Multi-modal contrastive learning adapts to *intrinsic dimension* of shared latent variables

Yu Gui[1] Cong Ma[2] Zongming Ma[3]

[1] Department of Statistics and Data Science, University of Pennsylvania, [2] Department of Statistics, University of Chicago, [3] Department of Statistics and Data Science, Yale University

yugui@wharton.upenn.edu, congm@uchicago.edu, zongming.ma@yale.edu

Growing availability of multi-modal measurements

$\Downarrow$

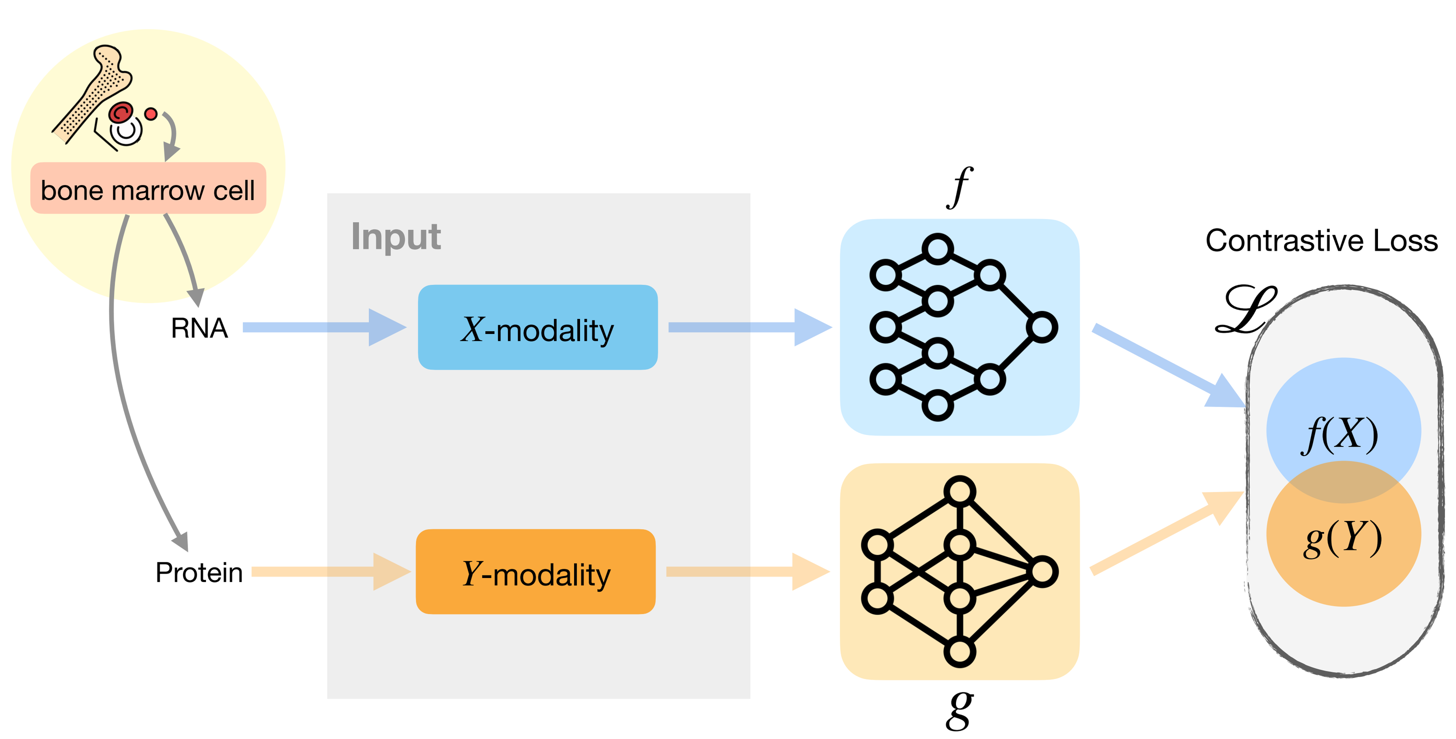
How can one efficiently integrate data from multi-modalities?

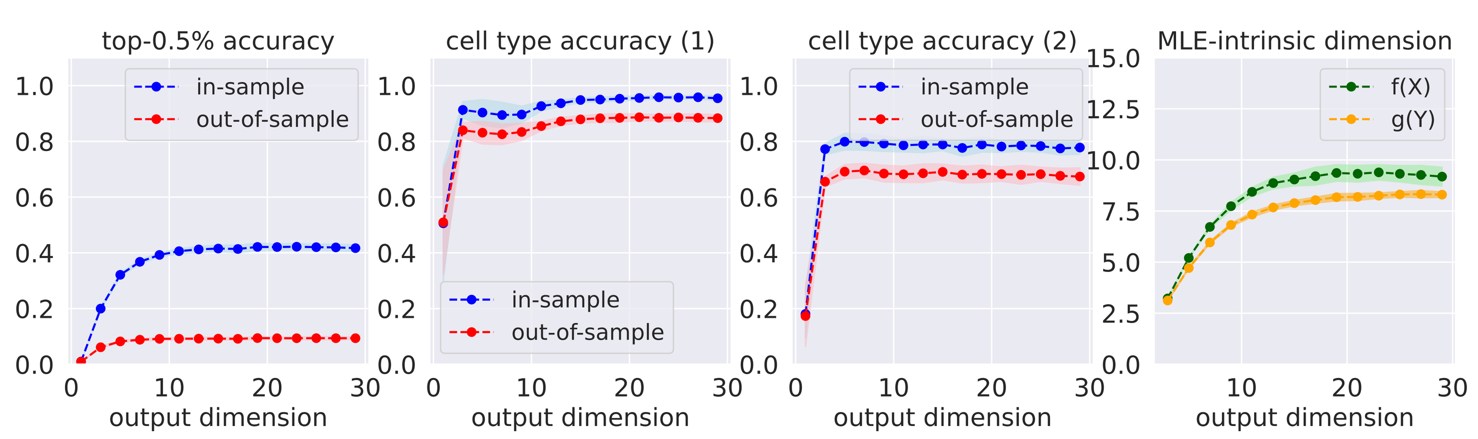
Multi-modal Contrastive Learning

maximize $\mathrm{sim}(f(X_i), g(Y_i))$, minimize $\mathrm{sim}(f(X_i), g(Y_j)),\ i \neq j$

- Contrastive language-image pre-training (CLIP)[2] has been the SOTA pipeline for multi-modal learning
- infoNCE loss $\mathcal{L}^N(f, g, \tau)$ with **temperature optimization**

$$-\frac{1}{N}\sum_{i\in[N]} \log \frac{\exp\left(\frac{\sigma(f(X_i),g(Y_i))}{\tau}\right)}{N^{-1}\sum_{j\in[N]}\exp\left(\frac{\sigma(f(X_i),g(Y_j))}{\tau}\right)} + \text{symmetric term}$$



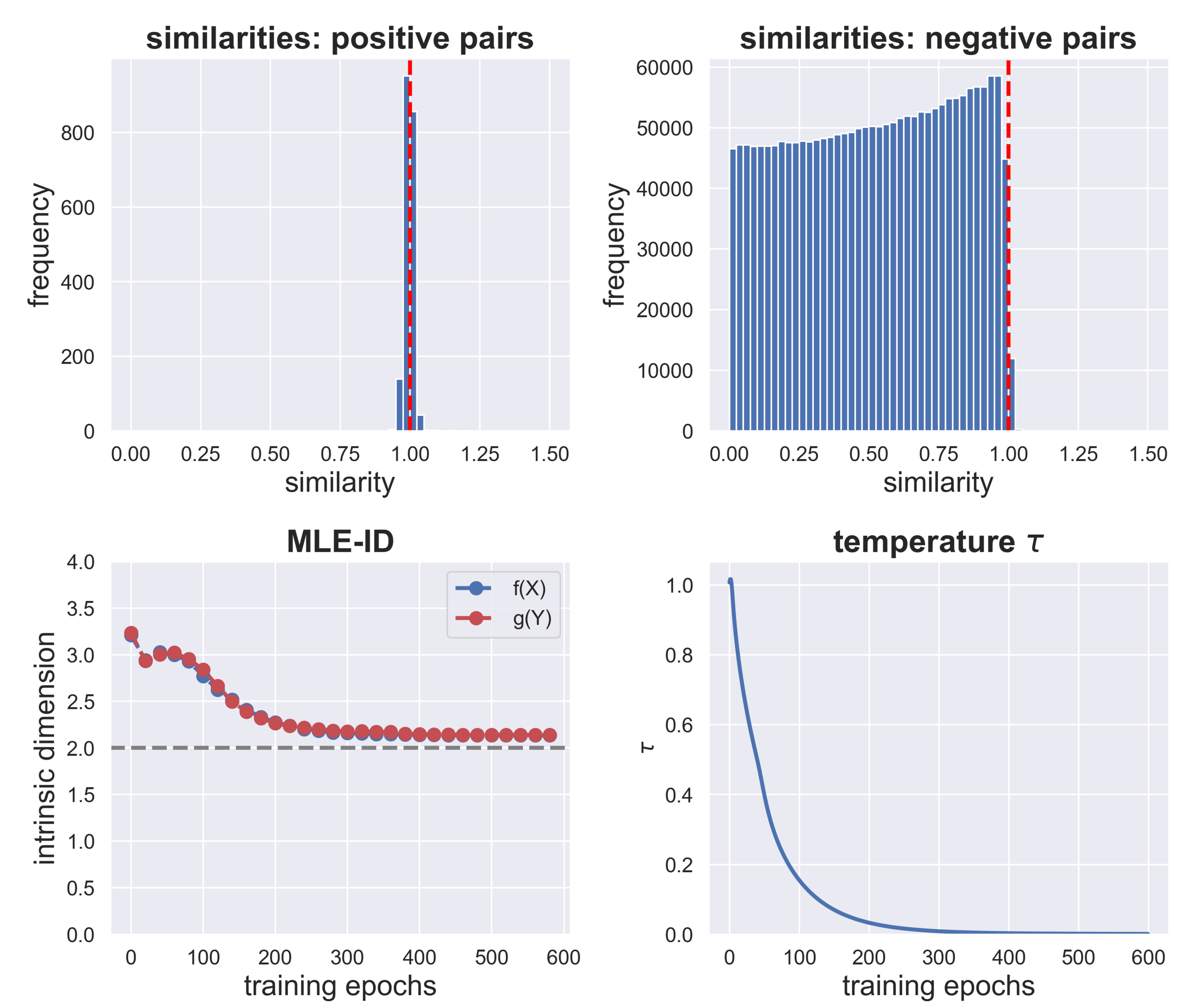


Question

- What representations does infoNCE learn in CLIP?
- How to tune or optimize temperature τ?

A toy example

$$Y_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{20}), \quad \xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{20-k^*}), \quad X_i = (Y_{i1}, \cdots, Y_{ik^*}, \xi_i^\top)^\top$$



- **Similarity concentration**: For positive pairs, cosine similarities concentrates around 1, while negative pairs are capped by 1
- **Intrinsic dimension adaptation**: Although output $d = 3$, representations with intrinsic dimension $k^* = 2$ are preferred
- **Temperature convergence**: The optimized temperature $\tau \to 0$

Intrinsic dimension: $\mathtt{ID}(f)$ is the smallest integer k s.t. there exist a measurable function $h: \mathbb{R}^{d_1} \to \mathbb{R}^d$ with $\dim(R(h)) = k$ and an injective measurable function $\phi: R(h) \to \mathbb{R}^d$ s.t. $f(x) = (\phi \circ h)(x)$ almost everywhere

Ideal Representations

- **Alignment**: with $m_\sigma(f,g) = \operatorname{ess\,sup}_{X \perp\!\!\!\perp \widetilde{Y}} \sigma(f(X), g(\widetilde{Y}))$

$$\mathcal{A}(\mathcal{H}) = \left\{(f,g) \in \mathcal{H}: \frac{f(X)}{\mathbb{E}\|f(X)\|} = \frac{g(Y)}{\mathbb{E}\|g(Y)\|},\quad \sigma(f(X), g(Y)) = m_\sigma(f,g)\ \text{a.s.}\right\}$$

- **Mutual information maximization**: $I_M^*(\mathcal{H}) = \sup_{\mathcal{H}} I(f_M(X); g_M(Y))$

$$\mathcal{W}(\mathcal{H}) = \left\{(f,g) \in \mathcal{H}: \liminf_{M\to+\infty} \left(I(f_M(X); g_M(Y)) - I_M^*(\mathcal{H})\right) \geq 0\right\}$$

$$\mathcal{V}(\mathcal{H}) = \mathcal{A}(\mathcal{H}) \cap \mathcal{W}(\mathcal{H})$$

- **Intrinsic dimension adaptation**:

Suppose $\mathcal{V}(\mathcal{H}) \neq \varnothing$. Then, for all $(f,g) \in \mathcal{V}(\mathcal{H})$, we have $\mathtt{ID}(f) = \mathtt{ID}(g) = k^*$, i.e., maps in $\mathcal{H}$ have the same intrinsic dimension k^*

Is any (approximate) minimizer of CLIP ideal?

$$\mathcal{O}_{\mathcal{L},\eta}(\mathcal{H}) = \left\{(f,g) \in \mathcal{H}: \exists\, \tau \geq \varepsilon(\eta),\ \limsup_{M\to+\infty} \left(\mathcal{L}(f_M, g_M, \tau) + 2I_M^*(\mathcal{H})\right) \leq 2\eta\right\}$$

Main results [1]

$$\mathcal{V}(\mathcal{H}) \neq \varnothing \implies \bigcap_{\eta\geq 0} \mathcal{O}_{\mathcal{L},\eta}(\mathcal{H}) \neq \varnothing.$$

In addition, for any $(f,g) \in \bigcap_{\eta\geq 0} \mathcal{O}_{\mathcal{L},\eta}(\mathcal{H})$,

- (similarity maximization) $\sigma(f(X), g(Y)) = m_\sigma(f,g)$ almost surely
- (intrinsic dimension adaptation) $\mathtt{ID}(f) = \mathtt{ID}(g) = k^*$
- (monotonicity in temperature) $\mathcal{L}(f, g, \tau)$ is increasing in τ
- (mutual information maximization) $(f,g) \in \mathcal{W}(\mathcal{H})$

References

[1] Gui, Y., Ma, C., and Ma, Z. (2025). Multi-modal contrastive learning adapts to intrinsic dimensions of shared latent variables. *NeurIPS*.

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.