Towards a Learning Theory of Representation Alignment

Francesco Insulla
Stanford University

Joint work with Shuo Huang (IIT, Italy) and Lorenzo Rosasco (MaLGa, IIT, MIT)

March 30, 2025

Data Modalities

LLMs:
Attention, GPT-3.5, $\cdots$

[Vaswani et al., 2017, Brown et al., 2020]

Vision:
ResNet, Vision Transformer (ViT), $\cdots$

[He et al., 2016, Dosovitskiy et al., 2020]

Multimodal data:
CLIP, GPT-4, Google's Gemini, $\cdots$

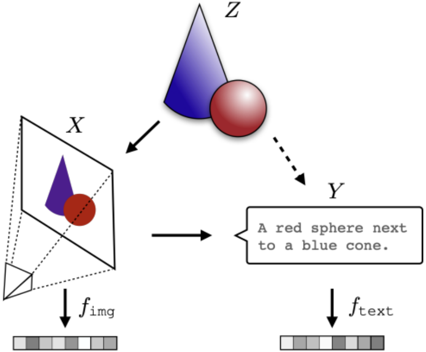
[Radford et al., 2021, OpenAI, 2023, Google, 2023]

Università di Genova MaLGa Stanford University MIT iit ISTITUTO ITALIANO DI TECNOLOGIA

Platonic Representation Hypothesis (PRH)

"Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces."

—– [Huh et al., 2024]



Università di Genova MaLGa Stanford University

How do we mathematically quantify and evaluate this alignment of feature learning across modalities?

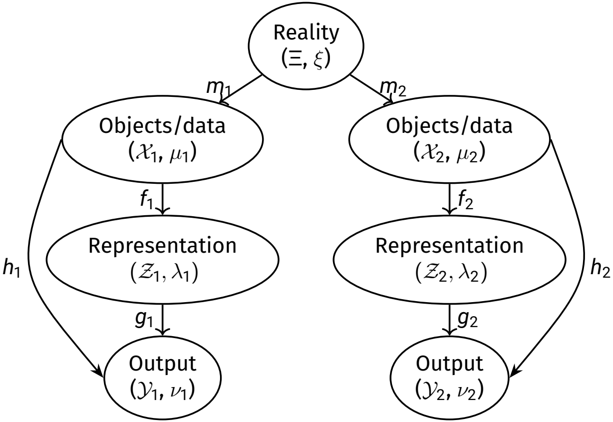
Outline

Review of Representation Alignment

Stitching: Task-Aware Alignment

Future Works

Setup and Notations



[Insulla et al., 2025]

Function spaces $\mathcal{H}_q := \{h_q : \mathcal{X}_q \to \mathcal{Y}_q \mid h_q = g_q \circ f_q\}, q = 1, 2.$

From Representation to Kernel

- Feature map $f : \mathcal{X} \to \mathcal{Z}$, Kernel $K(x, x') = \langle f(x), f(x') \rangle$, RKHS $\mathcal{H}_K$
- Kernel matrix $K_n \in \mathbb{R}^{n \times n}$ with sample $\{x_i\}_{i=1}^n$
- Integral operator $L_k : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ as

$$L_K g(x) = \int_{\mathcal{X}} K(x, x') g(x') d\mu(x')$$

Università di Genova MaLGa Stanford University MIT iit

Kernel Alignment (KA)

Empirical KA with kernel matrix $(K_{q,n})_{ij} = K_q(x_q^i, x_q^j), \quad i = 1, \dots, n$:

$$\widehat{A}(K_{1,n}, K_{2,n}) = \frac{\langle K_{1,n}, K_{2,n} \rangle_F}{\sqrt{\langle K_{1,n}, K_{1,n} \rangle_F \langle K_{2,n}, K_{2,n} \rangle_F}}$$

[Cristianini et al., 2001]

Population KA:

$$A(K_1, K_2) = \frac{\mathrm{Tr}\left(L_{K_1} L_{K_2}\right)}{\sqrt{\mathrm{Tr}\left(L_{K_1}^2\right) \mathrm{Tr}\left(L_{K_2}^2\right)}}$$

with

$$\mathrm{Tr}\left(L_{K_1} L_{K_2}\right) = \int K_1(x_1, x_1') K_2(x_2, x_2') d\mu(x_1, x_2) d\mu(x_1', x_2').$$

Statistic property: With probability at least $1 - \delta$, we have

$$|\widehat{A}(K_{1,n}, K_{2,n}) - A(K_1, K_2)| \leq \sqrt{(32/n) \log(2/\delta)}.$$

Spectral Interpretation of KA

- Eigen-pairs (η_ℓ, ϕ_ℓ) of L_K, then Mercer's theorem:

$$K(x, x') = \sum_\ell \eta_\ell \phi_\ell(x)\phi_\ell(x').$$

- Let $f_\ell = \sqrt{\eta_\ell}\phi_\ell$, then

$$A(K_1, K_2) = \frac{\sum_{i,j}\langle f_{1,i}, f_{2,j}\rangle}{\sqrt{\sum_i \eta_{1,i}^2 \sum_i \eta_{2,i}^2}} = \frac{\sum_{i,j} \eta_{1,i}\eta_{2,j}\langle \phi_{1,i}, \phi_{2,j}\rangle^2}{\sqrt{\sum_i \eta_{1,i}^2 \sum_i \eta_{2,i}^2}}.$$

Remark

1. Similarity between the eigenfunctions of the two integral operators
2. Let $[\Phi_{1,2}]_{i,j} = \langle \phi_{1,i}, \phi_{2,j}\rangle$. If $\Phi_{1,2} = I$, then $A(K_1, K_2) = \langle \hat{\eta}_1, \hat{\eta}_2\rangle$

Università di Genova MaLGa Stanford University

Distance Alignment (DA)

Suppose $d_q^2 = 2(1 - K_q)$, i.e. $d_q^2(x_q, x_q') = \|f_q(x_q) - f_q(x_q')\|^2$ and $K_q(x_q, x_q) = 1$, define

$$D(d_1, d_2) = \int (d_1^2(x, x') - d_2^2(x, x'))^2 d\mu(x) d\mu(x').$$

[Igel et al., 2007]

Equivalence: If $\|K_q\| = C$,

$$D(d_1, d_2) = 8C(1 - A(K_1, K_2)).$$

Università di Genova MaLGa Stanford University MIT IIT ISTITUTO ITALIANO DI TECNOLOGIA

KA to Independence Testing

- Hilbert-Schmidt Independence Criterion (**HSIC**): Cross-covariance operator $C_{1,2}[h_1, h_2] = \mathbb{E}_{x_1,x_2}[(h_1(x_1) - \mathbb{E}_{x_1}(h_1(x_1))(h_2(x_2) - \mathbb{E}_{x_2}(h_2(x_2))]$ for $h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2$.

$$\text{HSIC}(\mu, \mathcal{H}_1, \mathcal{H}_2) = \|C_{1,2}\|_{HS}^2,$$

[Gretton et al., 2005]

- Centered KA (**CKA**): replace L_K by HL_KH, $H = I - \mathbb{E}[\cdot]$

$$\text{CKA}(K_1, K_2) = \frac{\text{Tr}(HL_{K_1}HL_{K_2}H)}{\sqrt{\text{Tr}((HL_{K_1}H)^2)\text{Tr}((HL_{K_2}H)^2)}}$$

- **Equivalence** between HSIC and CKA:

$$\text{CKA}(K_1, K_2) = \frac{\text{HSIC}(\mathcal{H}_1, \mathcal{H}_2)}{\sqrt{\text{HSIC}(\mathcal{H}_1, \mathcal{H}_1)\text{HSIC}(\mathcal{H}_2, \mathcal{H}_2)}}.$$

[Kornblith et al., 2019]

KA to Measure Alignment

Maximum Mean Discrepancy (**MMD**):

$$\mathrm{MMD}(\mu_1, \mu_2; \mathcal{H}) = \sup_{h \in \mathcal{H}} \mathbb{E}\left[h(x_1) - h(x_2)\right].$$

Relationships

$$\mathrm{MMD}(\mu, \mu_1 \otimes \mu_2; \mathcal{H}_1 \otimes \mathcal{H}_2)^2 = \mathrm{HSIC}(\mu, \mathcal{H}_1, \mathcal{H}_2) = \left\|\mathbb{E}\left[K_{x_1} \otimes K_{x_2}\right]\right\|^2 = \|\Sigma_{1,2}\|_{HS}^2$$

Wrap-up

- Mathematically formalized the setup.
- KA: One metric of representation similarity (statistics property).
- KA and DA are mathematically equivalent.
- KA, HSIC, and MMD form a unified theoretical framework (integral operators, covariance operators, spectral analysis).

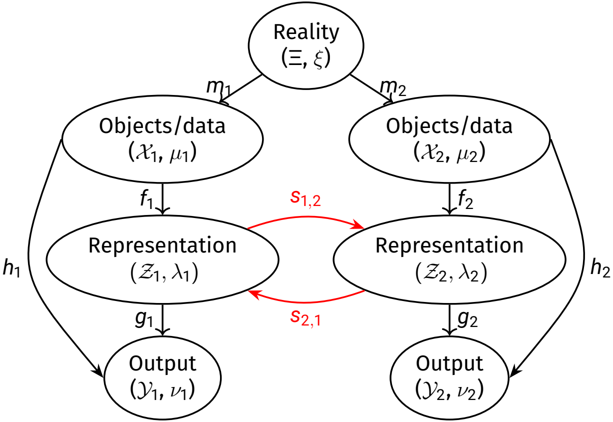
Outline

Review of Representation Alignment

Stitching: Task-Aware Alignment

Future Works

Stitching



- $\mathcal{H}_q := \{h_q : \mathcal{X}_q \to \mathcal{Y}_q | h_q = g_q \circ f_q,\ g_q \in \mathcal{G}_q, f_q \in \mathcal{F}_q\}$ with $q = 1, 2$.
- $\mathcal{H}_{1,2} := \{h_{1,2} = g_2 \circ s_{1,2} \circ f_1 : s_{1,2} \in \mathcal{S}_{1,2}\}$

Stitching Error

- The **risk** (least squares loss)

$$\mathcal{R}_q(h_q) = \int_{\mathcal{X}_q \times \mathcal{Y}_q} \|h_q(x) - y\|^2 \mathrm{d}\rho_q(x, y), \quad h_q \in \mathcal{H}_q.$$

- The **stitching error**

$$\mathcal{R}^{\text{stitch}}_{1,2}(s_{1,2}) := \mathcal{R}_2(g_2 \circ s_{1,2} \circ f_1) = \mathcal{R}_2(h_{1,2})$$

and the minimum

$$\mathcal{R}^{\text{stitch}}_{1,2}(\mathcal{S}_{1,2}) := \min_{s_{1,2} \in \mathcal{S}_{1,2}} \mathcal{R}_2(h_{1,2}) = \mathcal{R}_2(\mathcal{H}_{1,2}).$$

- The **excess stitching risk** (how usable f_1 is)

$$\mathcal{R}^{\text{stitch}}_{1,2}(\mathcal{S}_{1,2}) - \mathcal{R}_2(h_2).$$

Stitching between Output Layers

Lemma

Suppose $\dim(\mathcal{Y}_1) = \dim(\mathcal{Y}_2) = d$ *and* $\mathcal{R}_1 = \mathcal{R}_2$. *Let* $g_q \in \mathcal{G}_q$ ***be linear*** *with* $g_q(z_q) = W_q z_q$ *and* $W_q \in \mathbb{R}^{d \times d_q}$. *Let* $s_{1,2} : \mathcal{Z}_1 \to \mathcal{Z}_2$ *be linear with* $s_{1,2}(z_1) = S_{1,2} z_1$ *and* $S_{1,2} \in \mathbb{R}^{d_2 \times d_1}$. *Then* $\mathcal{R}^{stitch}_{1,2}(\mathcal{S}_{1,2}) = \mathcal{R}_1(\mathcal{H}_1)$.

Remark

- Linear stitching: stitching is not learning [Bansal et al., 2021]
- Stitching does not degrade performance.

[Insulla et al., 2025]

Stitching between Middle Layers

Let $\|M\|_\eta^2 = \langle M, M\mathrm{diag}(\eta)\rangle$.

Theorem

*Suppose g_2 **is κ_2-Lipschitz**. Again let $s_{1,2}$ be linear, identified with matrix $S_{1,2}$. With the spectral interpretations of $\Sigma_{1,2} = \mathbb{E}\left[f_1 f_2^T\right] = diag(\eta_1)^{1/2}\Phi_{1,2}diag(\eta_2)^{1/2}$ and $\tilde{A}_2 = \|I\|_{\eta_2} - \|\Phi_{1,2}\|_{\eta_2}^2$, we have*

$$\mathcal{R}_{1,2}^{stitch}(\mathcal{S}_{1,2}) \leq \mathcal{R}_2(h_2) + \kappa_2^2\tilde{A}_2 + 2\kappa_2(\tilde{A}_2\mathcal{R}_2(h_2))^{1/2}.$$

Remark

- $\tilde{A}_2$ is the misalignment
- If two representations are similar in the alignment sense, they are also similar in the stitching sense; however, the converse does not necessarily hold.

[Insulla et al., 2025]

Università di Genova MaLGa Stanford University MIT iit

Stitching Forward

Theorem

Let $\mathcal{Y}_1 = \mathcal{Y}_2$ *and* $\mathcal{R}_1 = \mathcal{R}_2 = \mathcal{R}$. *Let* $\mathcal{S}_{1,2} \circ \mathcal{G}_2 \subseteq \mathcal{G}_1$ *and let* g_q *be* κ_q*-Lipschitz for* $q = 1, 2$. *Then*

$$\mathcal{R}(\mathcal{H}_1) - \mathcal{R}(\mathcal{H}_2) \leq \mathcal{R}^{stitch}_{1,2}(\mathcal{S}_{1,2}) - \mathcal{R}(\mathcal{H}_2) \leq \kappa_2^2 \tilde{A}_2 + 2\kappa_2(\tilde{A}_2 \mathcal{R}(\mathcal{H}_2))^{1/2}.$$

[Insulla et al., 2025]

Remark

- It supports building universal models that share architectures across modalities as scale increases.
- Bansal et al., 2021 showed SGD minima have low stitching costs, which aligns with works that argue feature learning under SGD can be understood through the lens of adaptive kernels [Radhakrishnan et al., 2022].

Università di Genova MaLGa Stanford University

Outline

Review of Representation Alignment

Stitching: Task-Aware Alignment

Future Works

Future Work

- To understand which layer should be stitched
- Experimental verifications
- How to apply this to transfer learning

Thanks for listening!

Bibliography I

Bansal, Y., Nakkiran, P., and Barak, B. (2021).
Revisiting model stitching to compare neural representations.
Advances in Neural Information Processing Systems, 34:225–236.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020).
Language models are few-shot learners.
Advances in neural information processing systems, 33:1877–1901.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. (2001).
On kernel-target alignment.
Advances in Neural Information Processing Systems, 14.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020).
An image is worth 16x16 words: Transformers for image recognition at scale.
arXiv preprint arXiv:2010.11929.

Google (2023).
Gemini: a family of highly capable multimodal models.
arXiv preprint arXiv:2312.11805.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005).
Measuring statistical dependence with hilbert-schmidt norms.
In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer.

Bibliography II

He, K., Zhang, X., Ren, S., and Sun, J. (2016).
Deep residual learning for image recognition.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huh, M., Cheung, B., Wang, T., and Isola, P. (2024).
Position: The platonic representation hypothesis.
In *Forty-first International Conference on Machine Learning*.

Igel, C., Glasmachers, T., Mersch, B., Pfeifer, N., and Meinicke, P. (2007).
Gradient-based optimization of kernel-target alignment for sequence kernels applied to bacterial gene start detection.
IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(2):216–226.

Insulla, F., Huang, S., and Rosasco, L. (2025).
Towards a learning theory of representation alignment.
arXiv preprint arXiv:2502.14047.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019).
Similarity of neural network representations revisited.
In *International Conference on Machine Learning*, pages 3519–3529. PMLR.

OpenAI (2023).
GPT-4 technical report.
arXiv preprint arXiv:2303.08774.

Bibliography III

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.

Radhakrishnan, A., Beaglehole, D., Pandit, P., and Belkin, M. (2022).
Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn features.
arXiv preprint arXiv:2212.13881.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is all you need.
Advances in neural information processing systems, 30.