A Closer Look at Multimodal Representation Collapse

ICML
International Conference On Machine Learning
(Spotlight)




Abhra Chaudhuri [1] Anjan Dutta [2] Tu Bui [1] Serban Georgescu [1]

[1] FUJITSU

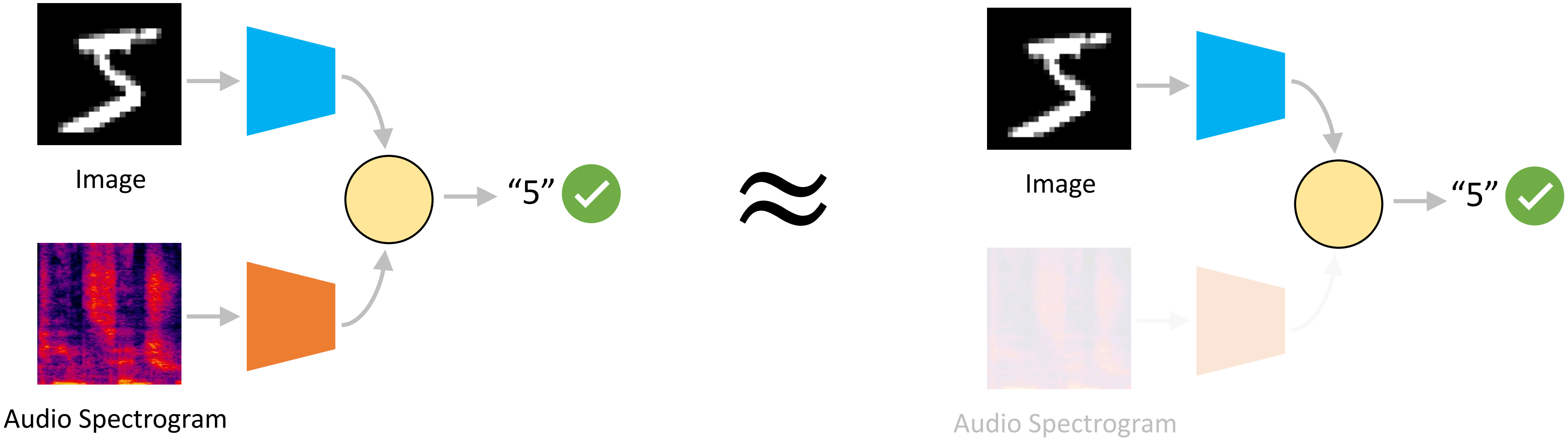
[2] UNIVERSITY OF SURREY

Outline

- Motivation
- Understanding Modality Collapse
- The Effect of Knowledge Distillation
- Explicit Basis Reallocation
- Experiments
- Conclusion and Open Problems

Outline

Modality Collapse



Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia

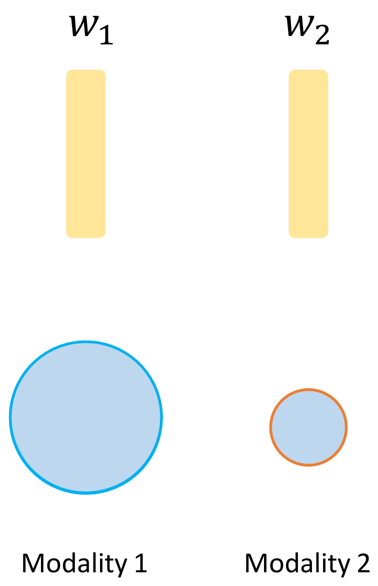
Modality Collapse

Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia

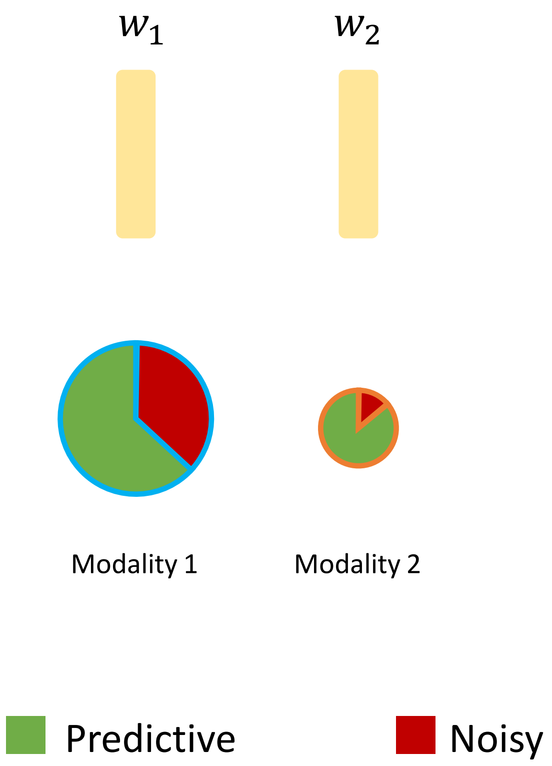
Outline

- Motivation
- Understanding Modality Collapse
- The Effect of Knowledge Distillation
- Explicit Basis Reallocation
- Experiments
- Conclusion and Open Problems

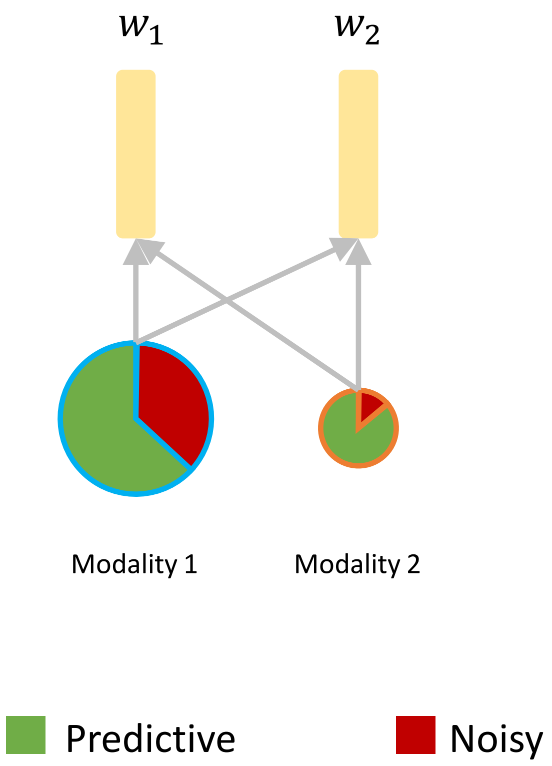
Cross-Modal Collisions due to Polysemanticity



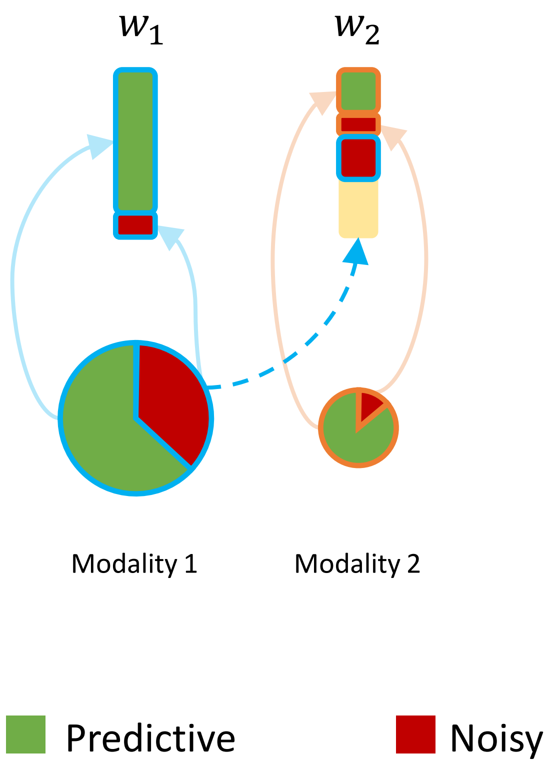
Cross-Modal Collisions due to Polysemanticity



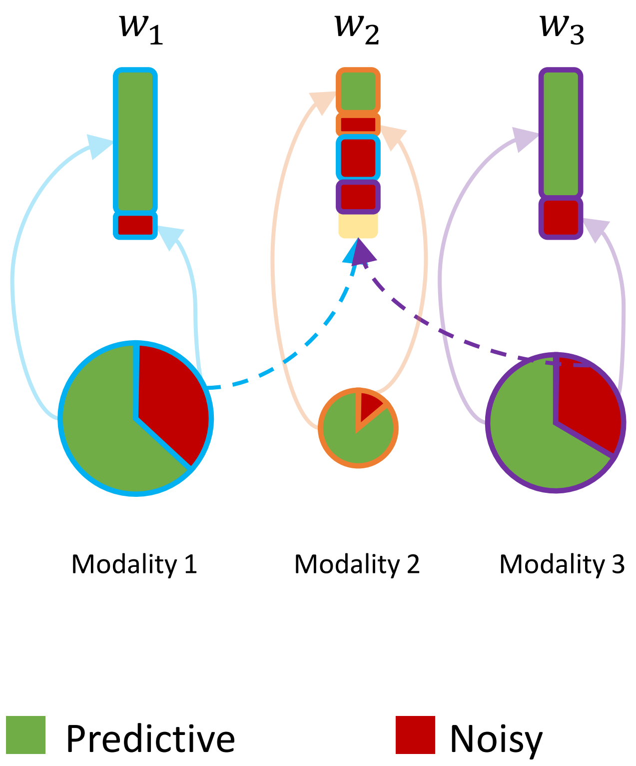
Cross-Modal Collisions due to Polysemanticity



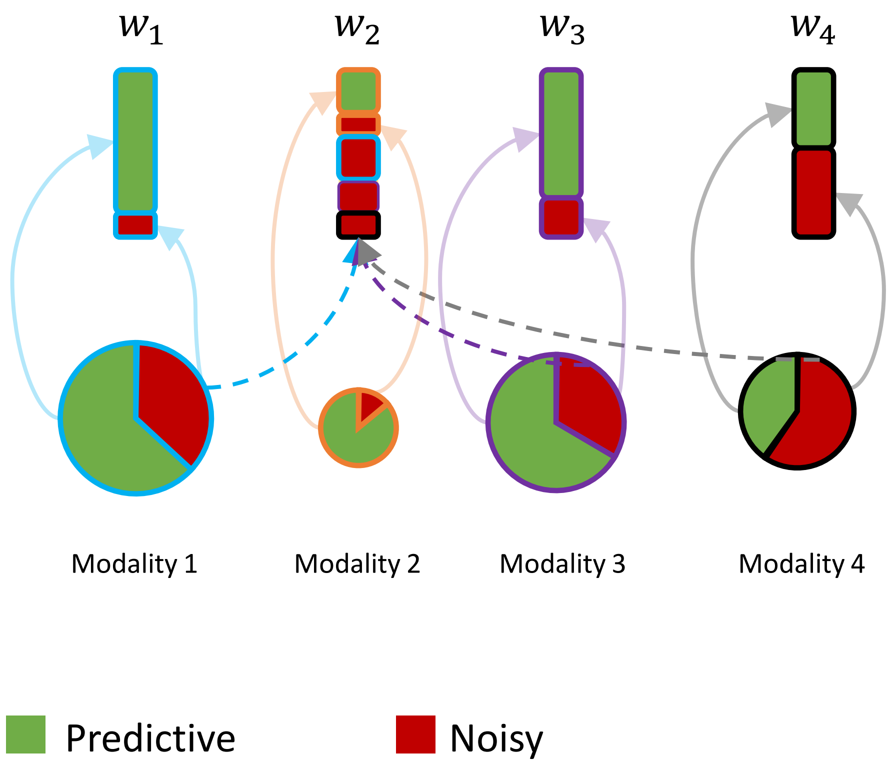
Cross-Modal Collisions due to Polysemanticity



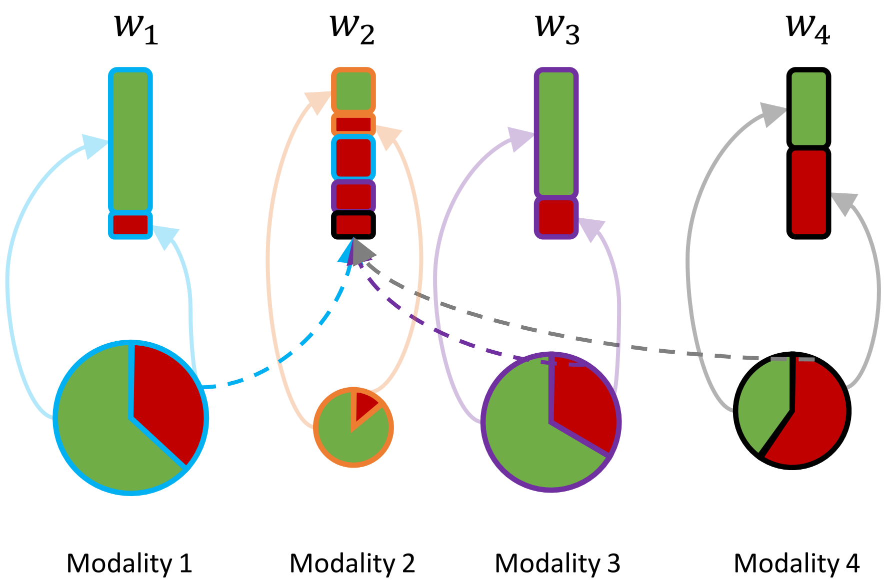
Cross-Modal Collisions due to Polysemanticity

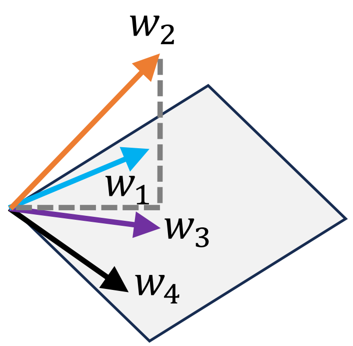


Cross-Modal Collisions due to Polysemanticity



Cross-Modal Collisions due to Polysemanticity

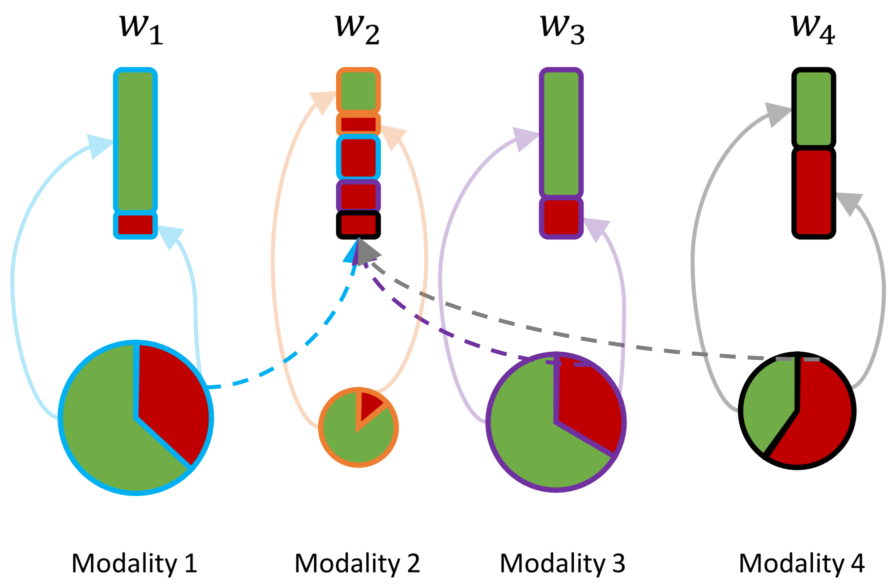




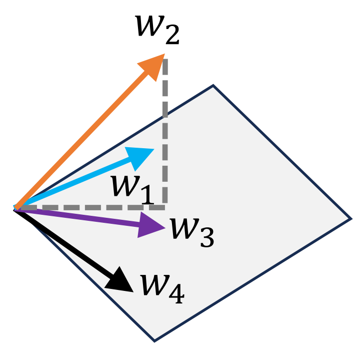
Predictive   Noisy

- Components along the subspace of other neurons (**rank bottleneck**) induce polysemanticity.

Cross-Modal Collisions due to Polysemanticity



Predictive Noisy

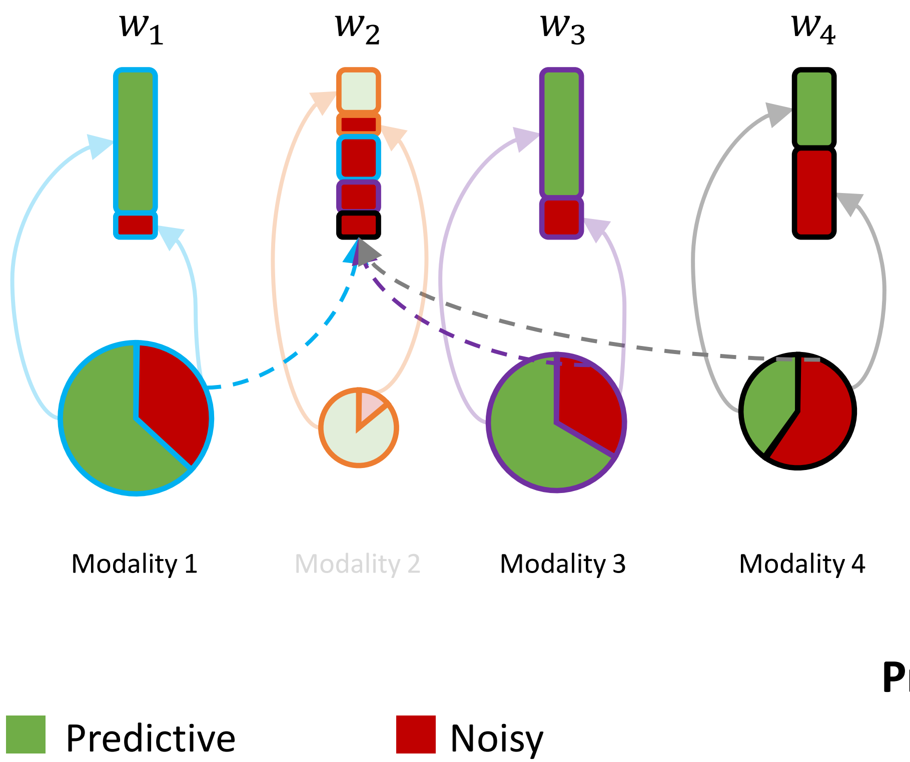


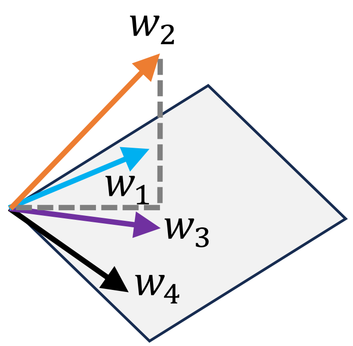
- Components along the subspace of other neurons (**rank bottleneck**) induce polysemanticity.

$$p(\mathbf{w}_p) \geq m(m-1)\frac{(\dim f_{min})^2}{\left(\sum_{i=1}^{m} \dim f_i\right)^2}$$

Probability of cross-modal polysemantic **collisions increase** with the **number of modalities**

Cross-Modal Collisions due to Polysemanticity



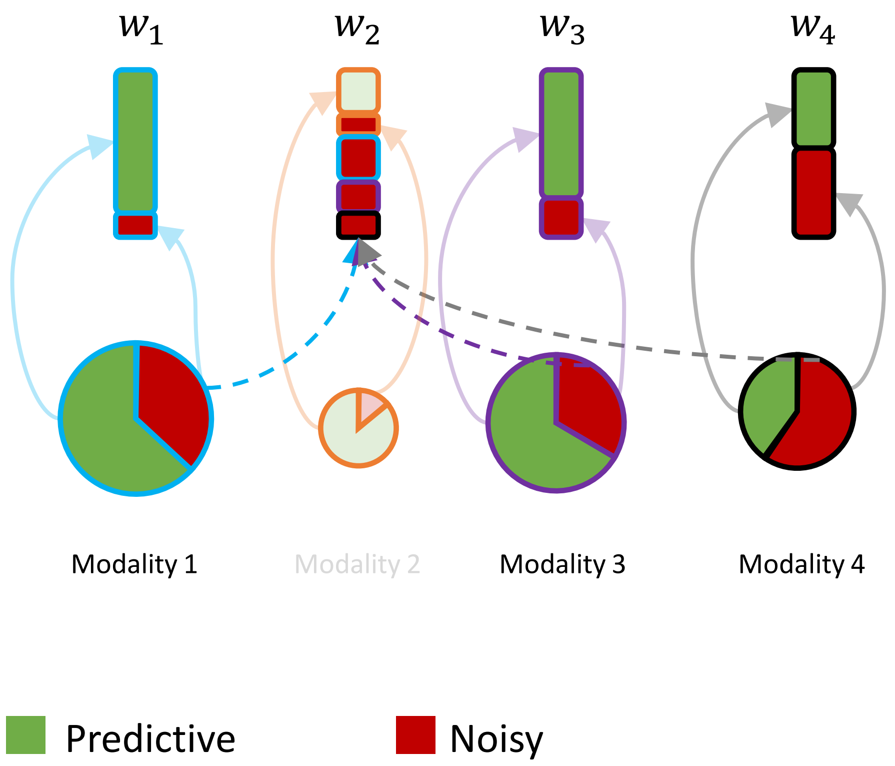


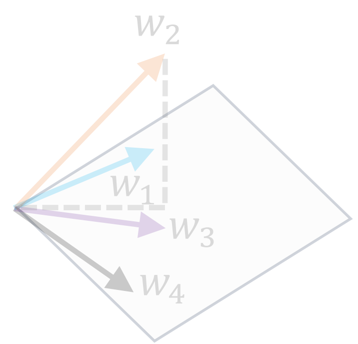
- Components along the subspace of other neurons (**rank bottleneck**) induce polysemanticity.
- **Increasing proportion of noisy features** in a neuron leads to **collapse** of the modality, the predictive features of which it is supposed to encode.

$$p(\mathbf{w}_p) \geq m(m-1)\frac{(\dim f_{min})^2}{\left(\sum_{i=1}^{m} \dim f_i\right)^2}$$

Probability of cross-modal polysemantic **collisions increase** with the **number of modalities**

Cross-Modal Collisions due to Polysemanticity





- Components along the subspace of other neurons (**rank bottleneck**) induce polysemanticity.
- **Increasing proportion of noisy features** in a neuron leads to **collapse** of the modality, the ...eatures of which it ... to encode.

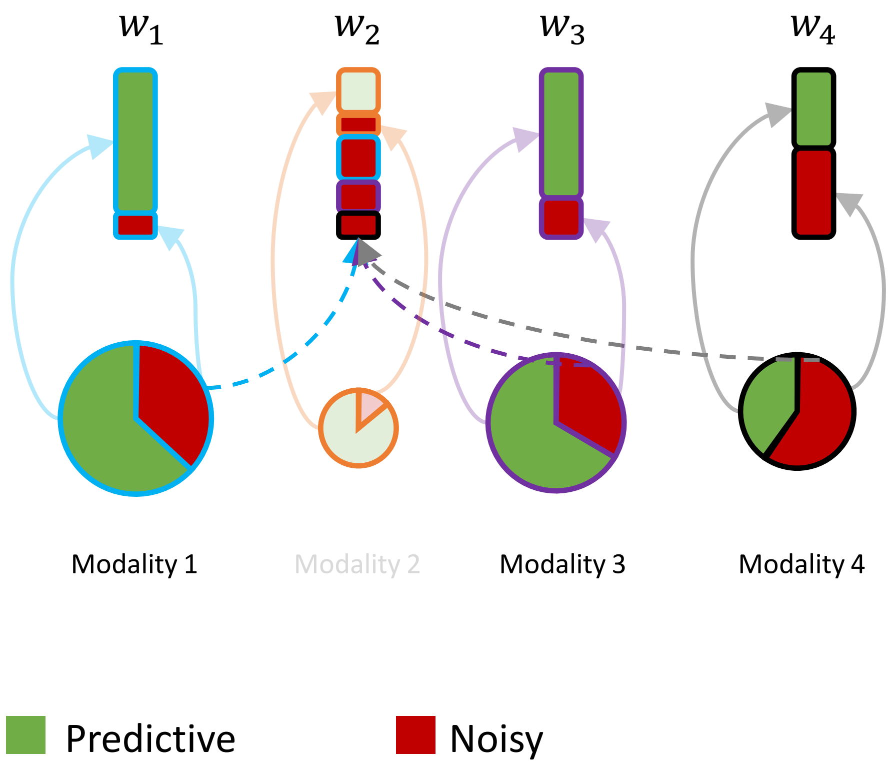
$$\left\| \mathbf{w} - \sum_{x \in X} \nabla\varphi_W(x)\nabla\varphi_W(x)^T \right\| \leq \gamma(\mathbf{w})^{-1/n}$$

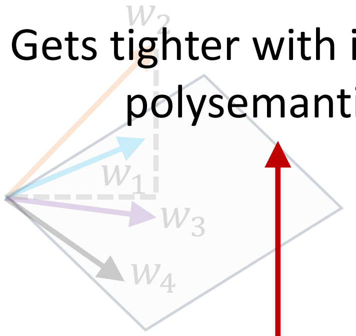
Polysemantic subspace

Weight matrix

Degree of polysemanticity

Cross-Modal Collisions due to Polysemanticity





Gets tighter with increasing polysemanticity

$$\left\| \mathbf{w} - \sum_{x \in X} \nabla \varphi_W(x) \nabla \varphi_W(x)^T \right\| \leq \gamma(\mathbf{w})^{-1/n}$$

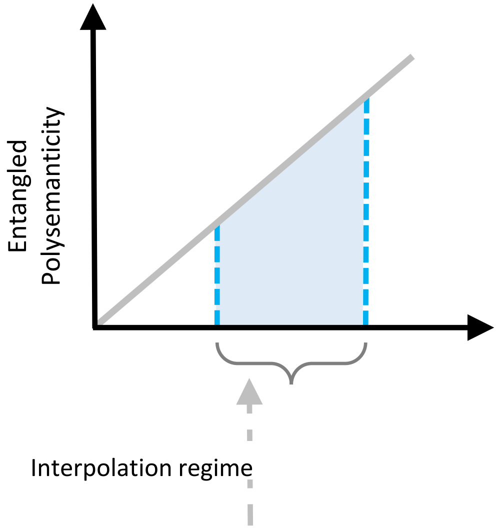
Polysemantic subspace

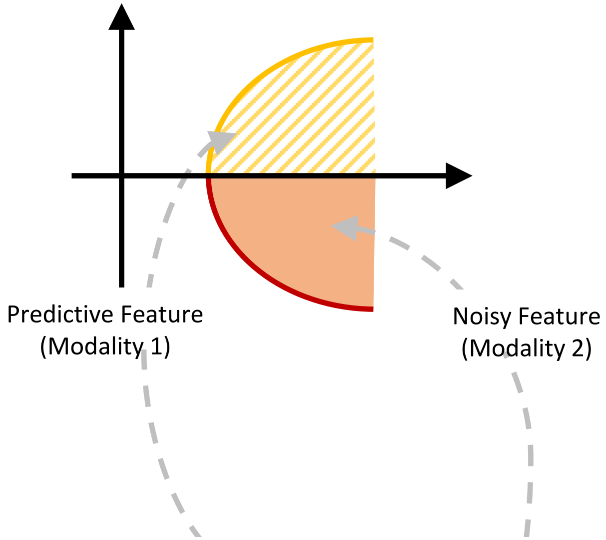
Weight matrix

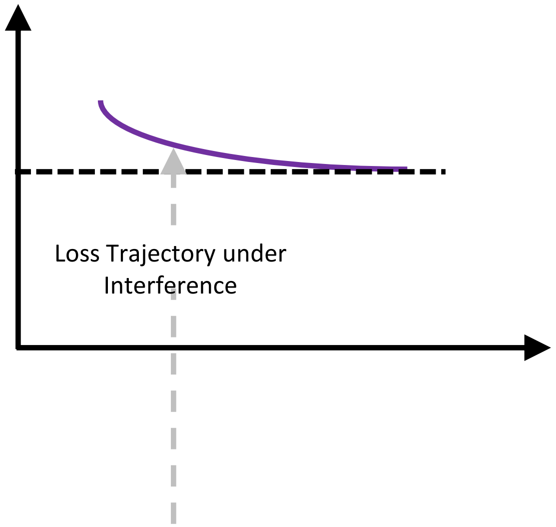
Degree of polysemanticity

- Components along the subspace of other neurons (**rank bottleneck**) induce polysemanticity.
- **Increasing proportion of noisy features** in a neuron leads to **collapse** of the modality, the ...atures of which it ...o encode.

Entangled *vs* Disentangled Polysemanticity

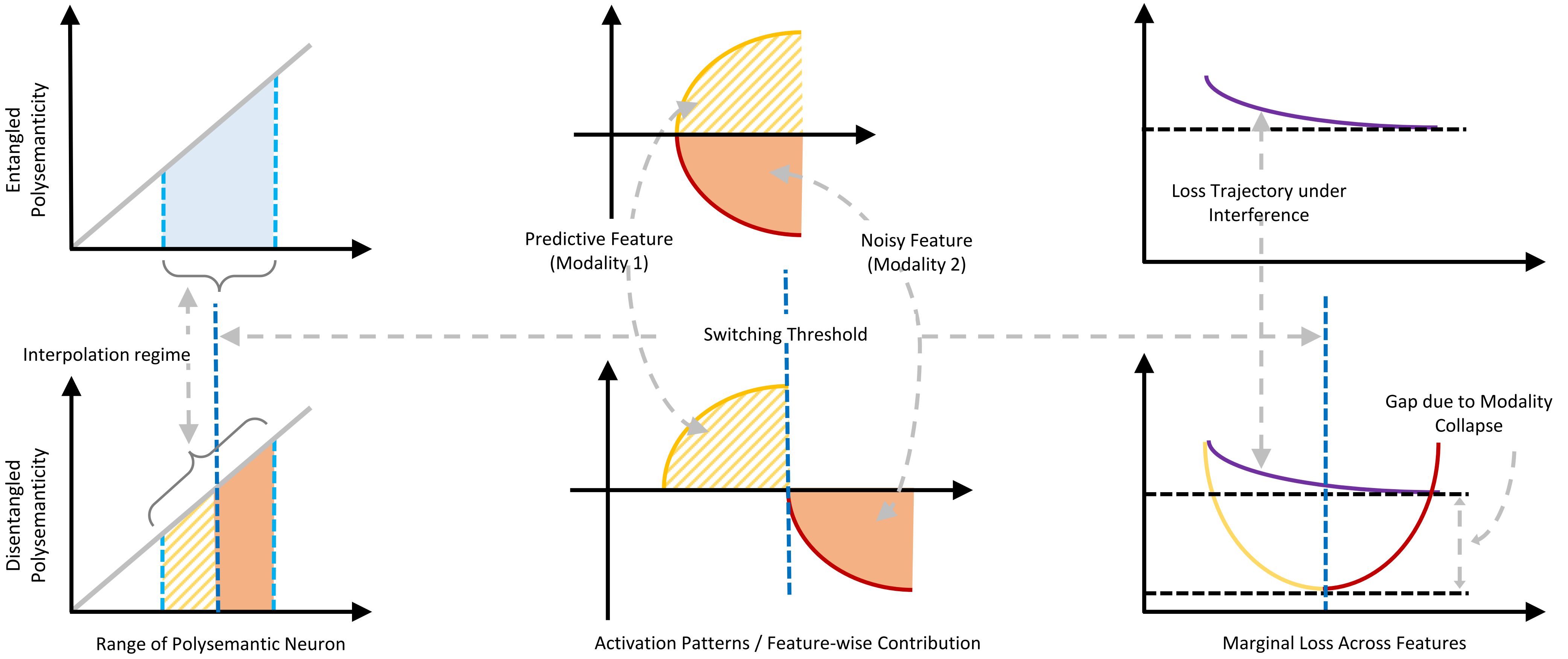






The vertical axes correspond to the value of the input feature.

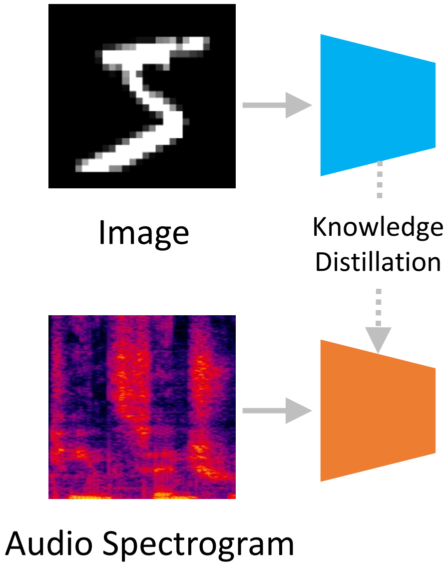
Entangled *vs* Disentangled Polysemanticity

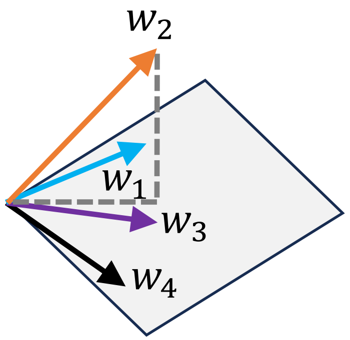


The vertical axes correspond to the value of the input feature.

Outline

Distillation Frees Up Rank Bottlenecks

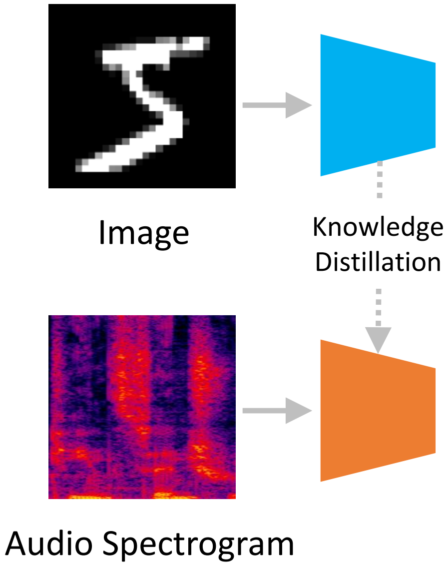


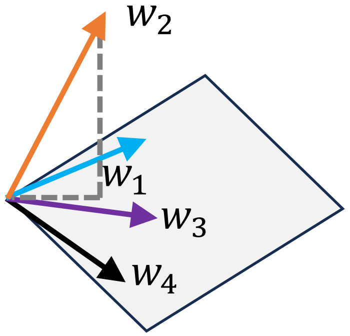


- Fusion head neuron corresponding to the student has fewer components along the other modalities.

Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia

Distillation Frees Up Rank Bottlenecks

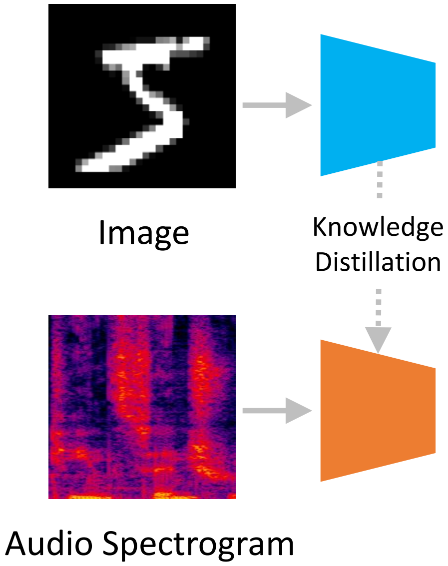


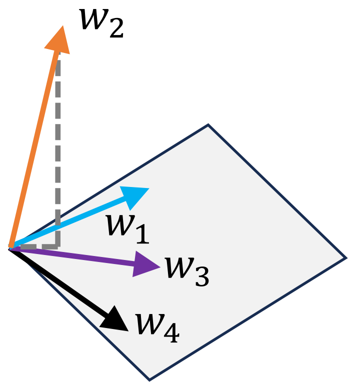


- Fusion head neuron corresponding to the student has fewer components along the other modalities.

Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia

Distillation Frees Up Rank Bottlenecks

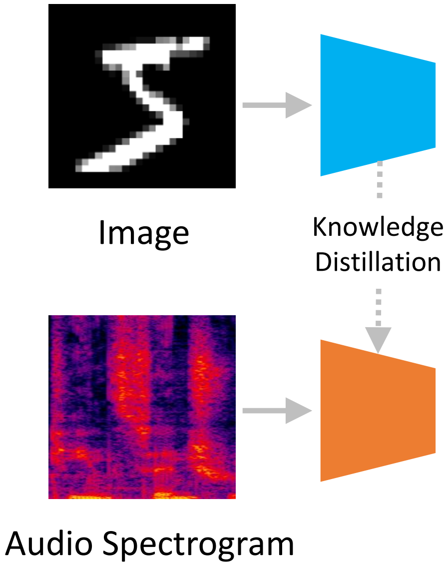


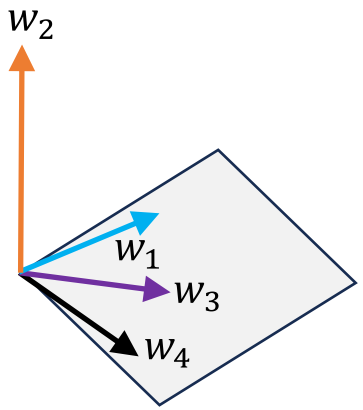


- Fusion head neuron corresponding to the student has fewer components along the other modalities.

Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia

Distillation Frees Up Rank Bottlenecks

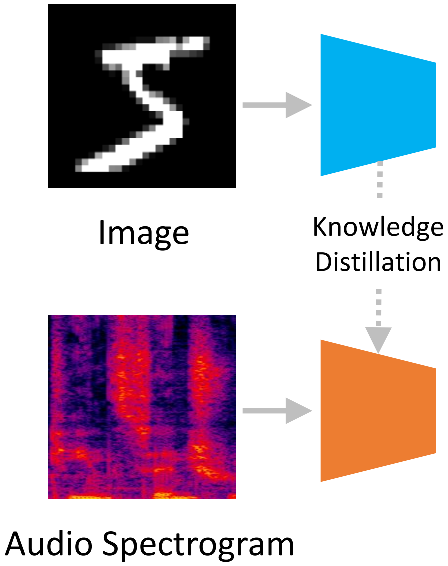


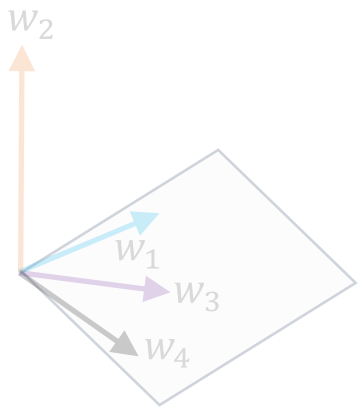


- Fusion head neuron corresponding to the student has fewer components along the other modalities.

Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia

Distillation Frees Up Rank Bottlenecks





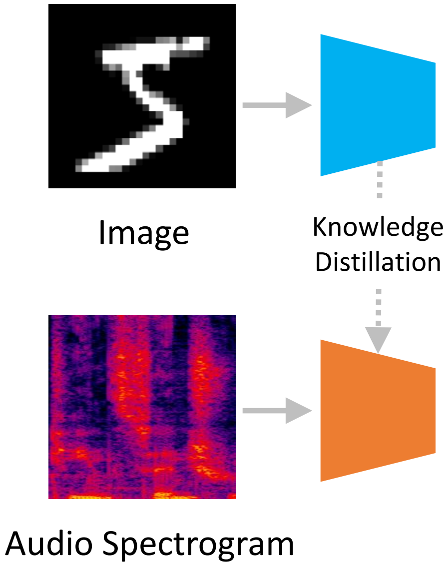
- Fusion head neuron corresponding to the student has fewer components along the other modalities.

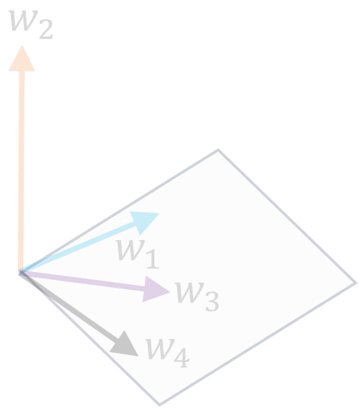
$$\boxed{\left\| \mathbf{w} - \sum_{x \in X} \nabla \varphi_W(x) \nabla \varphi_W(x)^T \right\| \leq \gamma(\mathbf{w})^{-1/n}}$$

Before

Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia

Distillation Frees Up Rank Bottlenecks





- Fusion head neuron corresponding to the student has fewer components along the other modalities.

$$\left\| \mathbf{w} - \sum_{x \in X} \nabla \varphi_W(x) \nabla \varphi_W(x)^T \right\| \leq \gamma(\mathbf{w})^{-1/n}$$

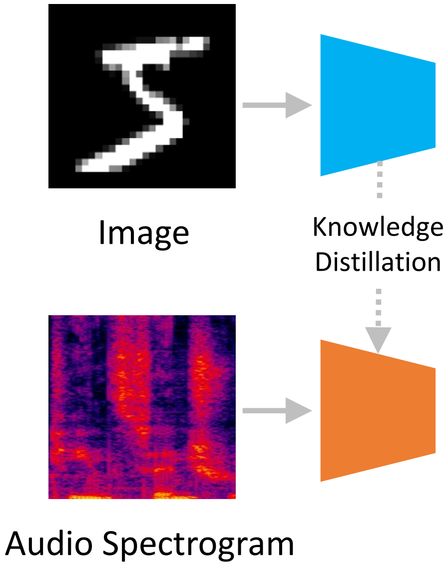
Before

After

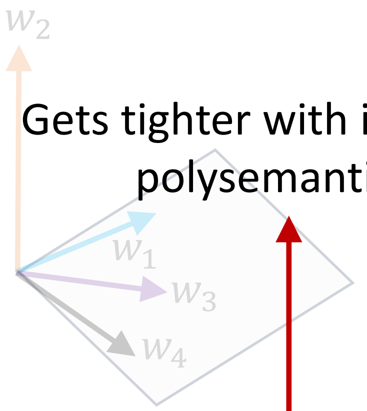
$$\lim_{d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \to \epsilon} \left\| \mathbf{w} - \sum_{\mathbf{x} \in X} \nabla \varphi_W(\mathbf{x}) \nabla \varphi_W(\mathbf{x})^T \right\| \leq \kappa^{-1/n}$$

Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia

Distillation Frees Up Rank Bottlenecks



Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia



- Fusion head neuron corresponding to the student has fewer components along the other modalities.

Gets tighter with increasing polysemanticity

Agnostic of the polysemantic subspace

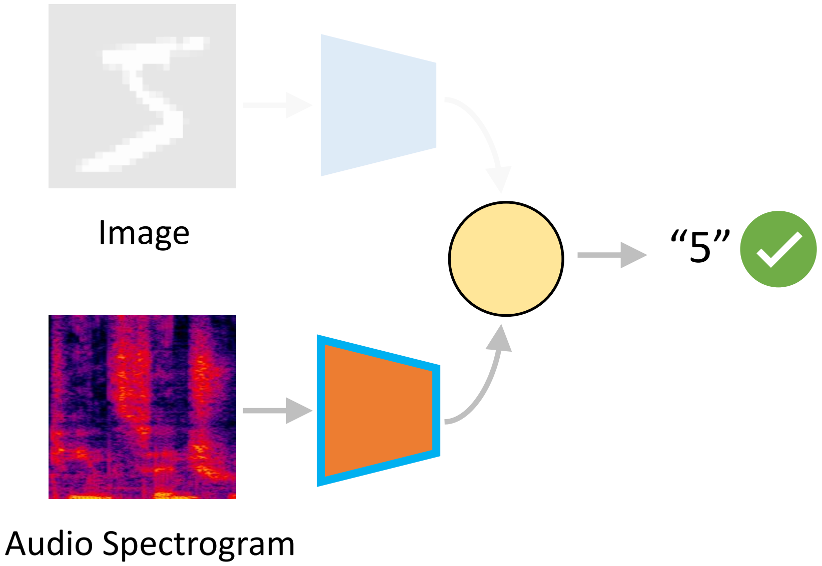
Before

$$\boxed{\left\| \mathbf{w} - \sum_{x\in X} \nabla\varphi_W(x)\nabla\varphi_W(x)^T \right\| \le \gamma(\mathbf{w})^{-1/n}}$$

After

$$\boxed{\lim_{d(\tilde{\mathbf{x}}_i,\tilde{\mathbf{x}}_j)\to\epsilon} \left\| \mathbf{w} - \sum_{\mathbf{x}\in X} \nabla\varphi_W(\mathbf{x})\nabla\varphi_W(\mathbf{x})^T \right\| \le \kappa^{-1/n}}$$

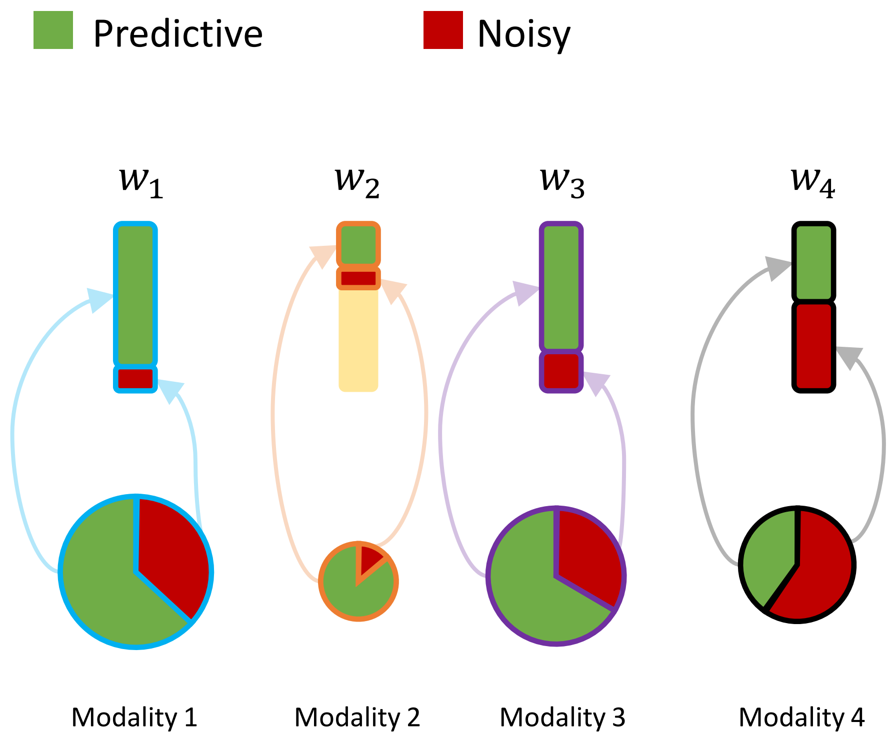
Distillation Frees Up Rank Bottlenecks

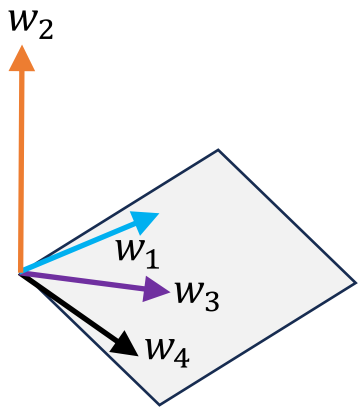


- Fusion head neuron corresponding to the student has fewer components along the other modalities.
- The student can function independently in absence of the teacher.

Visual Sources: Image – MNIST, Audio Spectrogram – Wikipedia

The Distillation Denoising Conjecture





- Knowledge distillation allows the **representation** of the **noise-components of the teacher** modalities as a **transformed version** of the **student noise**:

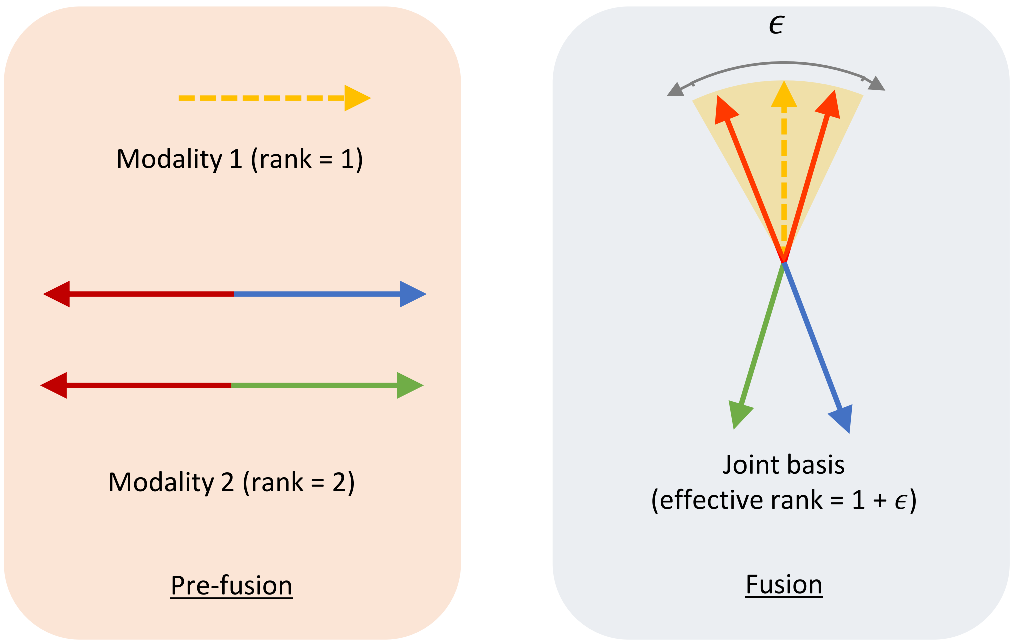
$$m_t\hat{\eta} = \phi_t(m_s\hat{\eta})$$

- This **eliminates** the need for encoding **noisy features** from every modality in the neurons encoding the **student modality**.

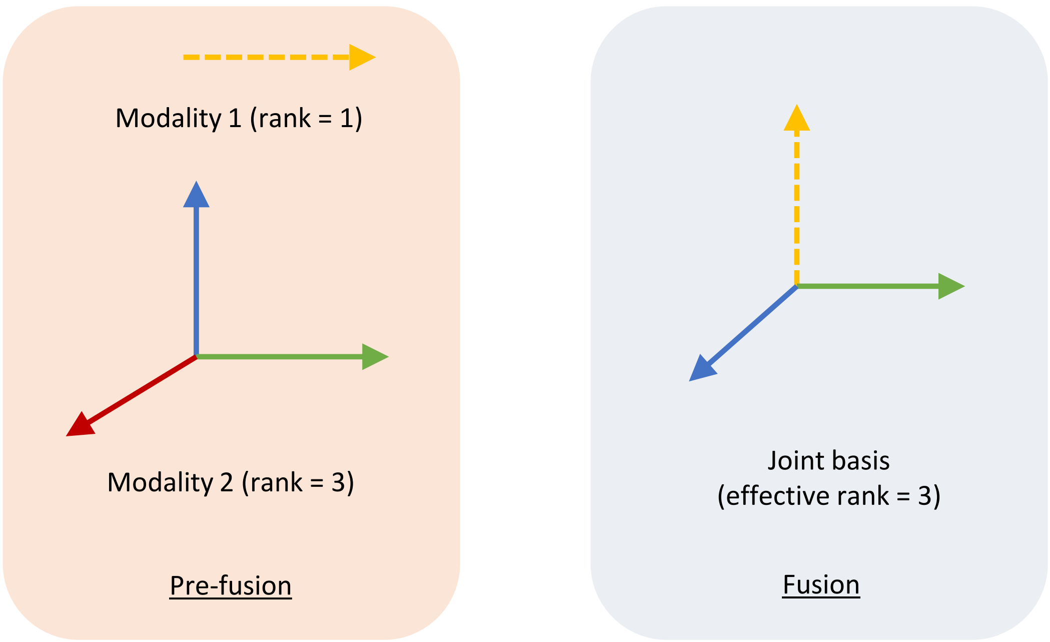
$m_t\hat{\eta}$: Teacher noise; $m_s\hat{\eta}$: Student noise; ϕ_t: Transformation function learned via distillation

Putting Things Together

Freeing up Rank Bottlenecks via Basis Reallocation

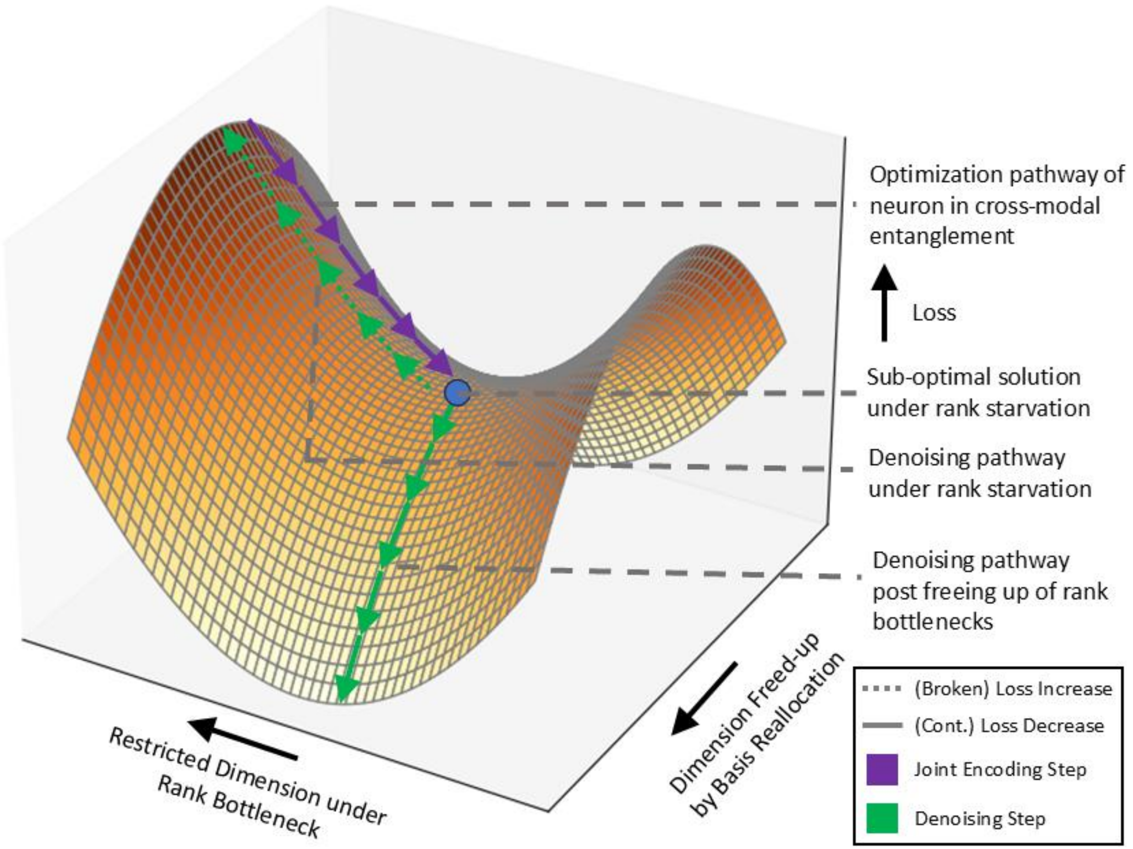


(a) Cross-Modal Interference due to Rank Bottleneck



(b) Rank Bottleneck Free-up via Basis Reallocation

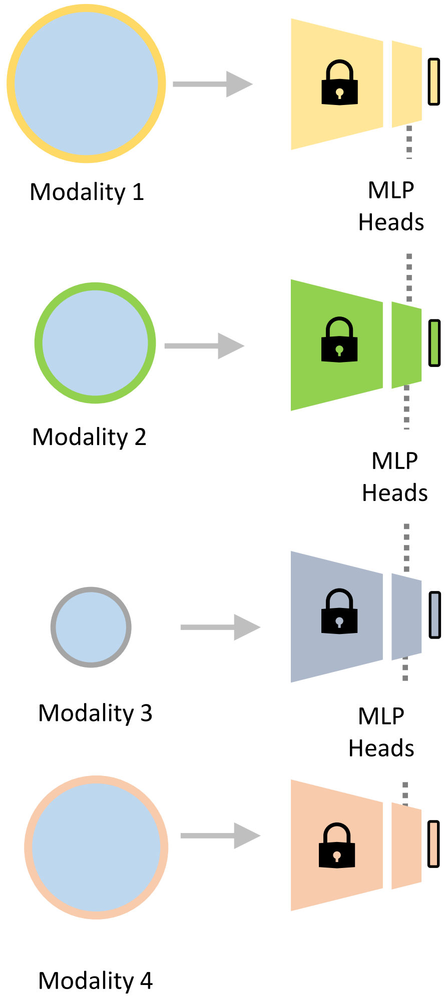
Putting Things Together



Outline

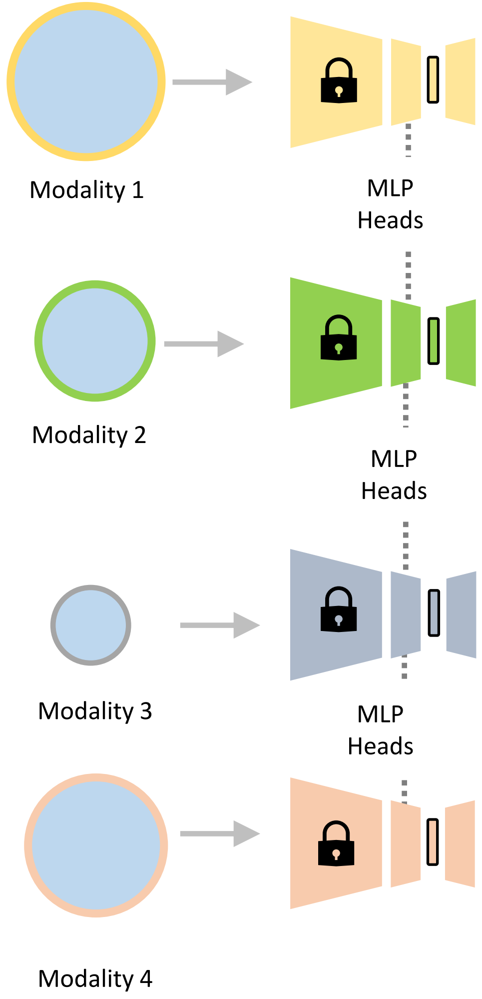
- Motivation
- Understanding Modality Collapse
- The Effect of Knowledge Distillation
- Explicit Basis Reallocation
- Experiments
- Conclusion and Open Problems

Explicit Basis Reallocation



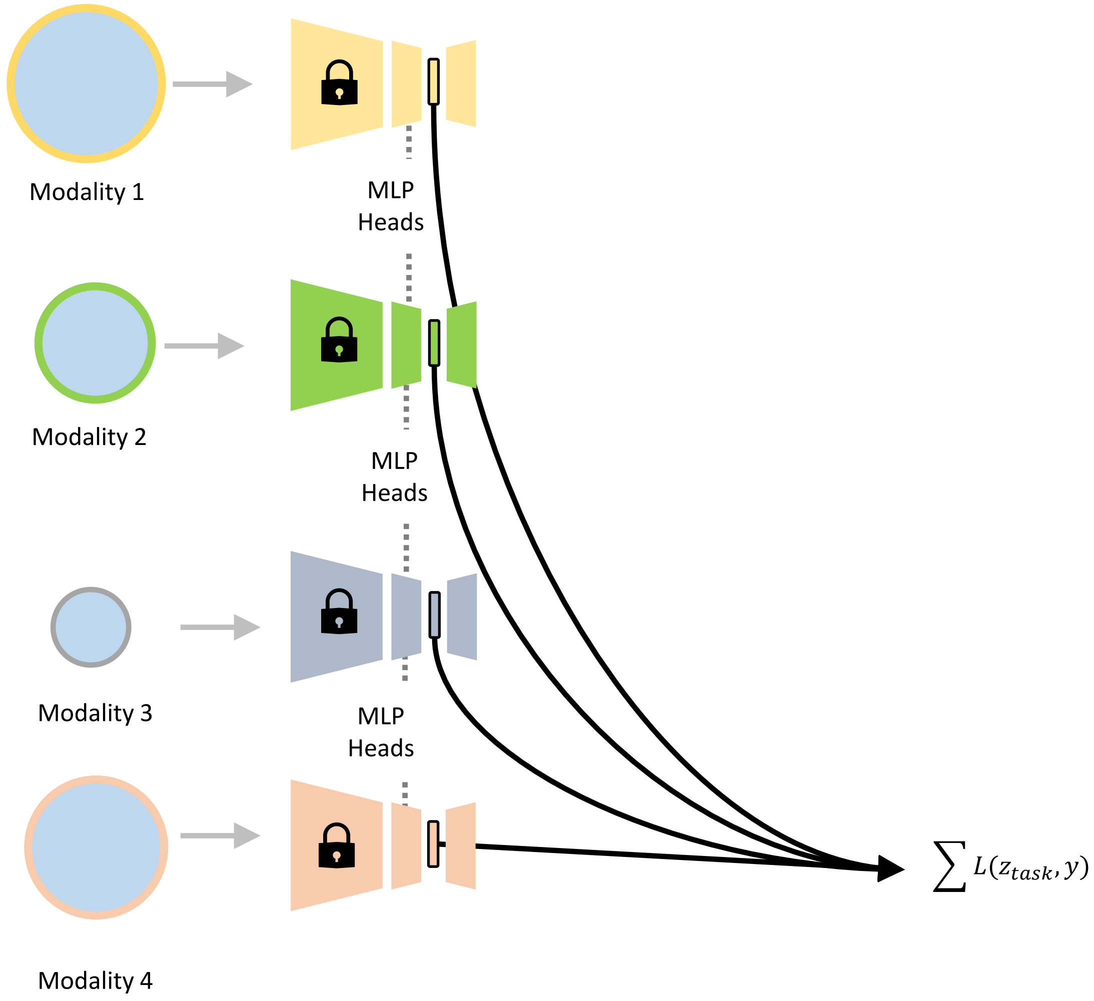
- Identify lower-dimensional latent properties.

Explicit Basis Reallocation



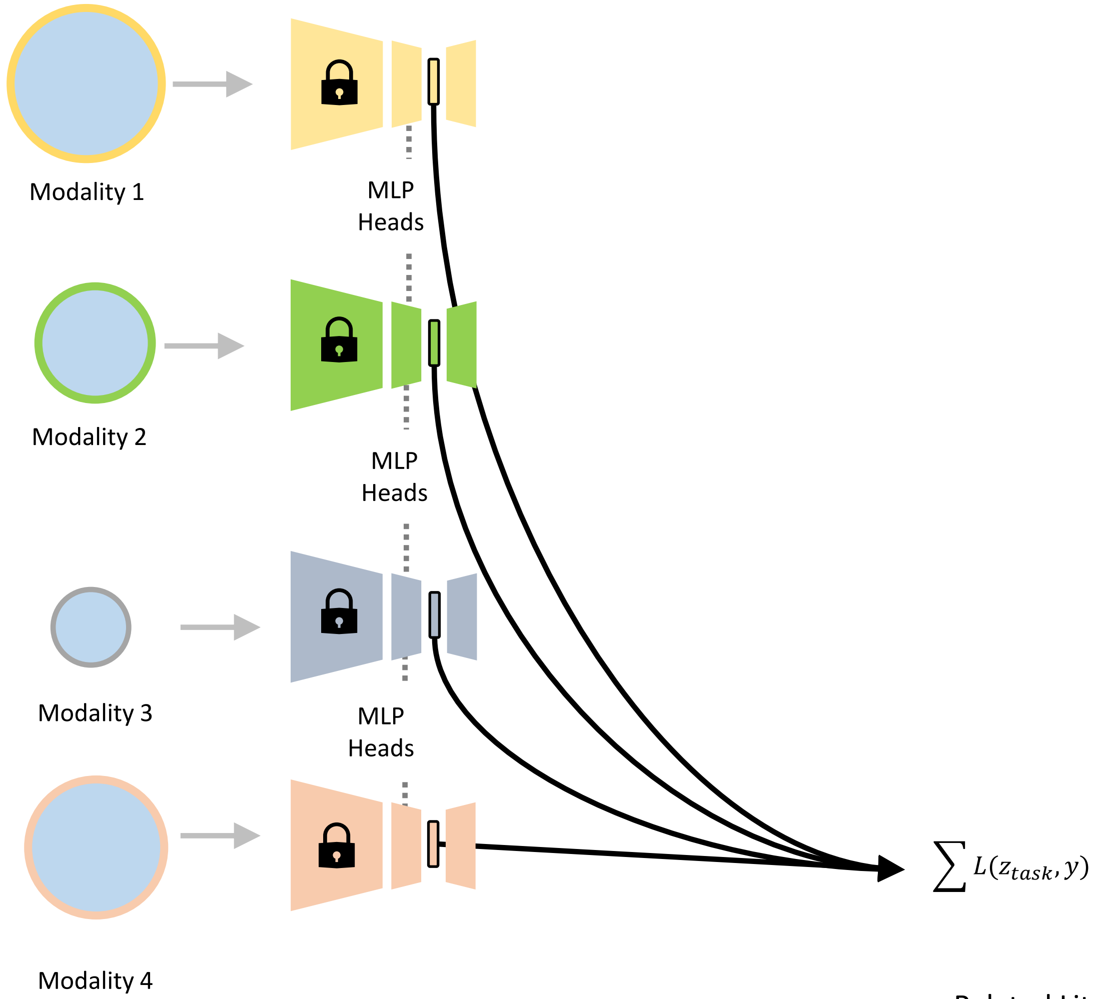
- Identify lower-dimensional latent properties.
- Reconstruct back to the input representation.

Explicit Basis Reallocation



- Identify lower-dimensional latent properties.
- Reconstruct back to the input representation.
- Ensure semantic consistency of latent properties.

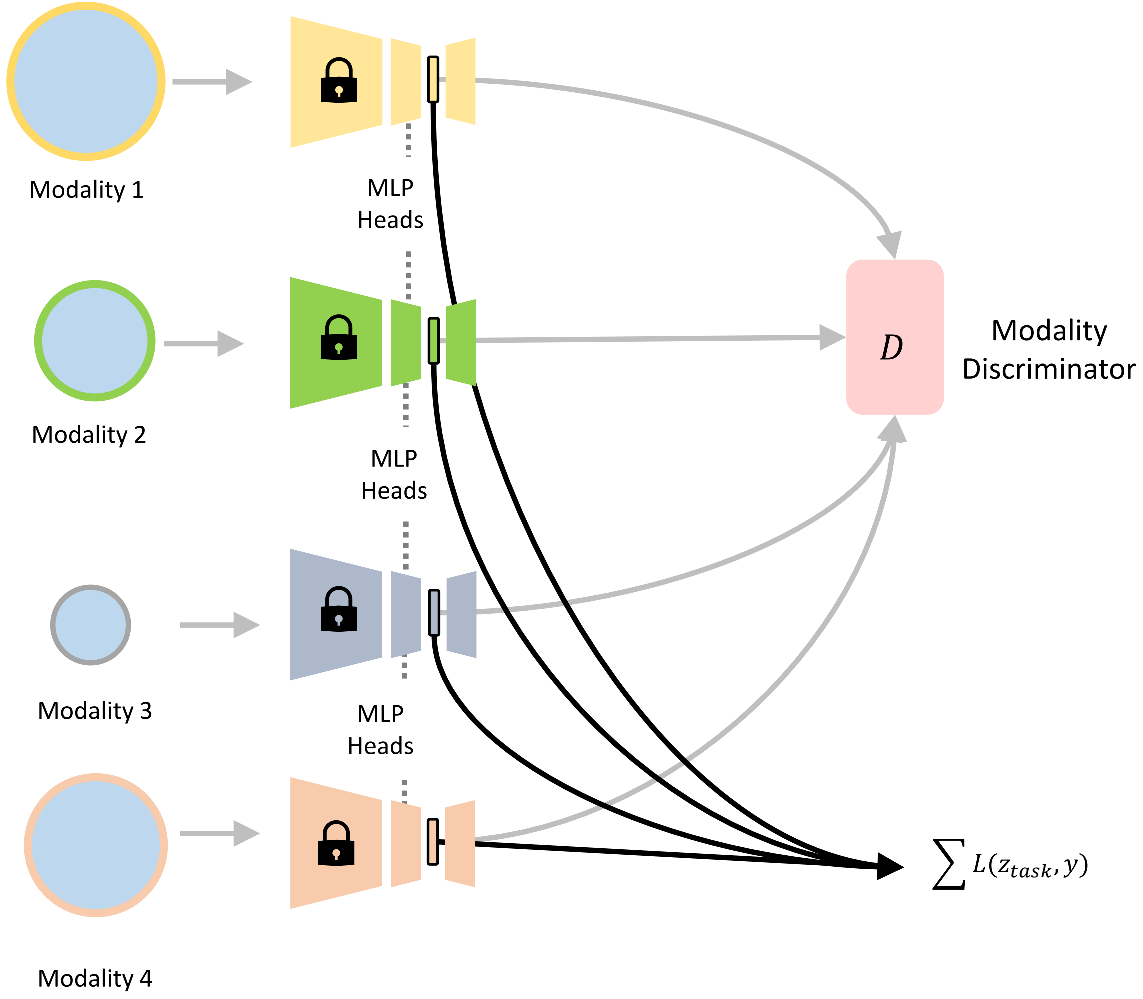
Explicit Basis Reallocation



- Identify lower-dimensional latent properties.
- Reconstruct back to the input representation.
- Ensure semantic consistency of latent properties.
  - Being able to **reconstruct** the input from the latent while **minimizing the task loss** in the latent space implies that:
    1. The latent encodes the causal factors.
    2. The reconstruction head implements the causal mechanisms.

Related Literature: Parascandolo et al., Learning Independent Causal Mechanisms, ICML 2018.

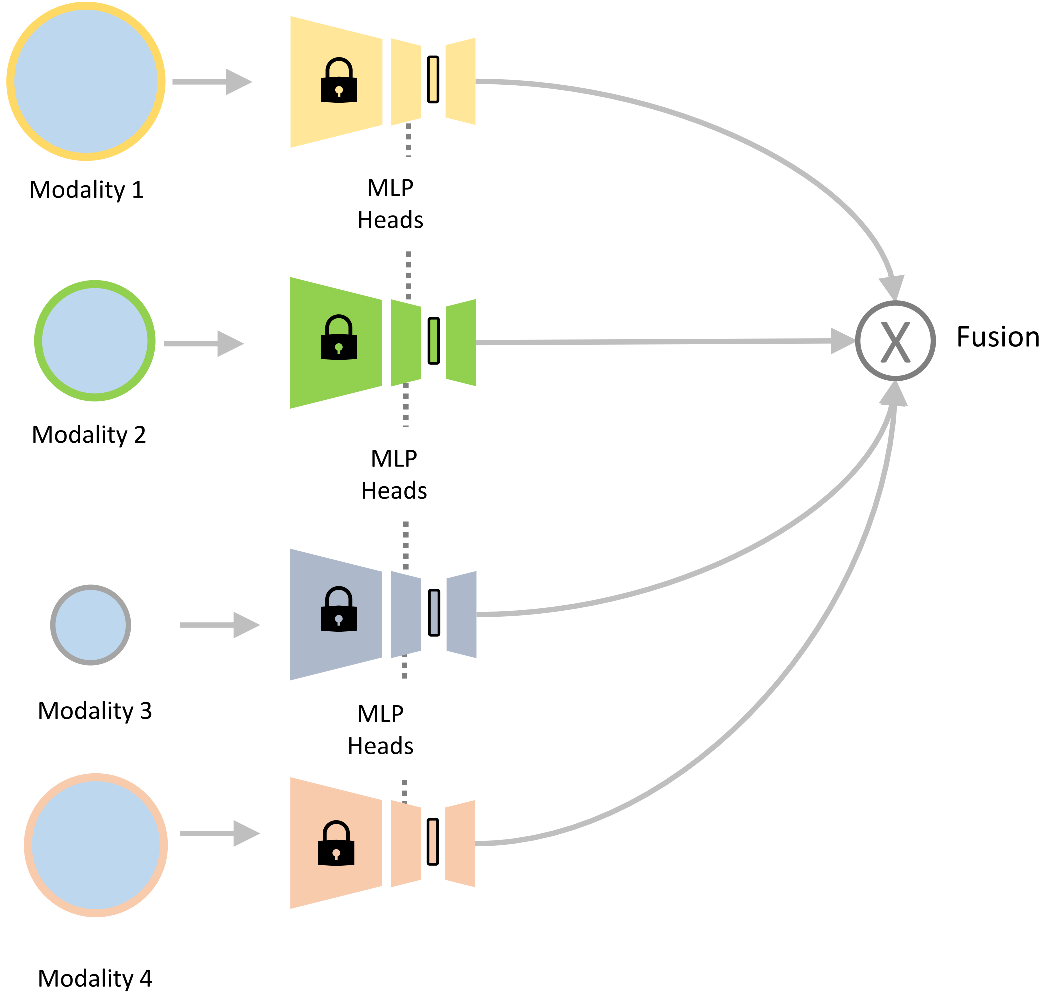
Explicit Basis Reallocation



- Identify lower-dimensional latent properties.
- Reconstruct back to the input representation.
- Ensure semantic consistency of latent properties.
- Semantics-preserving mechanism invariance through modality discriminator.
  - The modality discriminator is trained until the respective task validation accuracies start dropping.

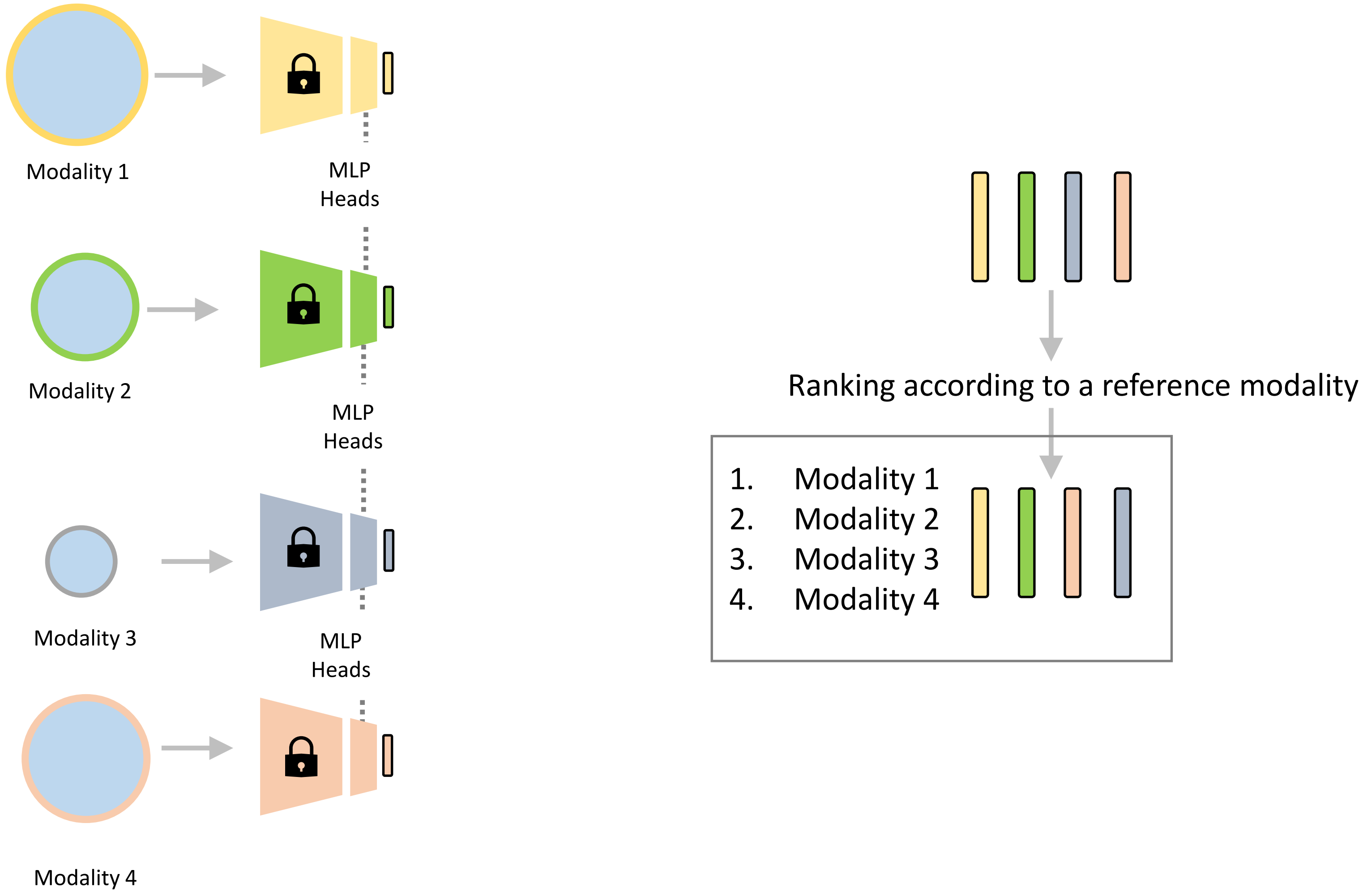
Related Literature: Arjovsky et al., Invariant Risk Minimization, 2020.

Tying Fusion Head to Factors and Mechanisms

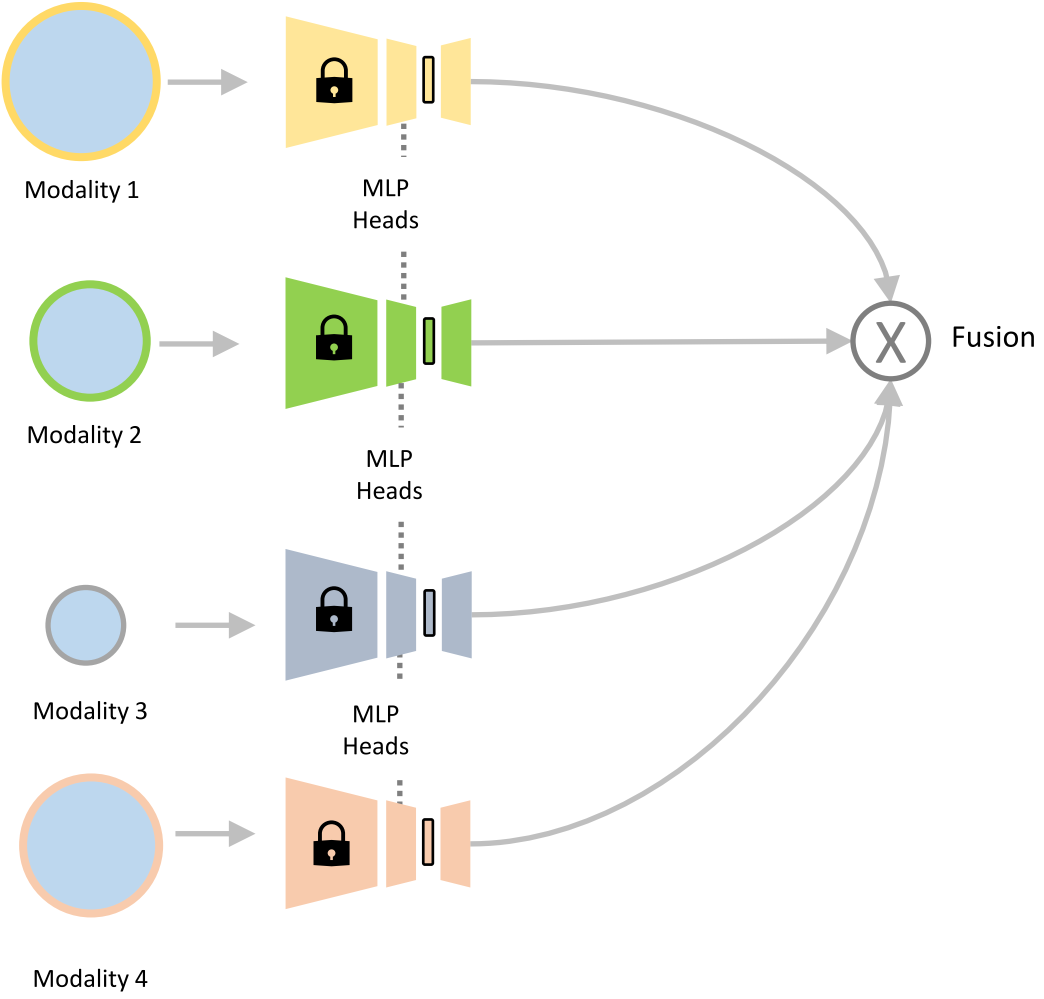


- The fusion head is trained on the representations obtained from the inverse mechanisms applied to the causal factors.
- This decouples the fusion head from the modalities and ties it to the recovered causal factors and mechanisms.

Similarity-Based Ordering of Causal Factors



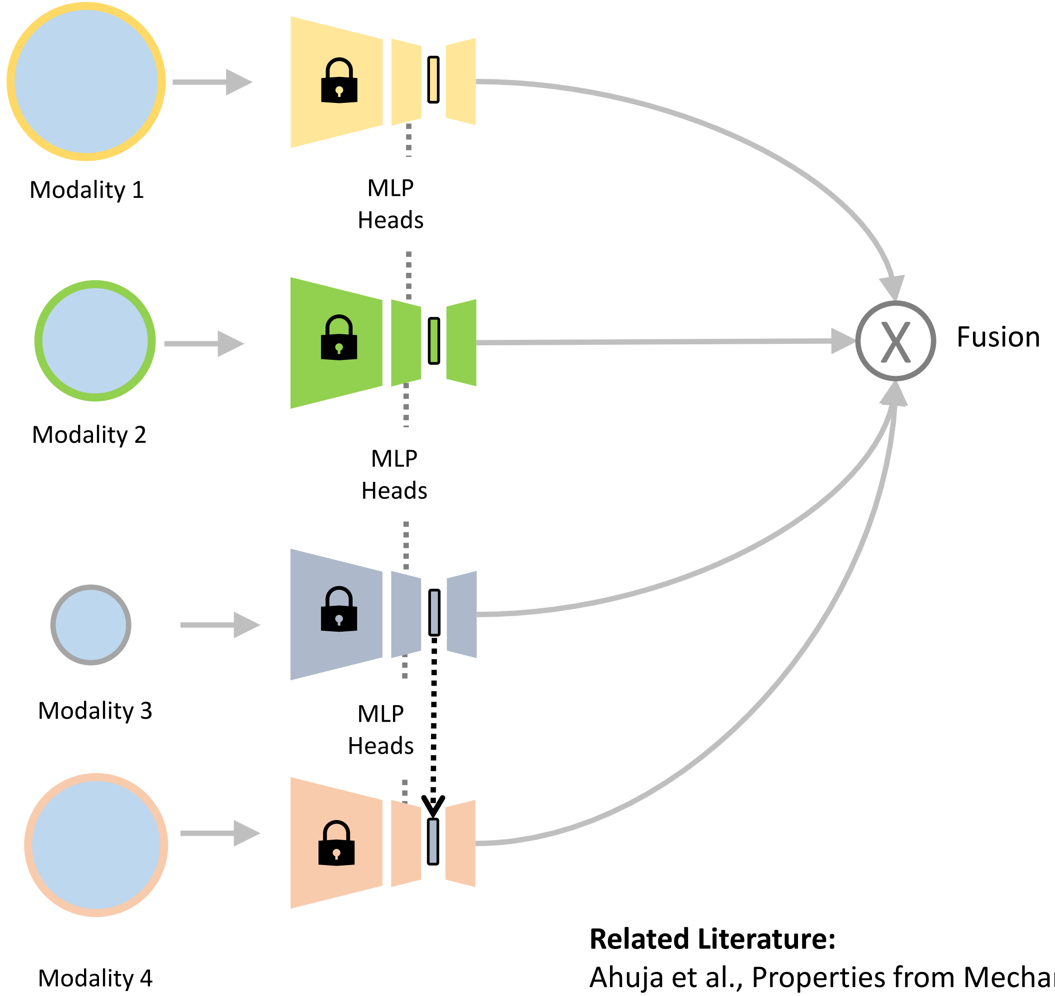
Similarity-Based Ordering of Causal Factors



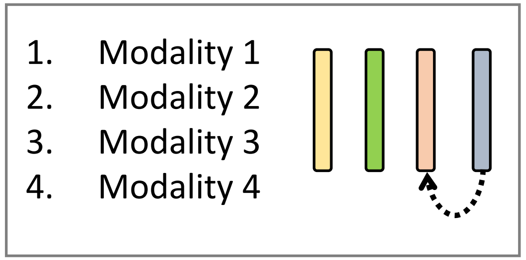
- When modalities go missing, check the rank list and substitute with the modality of the closest rank.

1. Modality 1
2. Modality 2
3. Modality 3
4. Modality 4

Substitution with the Closest Factor



- When modalities go missing, check the rank list and substitute with the modality of the closest rank.



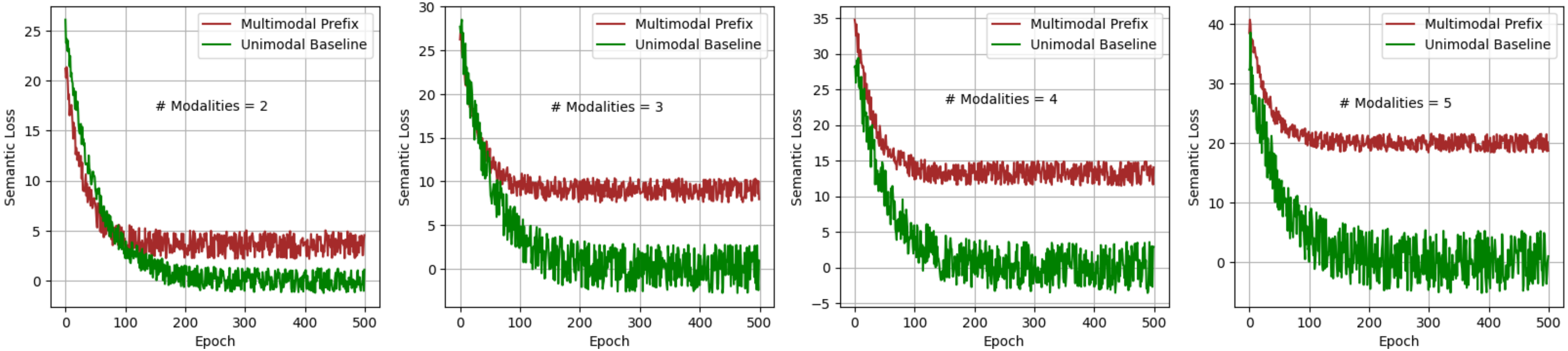
Related Literature:

Ahuja et al., Properties from Mechanisms: An Equivariance Perspective on Identifiable Representation Learning, ICLR 2023.

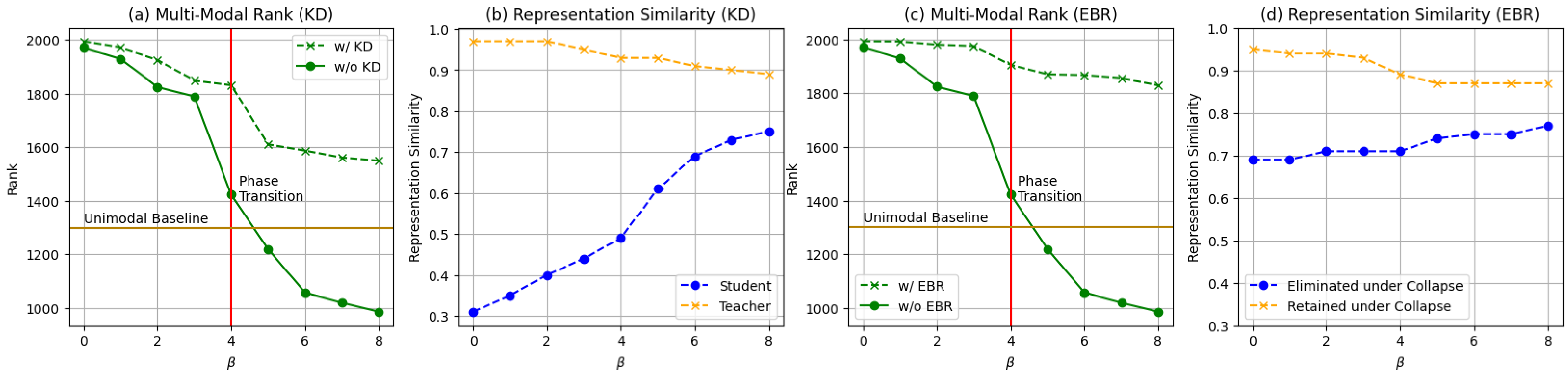
Gulrajani and Hashimoto, Identifiability Conditions for Domain Adaptation, ICML 2022.

Outline

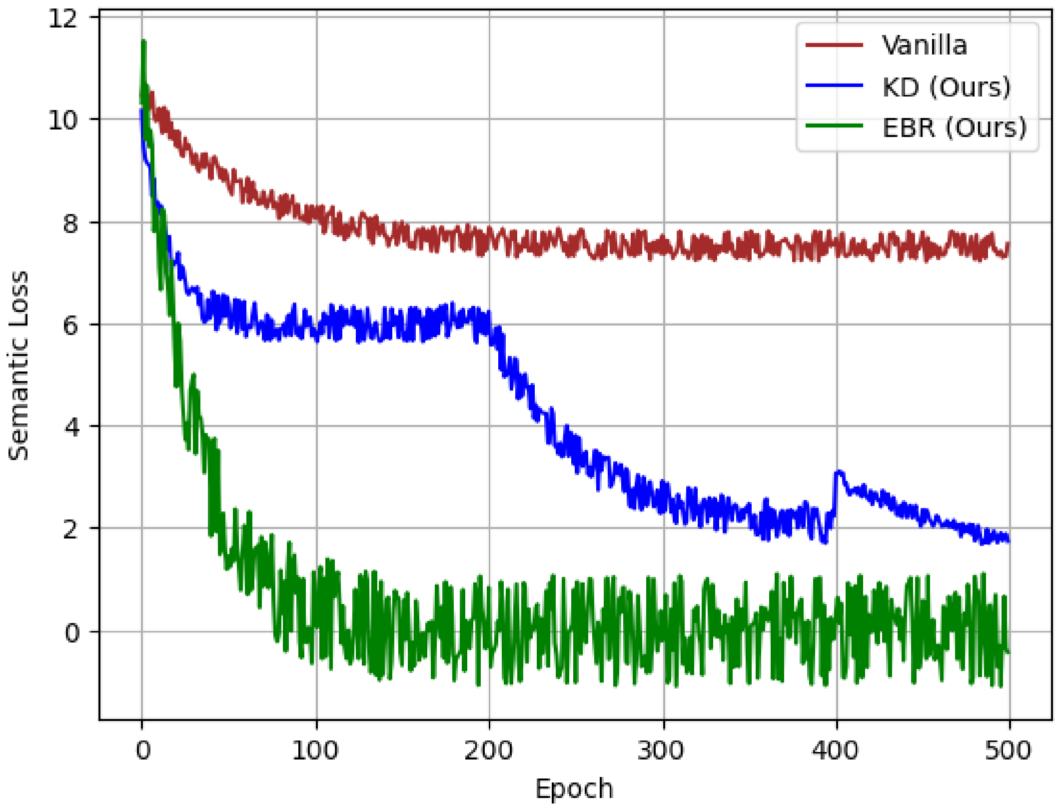
Experiments – Cross-Modal Entanglements



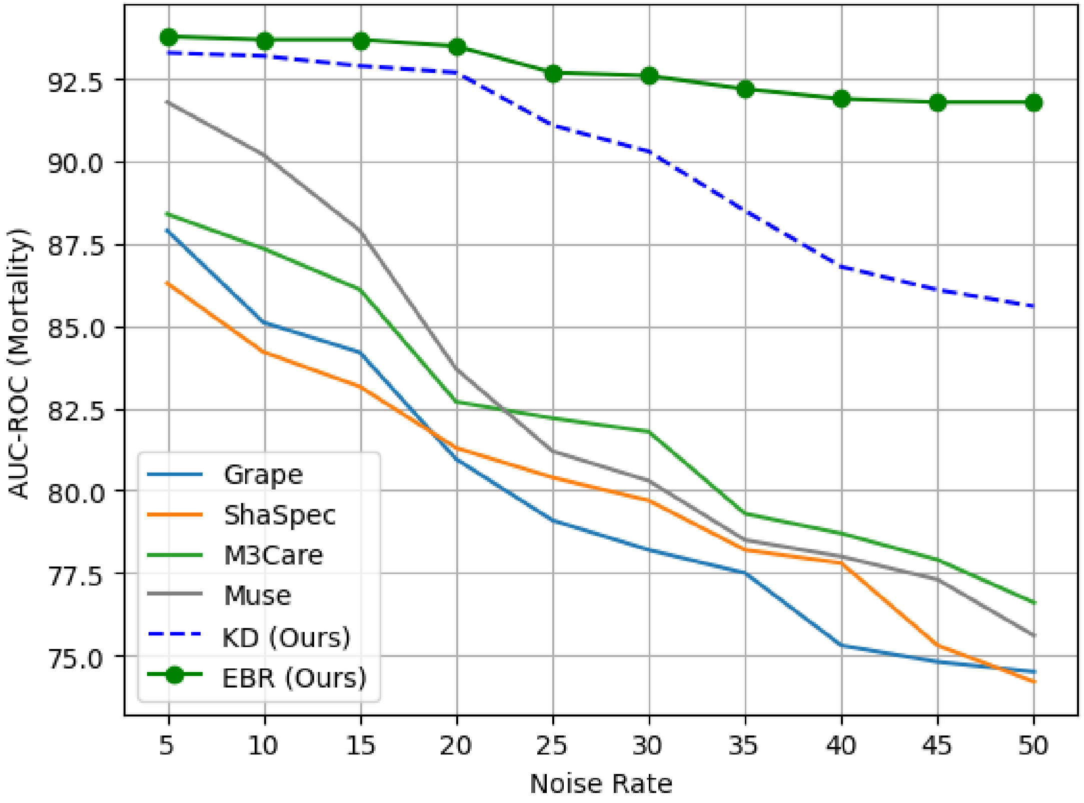
Experiments – Rank and Representation Similarity



Experiments - Convergence



Experiments – Denoising



Experiments – Comparison with SOTA

Method	Mortality		Readmission	
	AUC-ROC	AUC-PRC	AUC-ROC	AUC-PRC
Grape (NeurIPS ’20)	0.8837	0.4584	0.7085	0.4551
+ KD	0.9011	0.4620	0.7231	0.4610
+ EBR	**0.9102**	**0.4799**	**0.7488**	**0.4691**
M3Care (SIGKDD ’22)	0.8896	0.4603	0.7067	0.4532
+ KD	0.8950	0.4700	0.7080	0.4562
+ EBR	**0.8987**	**0.4850**	**0.7296**	**0.4832**
MUSE (ICLR’24)	0.9201	0.4883	0.7351	0.4985
+ KD	0.9350	0.4993	0.7402	0.5066
+ EBR	**0.9380**	**0.5001**	**0.7597**	**0.5138**

Vanilla Multimodal Learning

Method	Mortality		Readmission	
	AUC-ROC	AUC-PRC	AUC-ROC	AUC-PRC
CM-AE (ICML ’11)	0.7873 ± 0.40	0.3620 ± 0.22	0.6007 ± 0.31	0.3355 ± 0.25
SMIL (AAAI ’21)	0.7981 ± 0.11	0.3536 ± 0.12	0.6155 ± 0.09	0.3279 ± 0.15
MT (CVPR ’22)	0.8176 ± 0.10	0.3467 ± 0.06	0.6278 ± 0.09	0.2959 ± 0.05
Grape (NeurIPS ’20)	0.7657 ± 0.16	0.3733 ± 0.09	0.6335 ± 0.07	0.3120 ± 0.11
M3Care (SIGKDD ’22)	0.8265 ± 0.09	0.3830 ± 0.07	0.6020 ± 0.09	0.3870 ± 0.05
ShaSpec (CVPR ’23)	0.8100 ± 0.13	0.3630 ± 0.09	0.6216 ± 0.10	0.3549 ± 0.08
MUSE (ICLR’24)	0.8236 ± 0.09	0.39.87 ± 0.05	0.6781 ± 0.05	0.4185 ± 0.07
EBR (Ours)	**0.8533 ± 0.09**	**0.4277 ± 0.02**	**0.7030 ± 0.05**	**0.4290 ± 0.02**

Average across multiple missingness rates (random elimination of modalities during inference)

Outline

Conclusions

- **Modality collapse** is the result of **cross-modal polysemantic interference** between **predictive** features of one modality and **noisy** features from another.

Conclusions

- **Modality collapse** is the result of **cross-modal polysemantic interference** between **predictive** features of one modality and **noisy** features from another.
- It is a consequence of the **low-rank simplicity bias** in neural networks.

Conclusions

- **Modality collapse** is the result of **cross-modal polysemantic interference** between **predictive** features of one modality and **noisy** features from another.
- It is a consequence of the **low-rank simplicity bias** in neural networks.
- It can thus be **prevented by freeing up such bottlenecks** through implicit or explicit **basis reallocation**.

Open Problems

- Verification of **feature-wise separability** in disentangled polysemantic neurons.
- Effect of **unequal label information** across features.
- The **Distillation Denoising Conjecture**.
- **Geometry** of the **loss landscape** under **basis reallocation**.

A Closer Look at Multimodal Representation Collapse

Project Page

Get in touch:
Abhra Chaudhuri
abhra.chaudhuri@fujitsu.com

https://abhrac.github.io/mmcollapse/