Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning? (Provably)

Yu Huang[1], Junyang Lin[2], Chang Zhou[2], Hongxia Yang[2], Longbo Huang[1]

[1] IIIS, Tsinghua University *[2]DAMO Academy, Alibaba Group*

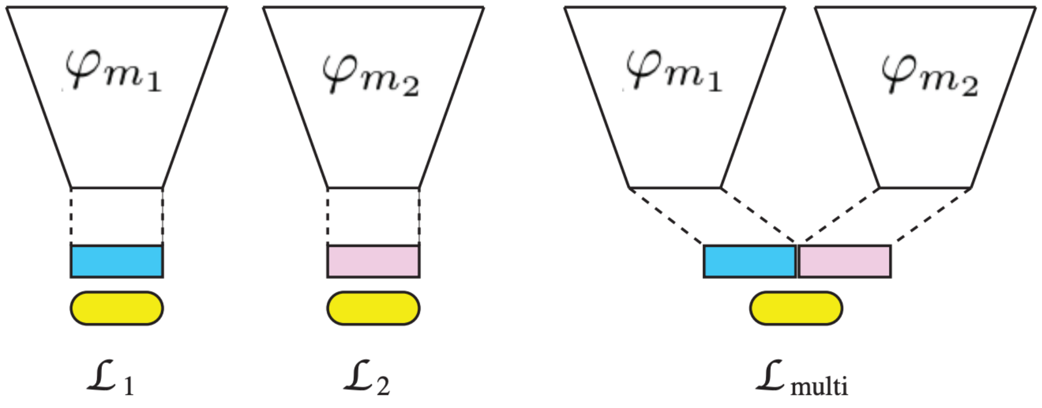
The remarkable success of deep multimodal learning

DALL·E 2
Text: *An astronaut riding a horse in a photorealistic style*

- Common belief: multi-modal is better than single since multiple signals generally bring more information. [Huang et al, 2021]
- However: the use of multimodal data in practice will reduce the performance of the model in some cases [Gat et al, 2020] [Han et al, 2021]

[1] Pictures from https://openai.com/dall-e-2/

Uni-modal networks consistently **outperform** multimodal networks in Practice [Wang et al, 2020]



Uni-modal v.s. Naive Joint Training

Dataset	Multi-modal	V@1	Best Uni	V@1	Drop
Kinetics	A + RGB	71.4	RGB	**72.6**	-1.2
	RGB + OF	71.3	RGB	**72.6**	-1.3
	A + OF	58.3	OF	**62.1**	-3.8
	A + RGB + OF	70.0	RGB	**72.6**	-2.6

- across different combinations of modalities and on different tasks and benchmarks

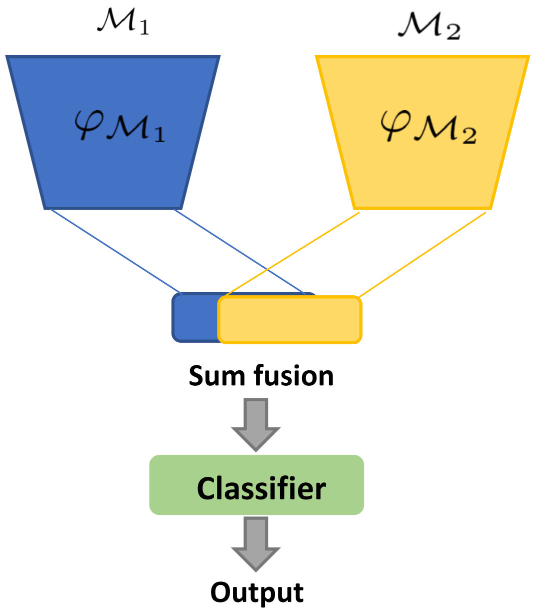
Our goal: theoretically explain this performance drop

Why Previous Analysis Cannot Explain?

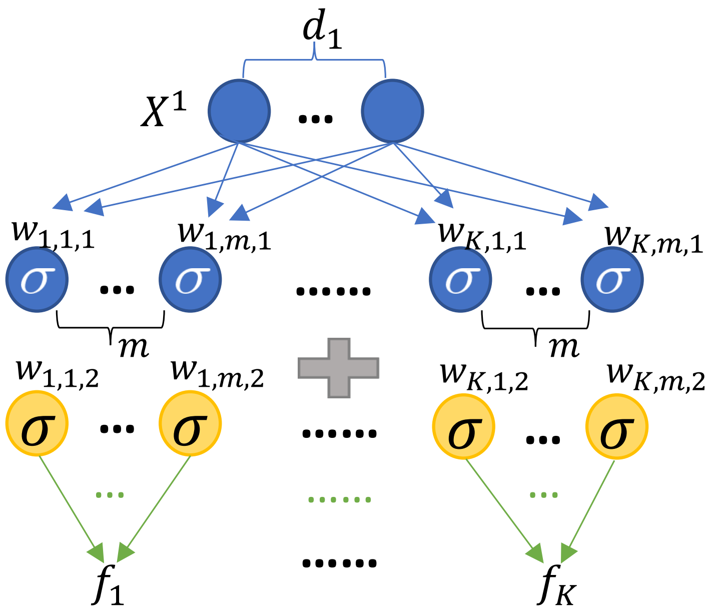
- Previous analysis: focus on the **generalization** side
- Cause: **Optimization issue**
- Recent efforts: [Du et al, 2021] do not consider neural networks architecture
- Our results: first theoretical treatment towards the degenerating aspect of multi-modal learning in **neural networks**

Setups

Late fusion framework



Learner Network



1. K-class classification
2. Each modality is generated from a sparse coding model:

$$\mathbf{X}^1 = \mathbf{M}^1 z^1 + \xi^1, \quad \mathbf{X}^2 = \mathbf{M}^2 z^2 + \xi^2$$
$$(z^1, z^2) \sim \mathcal{P}_z \quad \xi^r \sim \mathcal{P}_{\xi^r} \text{ for } r \in [2]$$

where z^1 and z^2 are sparse vectors and share some similarities.

3. Modality encoder: one-layer neural network, activated by smoothed ReLU

When only single modality is applied to training

- The uni-modal network will **focus on** learning the modality-associated features, which leads to good performance.

- Training error is zero:

$$\frac{1}{n}\sum_{(\mathbf{X}^r,y)\in\mathcal{D}^r}\mathbb{I}\left\{\exists j\neq y: f_y^{\mathrm{uni},r(T)}(\mathbf{X}^r)\leq f_j^{\mathrm{uni},r(T)}(\mathbf{X}^r)\right\}=0.$$

- The test error satisfies:

“Insufficient Structure”

$$\Pr_{(\mathbf{X}^r,y)\sim\mathcal{P}^r}\left(\exists j\neq y: f_y^{\mathrm{uni},r(T)}(\mathbf{X}^r)\leq f_j^{\mathrm{uni},r(T)}(\mathbf{X}^r)\right)=(1\pm o(1))\mu_r$$

When naive joint training is applied

- The neural network will not efficiently learn all features from different modalities:
  - Training error is zero:

$$\frac{1}{n}\sum_{(\mathbf{X},y)\in\mathcal{D}} \mathbb{I}\{\exists j \neq y : f_y^{(T)}(\mathbf{X}) \leq f_j^{(T)}(\mathbf{X})\} = 0.$$

  - For $r \in [2]$, with probability $p_{3-r} > 0$, the test error of $f^{r(T)}$ is high:

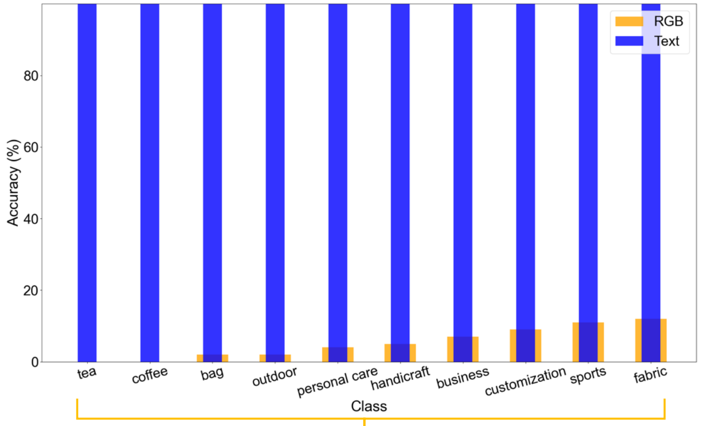
$$\Pr_{(\mathbf{X}^r,y)\sim\mathcal{P}^r}(\exists j \neq y : f_y^{r(T)}(\mathbf{X}^r) \leq f_j^{r(T)}(\mathbf{X}^r)) \geq \frac{1}{K}$$

where $p_1 + p_2 = 1 - o(1)$, and $p_r \geq m^{-O(1)}$, $\forall r \in [2]$.

Modality Competition:
multiple modalities will **compete** with each other. Only a subset of modalities that correlate more with their encoding network's random initialization will win.

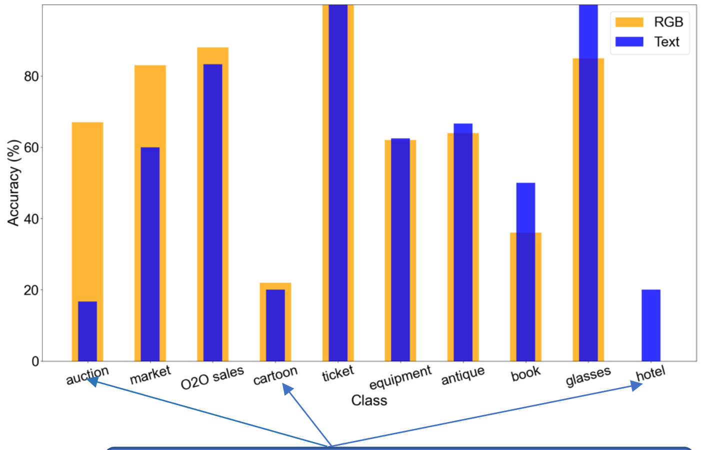
Insufficient Structure

Top 10 Improved Class Accuracy

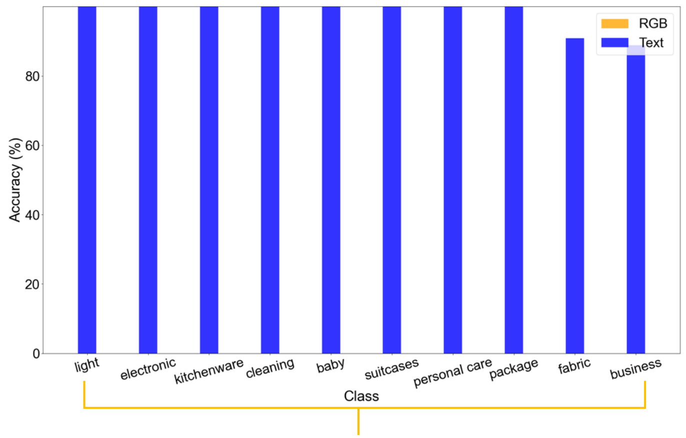


"Insufficient" structure for RGB

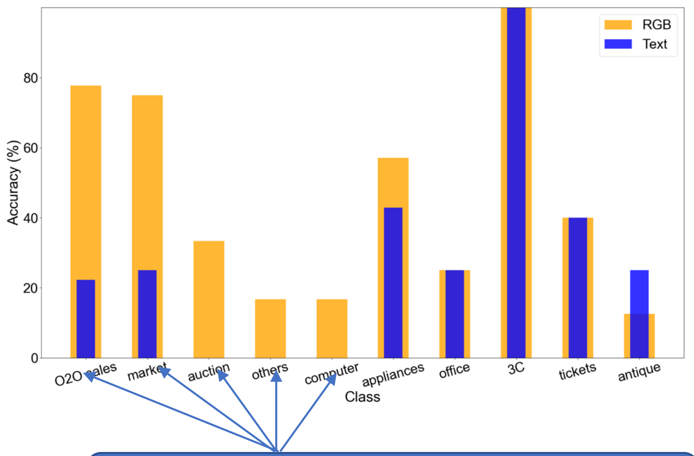
Top 10 Dropped Class Accuracy



"Insufficient" structure for Text



Evidence that RGB has not been learned



Text loses the competition and has not been explored

Original intention: the information provided by the remaining sufficient modalities can assist

Our results reveal: the modal not only fails to exploit the extra modalities, but also loses the expertise of the original modality.

Thanks!

References

- [1] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao and Longbo Huang. What Makes Multi-modal Learning Better than Single (Provably). NeurIPS 2021
- [2] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? CVPR 2020
- [3] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Yue Wang, Yang Yuan, and Hang Zhao. Modality laziness: Everybody's business is nobody's business. 2021.
- [4] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. arXiv preprint arXiv:2010.10802, 2020.
- [5] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. arXiv preprint arXiv:2102.02051, 2021