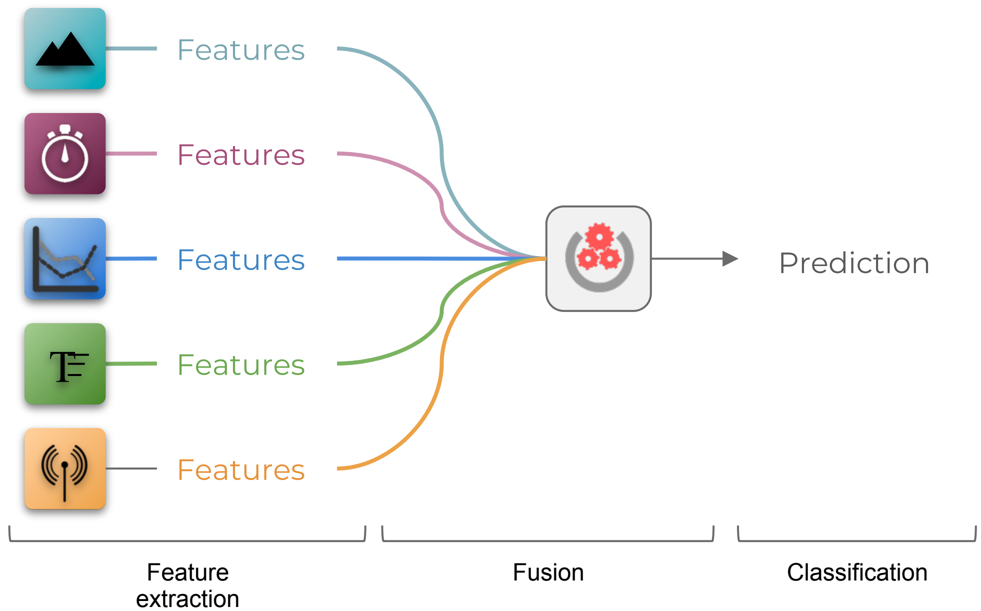


©

We need multimodal models

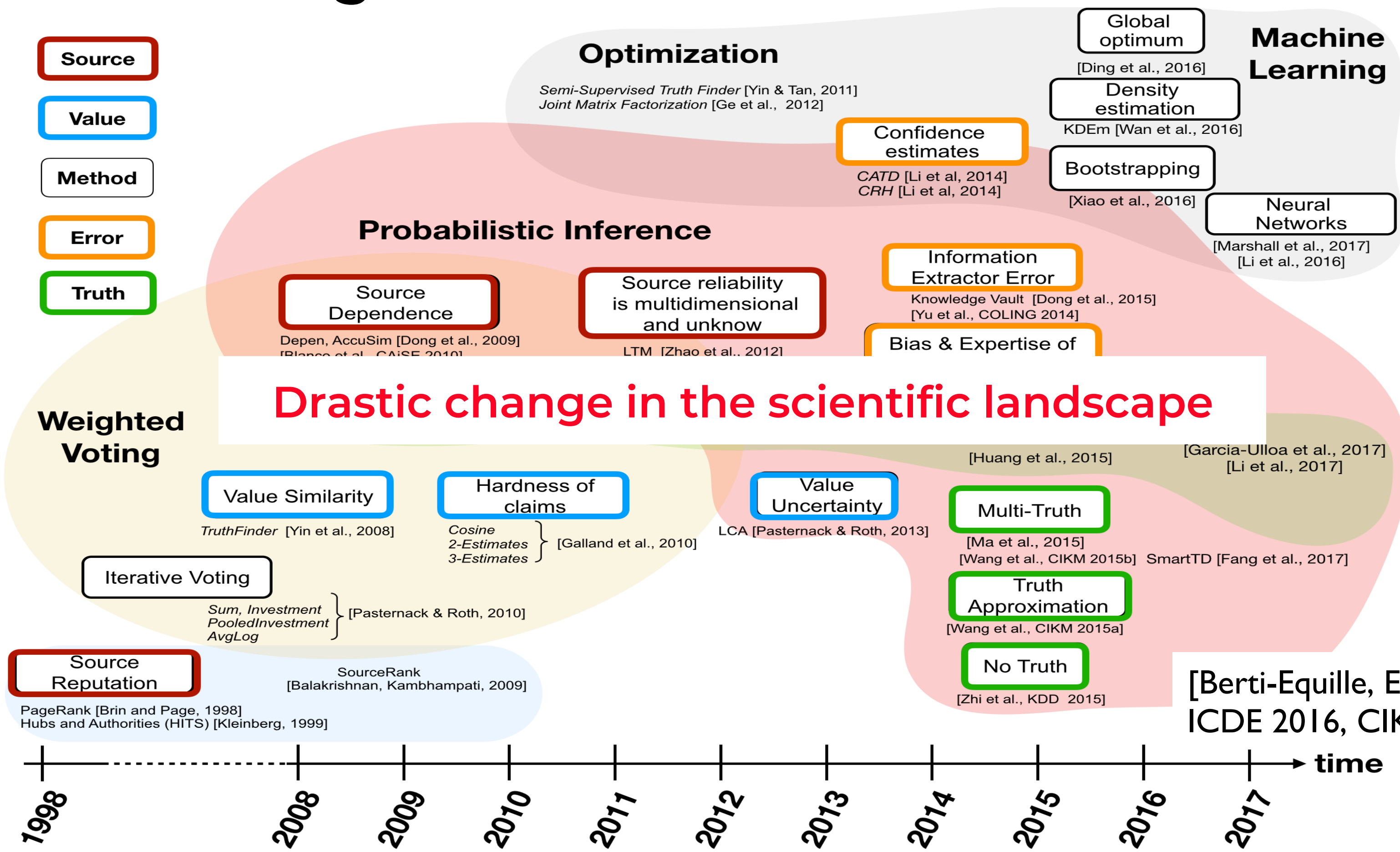
We need to take into account uncertainty and provide chains-of-evidences

Multimodal fact-checking: A classification problem



Goal: Fuse modalities efficiently to achieve high quality performance

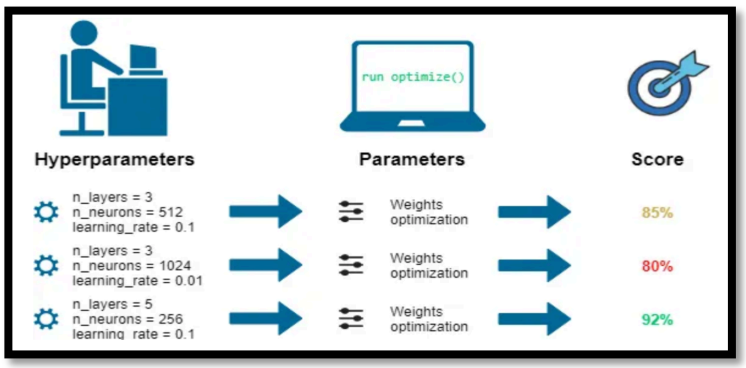
Truth finding from structured data before LLMs & MML



[Berti-Equille, Encyclopedia 2018, ICDE 2016, CIKM 2017 Tutorials]

Motivations (4/4)

- Ensure stable hyperparameter optimization for ML-based fact-checking



- Ensure resilience to multimodal data poisoning attacks

- Trace back LLM pre-training, fine-tuning, and prompt engineering

We need reproducibility

We need traceability and explainability

Theoretical, Technical, and Experimental Challenges

Multimodal Deep Learning

- Complex models
- Costly training
- Hard to communicate to non-experts

MM Network Architecture Search

Cost-driven and frugal models

Beyond XAI: Chain-of-evidences

(Multimodal) Uncertainty Quantification

- Quantify aleatoric and epistemic uncertainty in fact-checking
- Detect multimodal contradictions

image contradiction

Text/image contradiction

To use LLMs for fact-checking

We need to:

- Quantify **LLM hallucination & factuality** in perspective with the model/training size
- Detect **stereotype amplification** due to bias and low quality training corpus
- Evaluate **sensitivity** to prompt variations, noise, conflicting (multimodal) data or domain shift
- Evaluate LLM **vulnerability to adversarial attacks** (e.g., generated texts used in pretraining or prompts)
- Develop dedicated **benchmarks** and design **controlled experiments**

Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610-623. 2021.

Uncertainty Quantification in LLMs

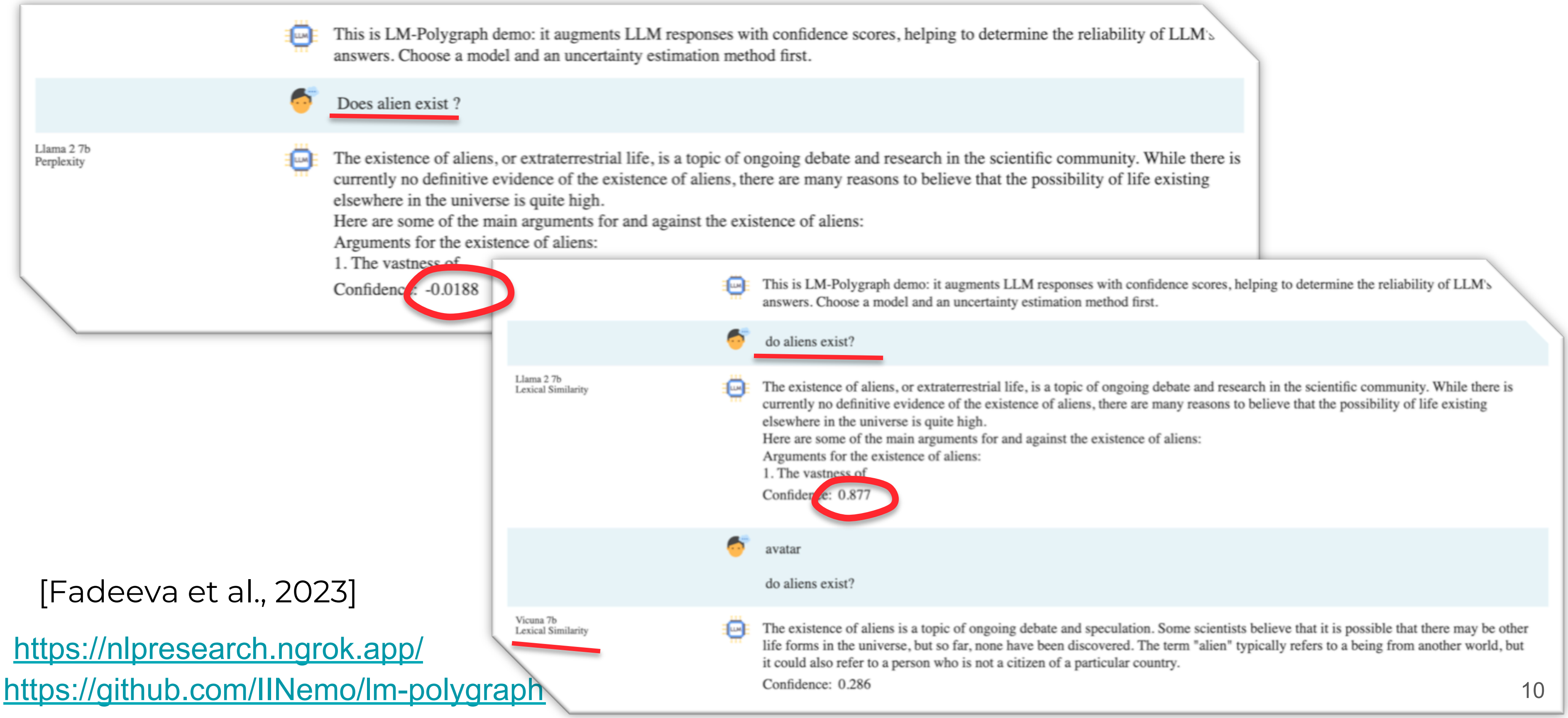
[Fadeeva et al., 2023]

Uncertainty Estimation Method	Type	Category	Compute	Memory	Need Training Data?
Maximum sequence probability	White-box	Information-based	Low	Low	No
Perplexity (Fomicheva et al., 2020)			Low	Low	No
Mean token entropy (Fomicheva et al., 2020)			Low	Low	No
Monte Carlo sequence entropy (Kuhn et al., 2023)			High	Low	No
Pointwise mutual information (PMI) (Takayama and Arase, 2019)			Medium	Low	No
Conditional PMI (van der Poel et al., 2022)			Medium	Medium	No
Semantic entropy (Kuhn et al., 2023)	White-box	Meaning diversity	High	Low	No
Sentence-level ensemble-based measures (Malinin and Gales, 2021)	White-box	Ensembling	High	High	Yes
Token-level ensemble-based measures (Malinin and Gales, 2021)			High	High	Yes
Mahalanobis distance (MD) (Lee et al., 2018)	White-box	Density-based	Low	Low	Yes
Robust density estimation (RDE) (Yoo et al., 2022)			Low	Low	Yes
Relative Mahalanobis distance (RMD) (Ren et al., 2023)			Low	Low	Yes
Hybrid Uncertainty Quantification (HUQ) (Vazhentsev et al., 2023a)			Low	Low	Yes
p(True) (Kadavath et al., 2022)	White-box	Reflexive	Medium	Low	No
Number of semantic sets (NumSets) (Lin et al., 2023)	Black-box	Meaning diversity	High	Low	No
Sum of eigenvalues of the graph Laplacian (EigV) (Lin et al., 2023)			High	Low	No
Degree matrix (Deg) (Lin et al., 2023)			High	Low	No
Eccentricity (Ecc) (Lin et al., 2023)			High	Low	No
Lexical similarity (LexSim) (Fomicheva et al., 2020)			High	Low	No

Table 1: UE methods implemented in LM-Polygraph.

https://github.com/IINemo/lm-polygraph

Uncertainty Quantification in LLMs: Lack of consistency



[Fadeeva et al., 2023]

https://nlpresearch.ngrok.app/
https://github.com/IINemo/lm-polygraph

Inadequacy of benchmarks for MM Fact-checking

Name	# Claims	# Labels	Data	Year
LIAR [4]	12836	6	Claim Text, Metadata (Speaker etc.)	2017
CREDBANK [8]	1049	5	Claim Text, Event, Topic	2015
The Lie Detector [9]	600	2	Claim Text	2009
Claim matching beyond english [10]	2343	3	Claim Text Pairs	2021
FEVER [1]	185445	3	Claim Text, Document Text	2018
MultiFC [12]	36534	40	Claim Text, Document url, Metadata	2019
Fakeddit [13]	1 million	2/3/6	Claim Text, Claim image	2019
Covid-19 Fake News dataset [11]	10700	2	Claim Text	2020
FakeNewsNet [14]	23921	2	Claim Text, Spatiotemporal info	2019
Whatsapp fact-checking dataset [15]	1032	3	Claim Image, Metadata	2020
Factify (ours)	50000	5	Claim Text, Claim Image, Document Text, Document Image, Images OCR	2021

Table 1
Details of related public datasets for automated fact-checking along with available meta data and release year.

Mishra et al. FACTIFY: A Multi-Modal Fact Verification Dataset, De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. https://ceur-ws.org/Vol-3199/paper18.pdf

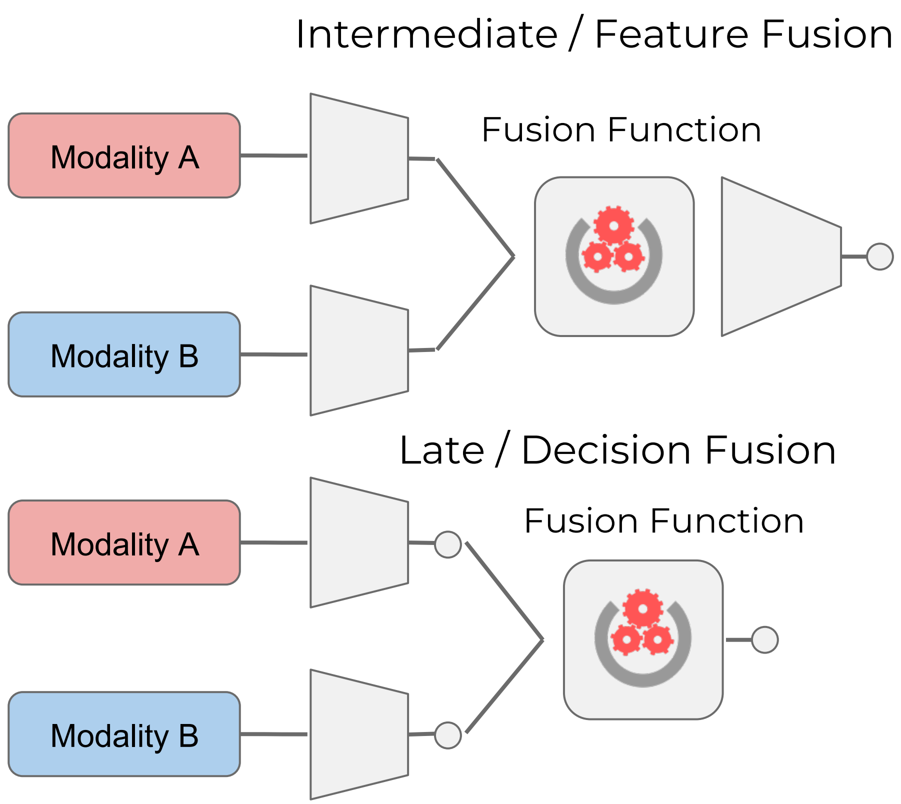
To address the challenges in MML

1. Design adaptive, conceptually, computationally simple, scalable multimodal deep learning architecture

2. Quantify the uncertainties in all stages of the pipeline (data / fusion / decision) and use them in multimodal fusion and fact-checking classification

3. Provide logical, evidence-based explanations of the results to practitioners for easier adoption and trust

Multimodal Data Fusion



- Design and test architectures that are:
  - Conceptually / Computationally simple
  - Scalable to many modalities
  - Easy to adapt to different modalities
  - Easy to interpret and communicate
- Introduce Uncertainty into fusion
- Test the designed approach on real world data

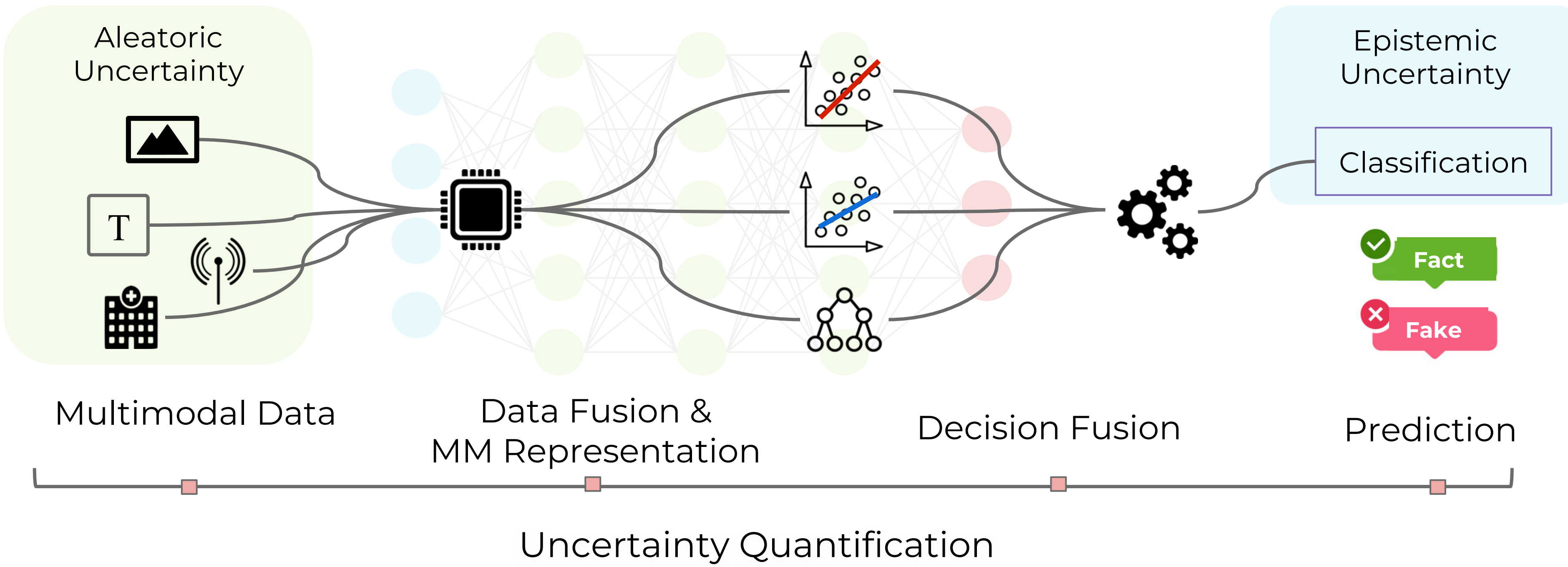
Our Contributions to MM Fusion

1. Use MLP-Mixers to design conceptually and computationally simple approach, with decent performance for MM fusion

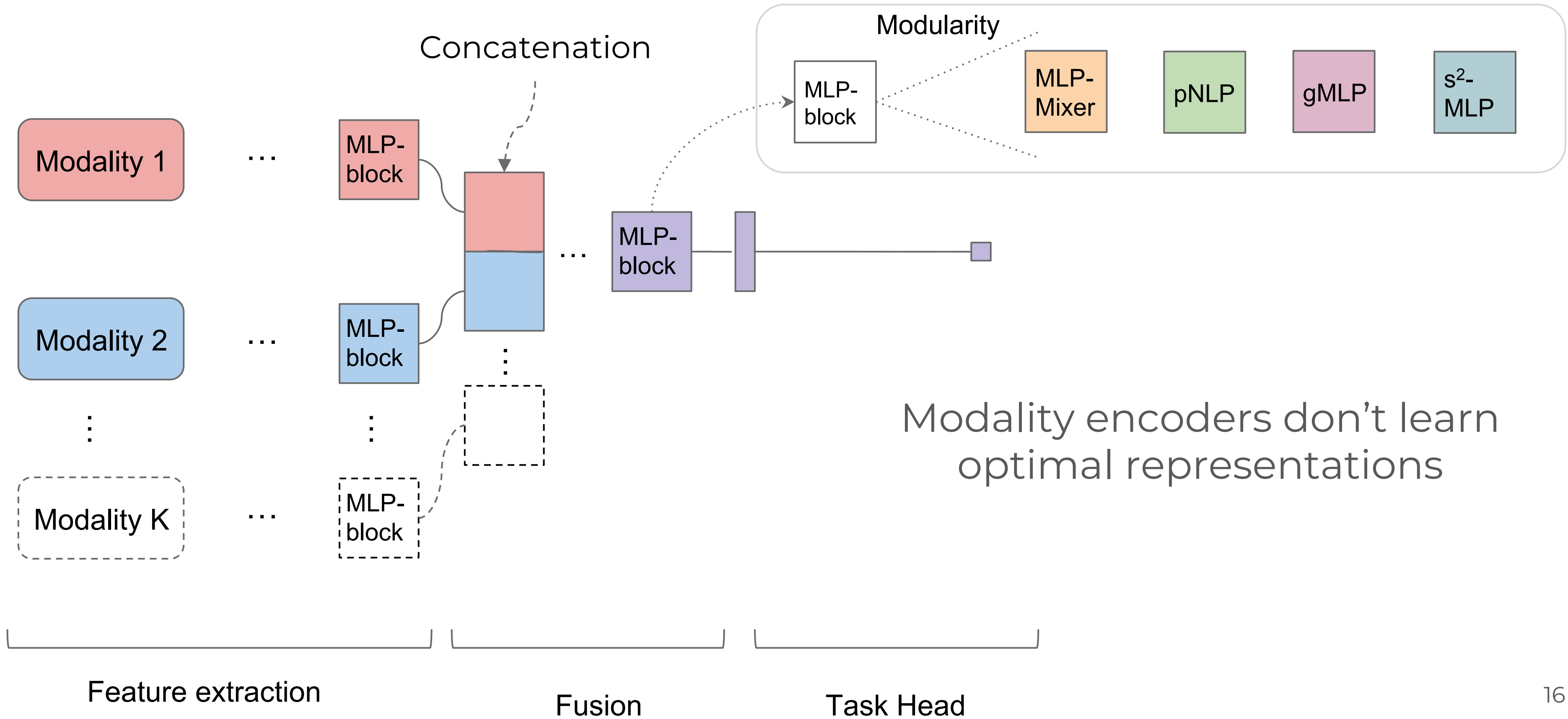
2. Force modalities to learn optimal representation with little to no time overhead

Grigor Bezirganyan, Sana Sellami, Laure Berti-Équille, Sébastien Fournier:
M2-Mixer: A Multimodal Mixer with Multi-head Loss for Classification from Multimodal Data. IEEE Big Data 2023: 1052-1058

Data and Decision Fusion with Uncertainty Quantification



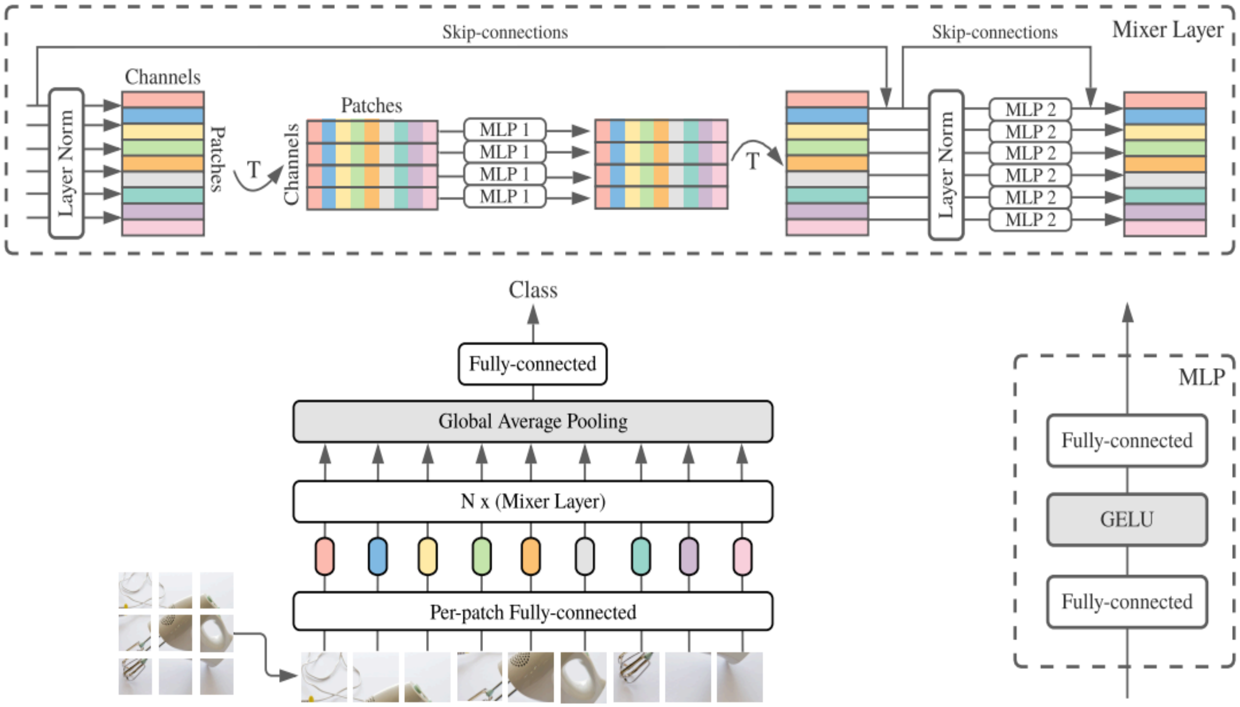
Multimodal Mixer: Use MLP-Blocks for Feature Extraction and Fusion



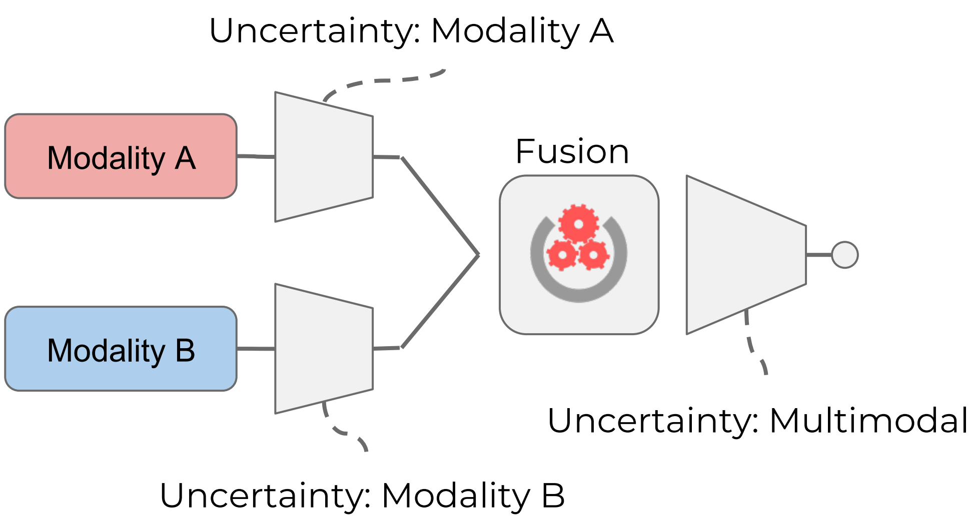
Modality encoders don't learn optimal representations

Application of MLP-Mixers

- ✅ Only MLPs
  [Tolstikhin et. al, 2021]
- ✅ Conceptually / Computationally Simple
- ✅ Competitive with Transformers / CNNs



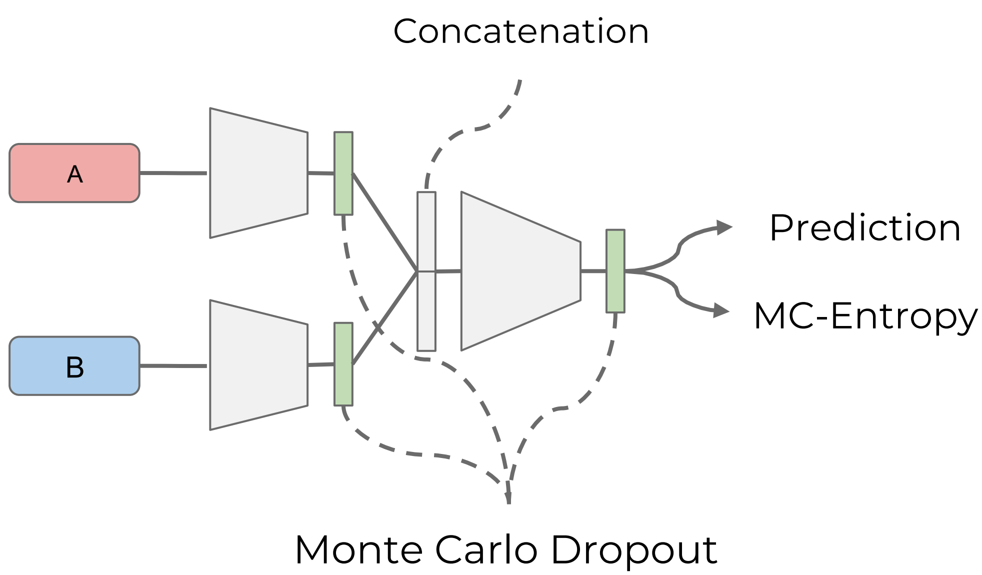
Multimodal Uncertainty Quantification



- Understand the propagation of uncertainty across multimodal network
- Use quantified uncertainty to improve multimodal fusion
- Reject uncertain predictions

Design Experiments with UQ-based Rejection

How can epistemic (model) uncertainty help to reject uncertain decisions ?

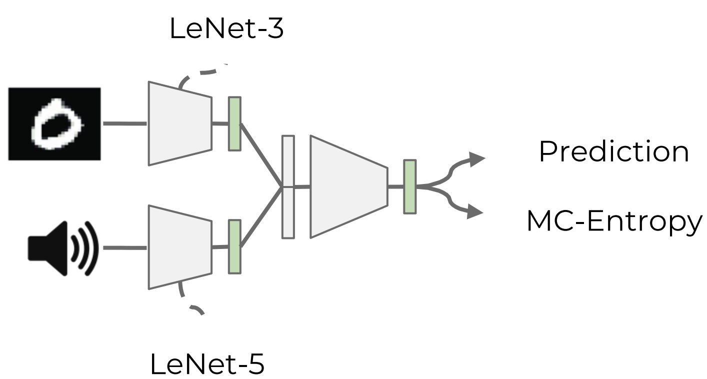


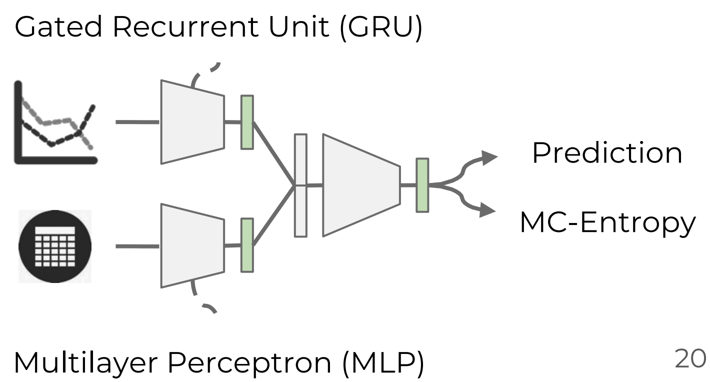
- Choose a threshold K
- Reject prediction if high entropy
- Accept prediction if low entropy

Training | Testing

Benchmarking Datasets and Models

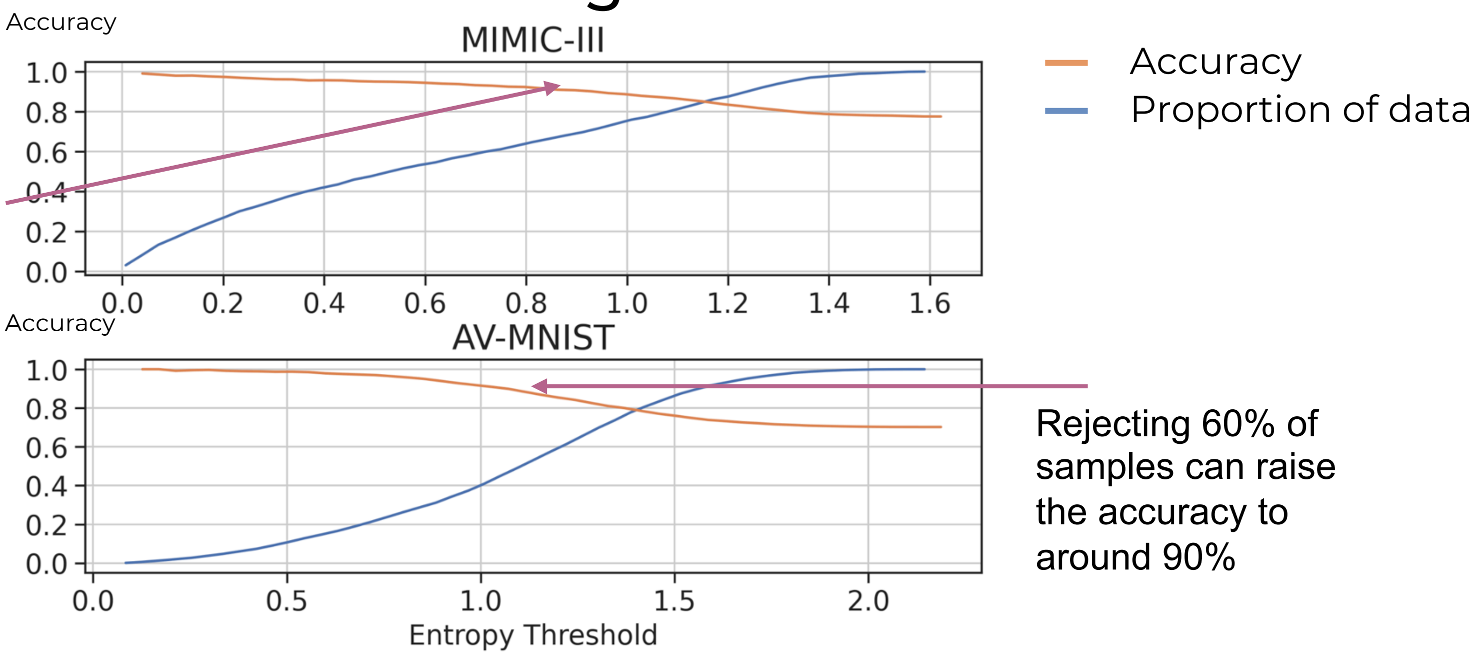
Dataset	AV-MNIST	MIMIC-III
Field	Multimedia	Healthcare
Modalities	Image / Audio	Time Series / Tabular
Samples	70,000	32,616
Parameters	267,000	33,700





Preliminary Results of Uncertainty-based Rejection (1/2) at training time

Rejecting 30% of samples can raise the accuracy to around 90%



Rejecting 60% of samples can raise the accuracy to around 90%

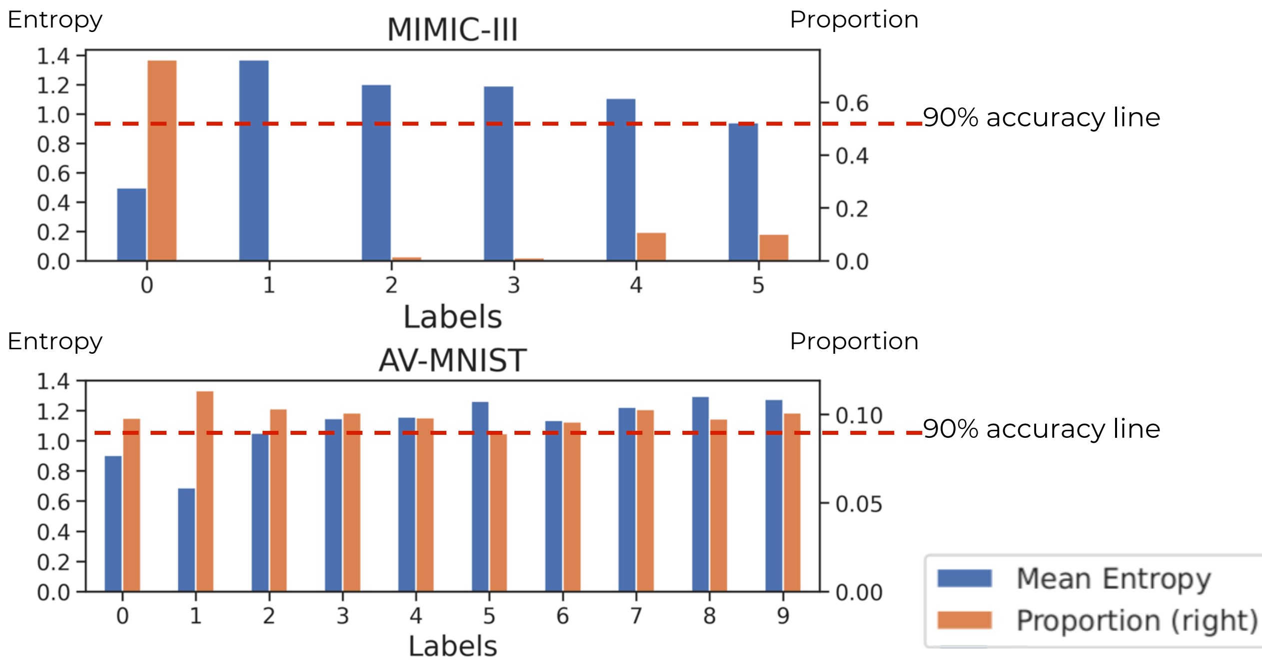
We need to find ways to reduce total uncertainty, without sacrificing most of our data.

Preliminary Results of Uncertainty-based Rejection (2/2)

Epistemic uncertainty is high when we have data scarcity

On MIMIC-III, the model is certain only on one class.

On AV-MNIST, we have more balanced dataset, so we have more balanced uncertainty



UQ can also help to debug the model at test time

Ongoing work

- Test multimodal NAS with various multimodal datasets
- Analyze uncertainty propagation across multimodal network layers
- Use UQ for improving multimodal fusion and decision
- Integrate uncertainty-based rejection
- Test the model on multimodal fact-checking datasets

Thanks!

References

1. Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., ... & Zhang, Y. (2023). Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv preprint arXiv:2310.07521
2. Munn, L., Magee, L., & Arora, V. (2023). Truth Machines: Synthesizing Veracity in AI Language Models. arXiv preprint arXiv:2301.12066.
3. Fadeeva, E., Vashurin, R., Tsvigun, A., Vazhentsev, A., Petrakov, S., Fedyanin, K., ... & Shelmanov, A. (2023). LM-Polygraph: Uncertainty Estimation for Language Models. arXiv preprint arXiv:2311.07383.
4. James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In Proc. of the 2019 EMNLP-IJCNLP, pages 2944–2953, Hong Kong, China. https://aclanthology.org/D19-1292
5. Atanasova, P., Wright, D., & Augenstein, I. (2020). Generating label cohesive and well-formed adversarial claims. arXiv preprint arXiv:2009.08205. https://github.com/copenlu/fever-adversarial-attacks
6. Gao, J., Hoffmann, H. F., Oikonomou, S., Kiskovski, D., & Bandhakavi, A. (2021). Logically at Factify 2022: Multimodal Fact Verification. arXiv preprint arXiv:2112.09253. https://arxiv.org/abs/2112.09253
7. Verschuuren, P. J., Gao, J., van Eeden, A., Oikonomou, S., & Bandhakavi, A. (2023). Logically at Factify 2023: A Multi-Modal Fact Checking System Based on Evidence Retrieval techniques and Transformer Encoder Architecture. arXiv preprint arXiv:2301.03127. https://arxiv.org/abs/2301.03127
8. FEVER 2.0 Adversarial Attacks Dataset, https://fever.ai/dataset/adversarial.html
9. Defactify Workshop https://aiisc.ai/defactify/
10. Factify Multi-Modal Fact Verification dataset, https://competitions.codalab.org/competitions/35153
11. Dan Saattrup Nielsen and Ryan McConville. “MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset.” arXiv preprint arXiv:2202.11684 (2022). https://mumin-dataset.github.io/
12. L Berti-Equille, ML Ba. Veracity of big data: challenges of cross-modal truth discovery. Journal of Data and Information Quality (JDIQ) 7 (3), 1-3