Introduction to Representational Convergence in Neural Networks

Data Mining & Quality Analytics Lab.

2026. 08. 14

발표자: 정진용

발표자 소개

❖ 정진용 (Jinyong Jeong)

- 고려대학교 산업경영공학과 석·박사 통합과정(2021.09~)
- Data Mining & Quality Analytics Lab. (김성범 교수님)

❖ 관심 연구 분야

- Out-Of-Distribution Generalization & Domain Generalization
- Semi-Supervised Learning & Class-Imbalanced Semi-Supervised Learning

❖ E-mail

- jy_jeong@korea.ac.kr

목차

1. Introduction
    - The Platonic Representation Hypothesis
2. Representations are converging
3. Why are representations converging?
4. What are the implications of convergence?
5. Counterexamples and limitations

Introduction

The platonic representation hypothesis

❖ Position: The Platonic Representation Hypothesis (ICML, 2024 position oral)

- 신경망들의 표현들이 수렴하고 있다는 가설을 제시하고 논증하는 논문
- 서로 다른 구조, 목적, 데이터로 학습된 모델들이 하나의 공유된 현실 표현으로 수렴한다고 주장함
- 수렴의 종착점을 플라톤 이데아에 빗대서 platonic representation이라고 명명함

Position: The Platonic Representation Hypothesis

Minyoung Huh [*1] Brian Cheung [*1] Tongzhou Wang [*1] Phillip Isola [*1]

Abstract

We argue that representations in AI models, particularly deep networks, are converging. First, we survey many examples of convergence in the literature: over time and across multiple domains, the ways by which different neural networks represent data are becoming more aligned. Next, we demonstrate convergence across data modalities: as vision models and language models get larger, they measure distance between datapoints in a more and more alike way. We hypothesize that this convergence is driving toward a shared statistical model of reality, akin to Plato's concept of an ideal reality. We term such a representation the *platonic representation* and discuss several possible selective pressures toward it. Finally, we discuss the implications of these trends, their limitations, and counterexamples to our analysis.

Project Page: phillipi.github.io/prh

Code: github.com/minyoungg/platonic-rep

1. Introduction

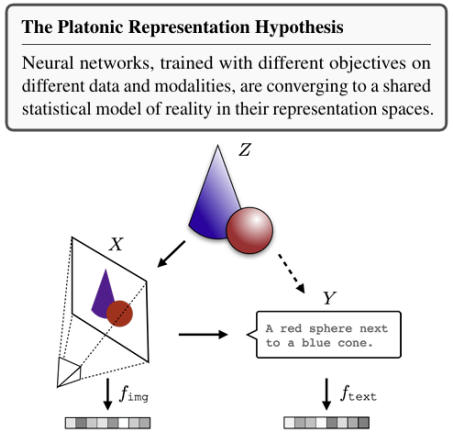


Figure 1. **The Platonic Representation Hypothesis:** Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z, and scaling model size, as well as data and task diversity, drives this convergence.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). Position: The Platonic Representation Hypothesis. International Conference on Machine Learning (ICML).

Introduction

The platonic representation hypothesis

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

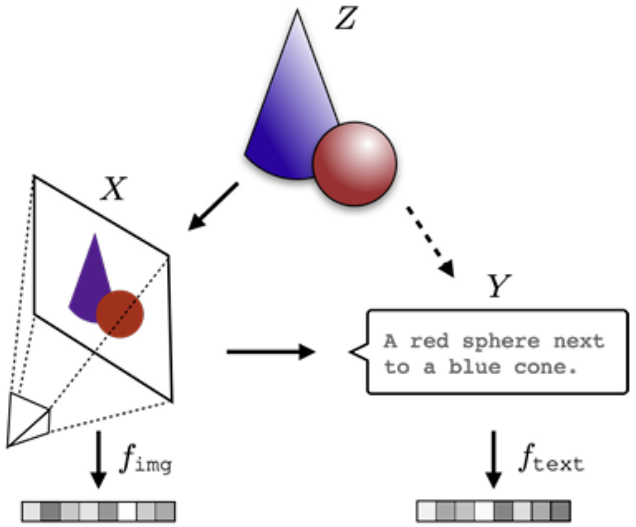
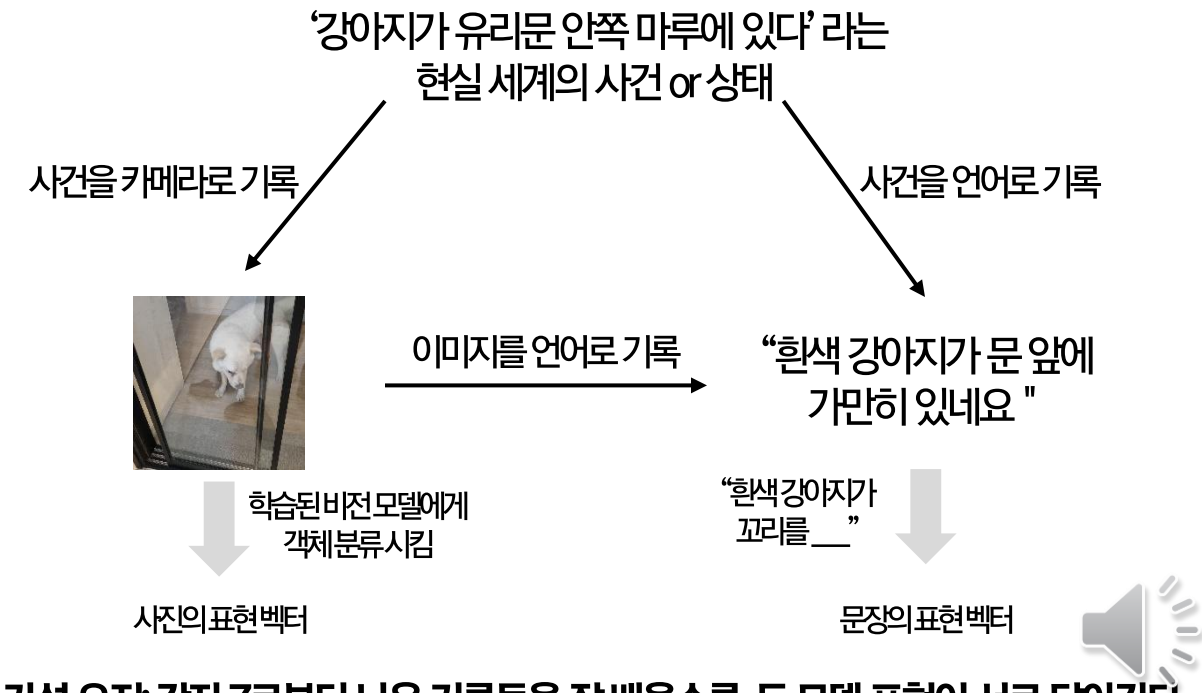


Figure 1. **The Platonic Representation Hypothesis:** Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z, and scaling model size, as well as data and task diversity, drives this convergence.



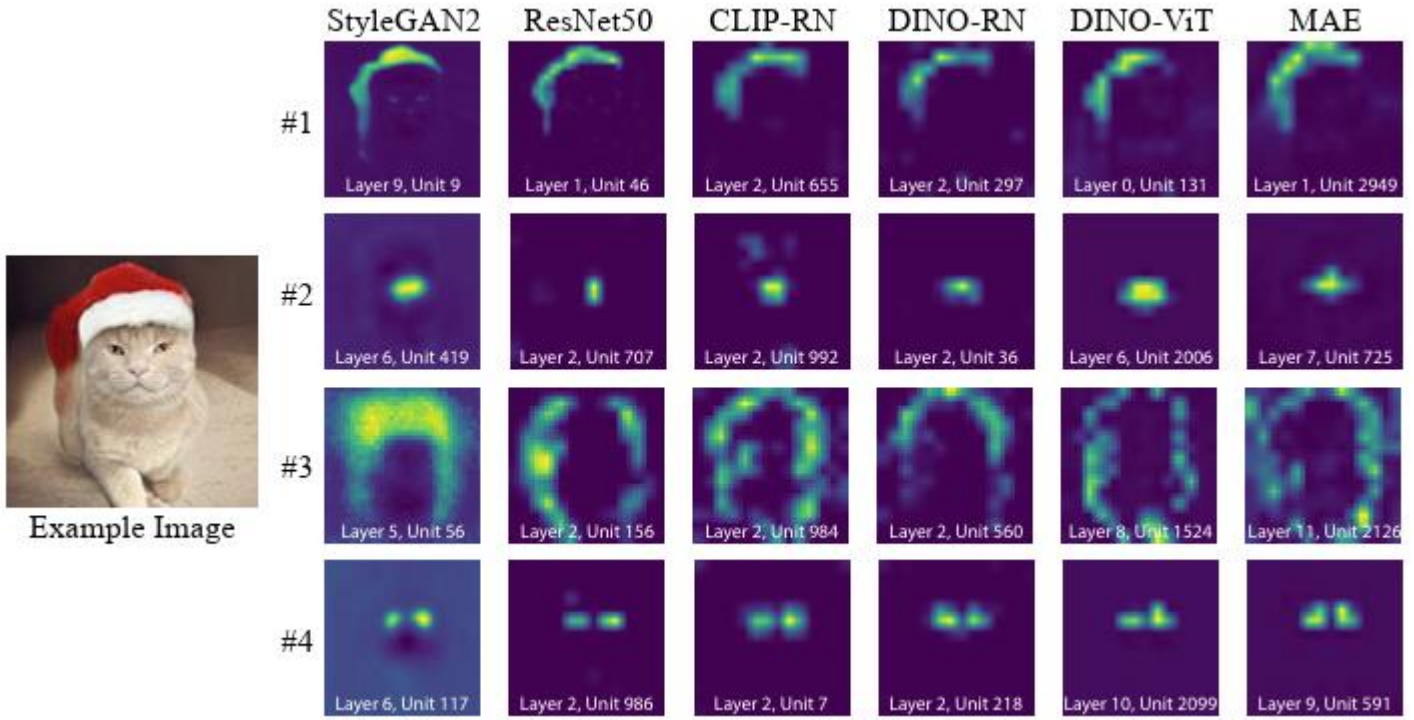
가설 요지: 각자 Z로부터 나온 기록들을 잘 배울수록, 두 모델 표현이 서로 닮아간다.

Representations are converging

Different models, with different architectures and objectives, can have aligned representations

❖ 서로 다른 비전 모델들이 똑같은 데이터에 대해서 어떻게 표현을 하고 있을까?

- 학습 방식이 전혀 다른 6개 모델에 같은 이미지를 입력해 봄
- 다양한 비전 모델들이 뉴런 수준에서 유사하게 작동하고 있는 것을 관찰할 수 있음 (Rosetta Neurons)



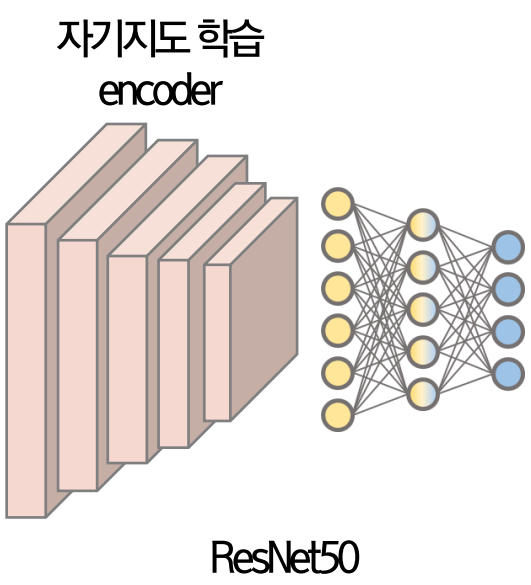
Dravid, A., Gandelsman, Y., Efros, A. A., & Shocher, A. (2023, October). Rosetta neurons: Mining the common units in a model zoo. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1934–1943). IEEE.

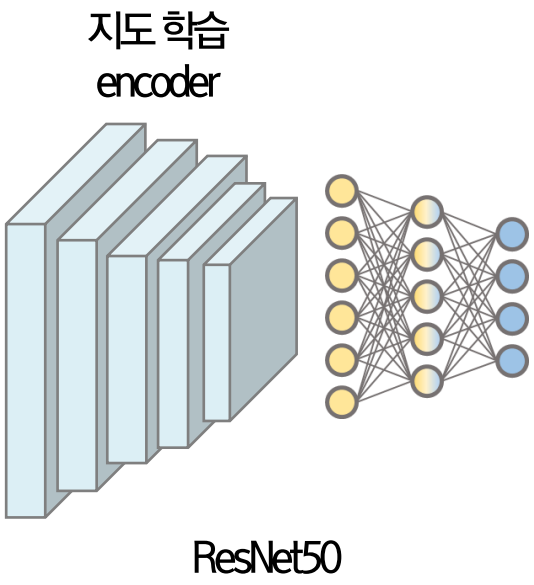
Representations are converging

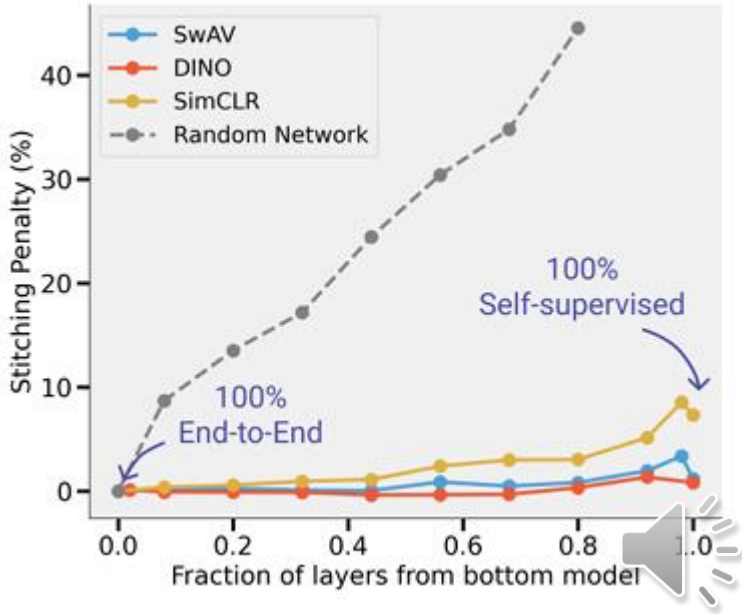
Different models, with different architectures and objectives, can have aligned representations

❖ 서로 다른 방식으로 학습된 비전 모델을 이어 붙여도 계속 잘 작동할 수 있을까?

- ImageNet으로 사전 학습된 모델들을 1x1 conv layer로 선형 변환 시켜서 이어 붙여도 성능 변화가 거의 없음
  - ➢ Stitching penalty : (이어 붙인 모델의 테스트 에러) – (오른쪽 모델, 지도 학습된 모델의 테스트 에러)
- 서로 다른 학습 방식을 적용해도, 두 표현들이 선형 변환 하나의 차이만큼 가까이에 있음







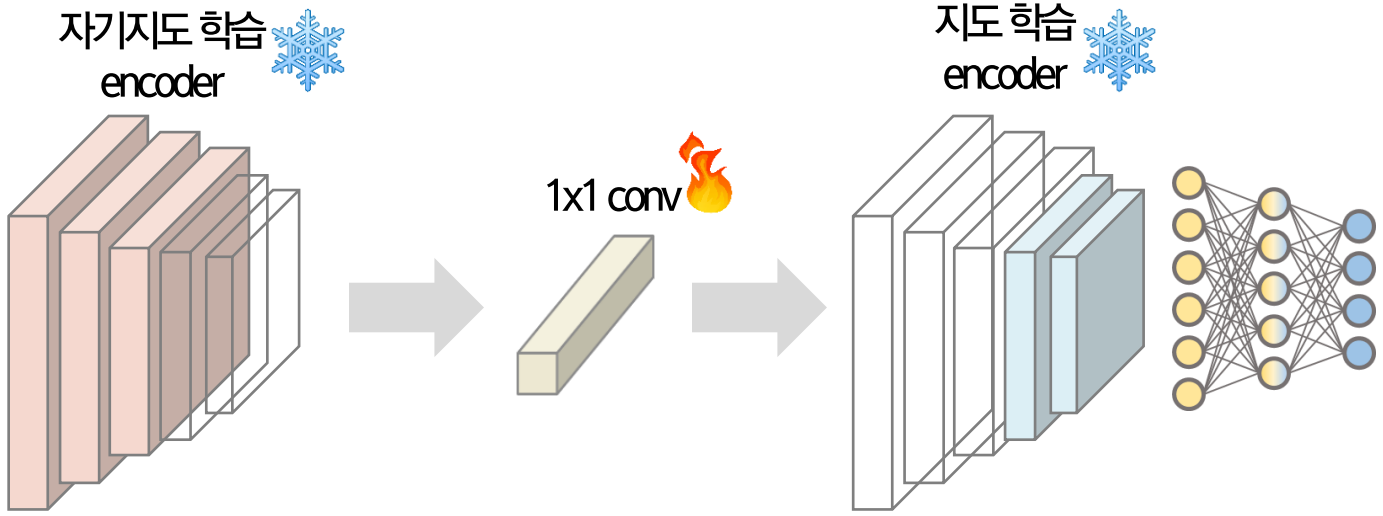
Bansal, Y., Nakkiran, P., & Barak, B. (2021). Revisiting model stitching to compare neural representations. Advances in neural information processing systems, 34, 225-236.

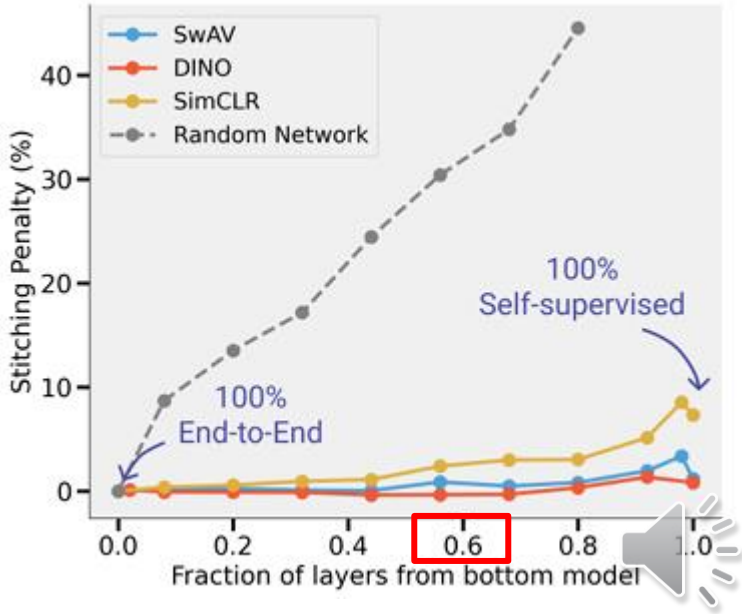
Representations are converging

Different models, with different architectures and objectives, can have aligned representations

❖ 서로 다른 방식으로 학습된 비전 모델을 이어 붙여도 계속 잘 작동할 수 있을까?

- ImageNet으로 사전 학습된 모델들을 1x1 conv layer로 선형 변환 시켜서 이어 붙여도 성능 변화가 거의 없음
  - ➢ Stitching penalty : (이어 붙인 모델의 테스트 에러) – (오른쪽 모델, 지도 학습된 모델의 테스트 에러)
- 서로 다른 학습 방식을 적용해도, 두 표현들이 선형 변환 하나의 차이만큼 가까이에 있음





→Label을 보고 배우든, 안 보고 배우든 두 모델은 서로 비슷한 표현을 보이고 있음

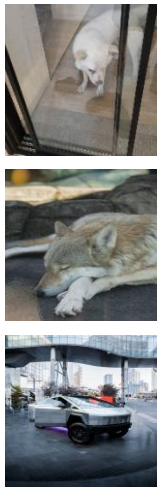
Bansal, Y., Nakkiran, P., & Barak, B. (2021). Revisiting model stitching to compare neural representations. Advances in neural information processing systems, 34, 225-236.

Measuring representational alignment

❖ 서로 다른 모델들의 표현이 닮아간다는 것은 어떻게 확인할 수 있을까?

- 커널 $K(x_i, x_j) = < f(x_i), f(x_j) >$를 사용해서 모든 표현들에 대한 관계를 나타냄
- 커널의 값들 자체를 전부 비교하는 CKA 지표 대신, mutual k-NN으로 가까운 이웃이 겹치는 비율만 측정
  - Representational alignment : 서로 다른 모델 간 표현이 얼마나 닮았는지 평가하는 지표

입력 K_{vision}



1.00	0.99	-0.37
0.99	1.00	-0.29
-0.37	-0.29	1.00

입력 K_{text}

"강아지 관련 문장"	1.00	0.96	-0.46
"늑대 관련 문장"	0.96	1.00	-0.21
"트럭 관련 문장"	-0.46	-0.21	1.00

가깝다 … 멀다

Measuring representational alignment

❖ 서로 다른 모델들의 표현이 닮아간다는 것은 어떻게 확인할 수 있을까?

- 커널 $K(x_i, x_j) = < f(x_i), f(x_j) >$를 사용해서 모든 표현들에 대한 관계를 나타냄
- 커널의 값들 자체를 전부 비교하는 CKA 지표 대신, mutual k-NN으로 가까운 이웃이 겹치는 비율만 측정
  - ➢ Representational alignment : 서로 다른 모델 간 표현이 얼마나 닮았는지 평가하는 지표

$$\text{sim}\left(K_{vision}, K_{text} \right) = ?$$

K_{vision}

1.00	0.99	-0.37
0.99	1.00	-0.29
-0.37	-0.29	1.00

K_{text}

1.00	0.96	-0.46
0.96	1.00	-0.21
-0.46	-0.21	1.00

Representations are converging

Different models, with different architectures and objectives, can have aligned representations

❖ 서로 다른 비전 모델들의 표현은 실제로 얼마나 닮아 있을까?

- 학습이 완료된 78개 비전 모델들에게 VTAB 문제들을 풀도록 하여서, 각 모델들의 범용 능력을 측정함
  - ➢ 모델 학습에 ImageNet-21k, ImageNet-1k, Places-365, synthetic dataset 사용됨
  - ➢ VTAB: 일상 사진 분류, 위성 및 의료 영상 분류, 개수 세기 및 거리와 방향 판단 등 총 19개 분류 task로 구성된 데이터셋
- 마찬가지로 Places-365 이미지 데이터셋을 활용해서, 각 모델들의 표현이 얼마나 가까운지 alignment를 측정함
- 성능이 좋은 모델들 사이에서는 서로 닮은 표현을 뱉고 있는 것을 확인할 수 있음

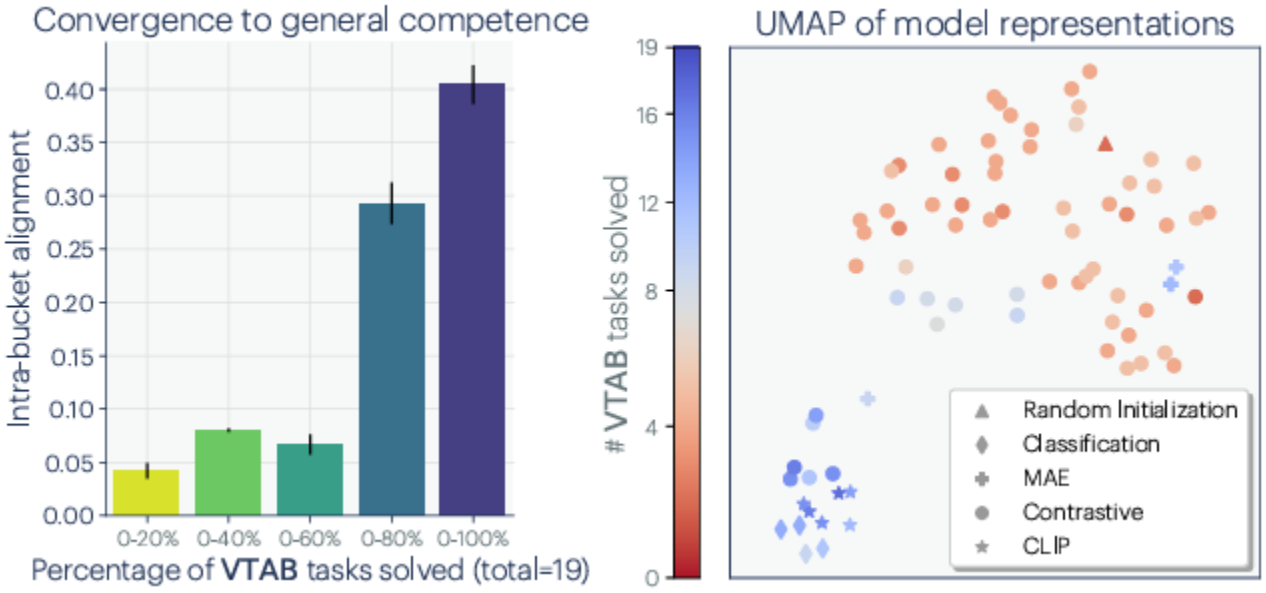


Figure 2. **VISION models converge as COMPETENCE increases:** We measure alignment among 78 models using mutual nearest-neighbors on Places-365 (Zhou et al., 2017), and evaluate their performance on downstream tasks from the Visual Task Adaptation Benchmark (VTAB; Zhai et al. (2019)). **LEFT:** Models that solve more VTAB tasks tend to be more aligned with each other. Error bars show standard error. **RIGHT:** We use UMAP to embed *models* into a 2D space, based on distance $= -\log(\text{alignment})$. More competent and general models (blue) have more similar representations.

Representations are converging

Different models, with different architectures and objectives, can have aligned representations

❖ 모달리티가 다른 언어 모델과 비전 모델의 표현도 서로 닮을 수 있을까?

- OpenWebText dataset을 사용해서 다양한 언어 모델들 성능을 평가함
- WIT 데이터셋을 사용해서 비전 모델과 다양한 언어 모델들 사이의 alignment를 측정함
- 언어를 잘 이해하는 모델일수록, 비전 모델과 표현이 닮아가는 것을 확인할 수 있음

WIT, 위키피디아의 (이미지, 캡션) 데이터셋 예시

"Half Dome, a granite dome at the eastern end of Yosemite Valley…"

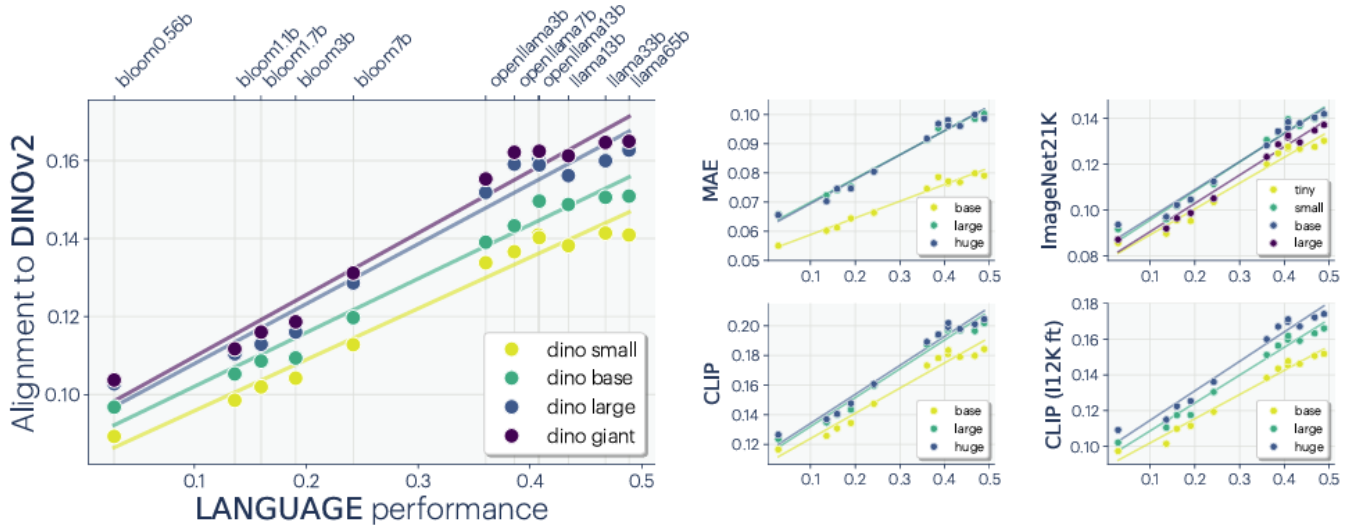


Figure 3. **LANGUAGE and VISION models align:** We measure alignment using mutual nearest-neighbor on the Wikipedia caption dataset (WIT) (Srinivasan et al., 2021). The x-axis is the language model performance measured over 4M tokens from the OpenWebText dataset (Gokaslan & Cohen, 2019) (see Appendix B for plots with model names). We measure performance using 1 − `bits-per-byte`, where `bits-per-byte` normalizes the cross-entropy by the total bytes in the input text string. The results show a linear relationship between language-vision alignment and language modeling score, where a general trend is that more capable language models align better with more capable vision models. We find that CLIP models, which are trained with explicit language supervision, exhibit a higher level of alignment. However, this alignment decreases after being fine-tuned on ImageNet classification (labeled CLIP (I12K ft)).

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., & Najork, M. (2021, July). Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval (pp. 2443-2449).

Representations are converging

Different models, with different architectures and objectives, can have aligned representations

❖ 이미지 데이터셋과 텍스트 데이터셋이 공유하는 정보가 많아질수록 표현들은 더욱 수렴할 수 있을까?

- DCI 데이터셋은 (이미지, 캡션)으로 구성되어 있으며, LLaMA3-8B-Instruct로 이미지 캡션 길이를 조절함
- Figure 3에서 사용된 언어 모델들 전체 alignment 평균을 측정함
- 공유하는 정보의 양이 많아질수록 표현 수렴 정도가 강해지는 것을 알 수 있음

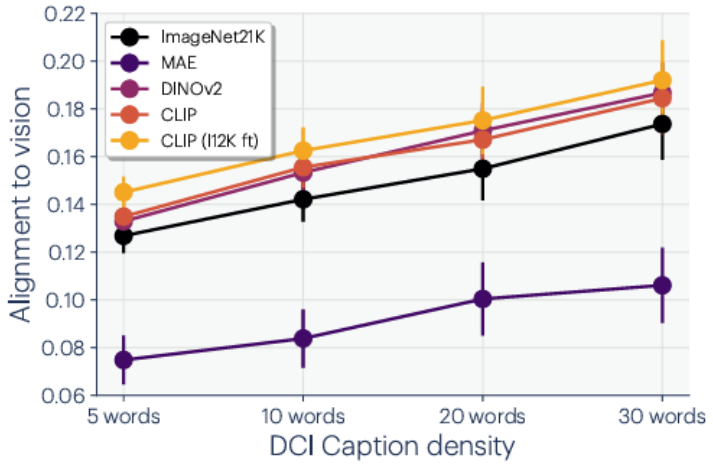


Figure 12. **Increasing caption density improves alignment:** We vary caption length using the Densely-Captioned-Images (DCI) dataset (Urbanek et al., 2023). Starting from a dense caption, we used LLaMA3-8B-Instruct (Meta, 2024) to summarize and generate coarse-grained captions. We compute the average alignment score across all vision and language models with standard deviation measured over the language models we evaluated. With denser captions, the mapping may become more bijective, leading to improved language-vision alignment scores.

"강아지."

↓

“갈색 강아지가 공원에서 공을 쫓는다."

↓

"늦은 오후의 공원 잔디밭에서, 갈색 리트리버가 노란 테니스공을 쫓아 전속력으로 달린다."

Representations are converging

Different models, with different architectures and objectives, can have aligned representations

❖ 이미지 데이터셋과 텍스트 데이터셋이 공유하는 정보가 많아질수록 표현들은 더욱 수렴할 수 있을까?

- DCI 데이터셋은 (이미지, 캡션)으로 구성되어 있으며, LLaMA3-8B-Instruct로 이미지 캡션 길이를 조절함
- Figure 3에서 사용된 언어 모델들 전체 alignment 평균을 측정함
- 공유하는 정보의 양이 많아질수록 표현 수렴 정도가 강해지는 것을 알 수 있음

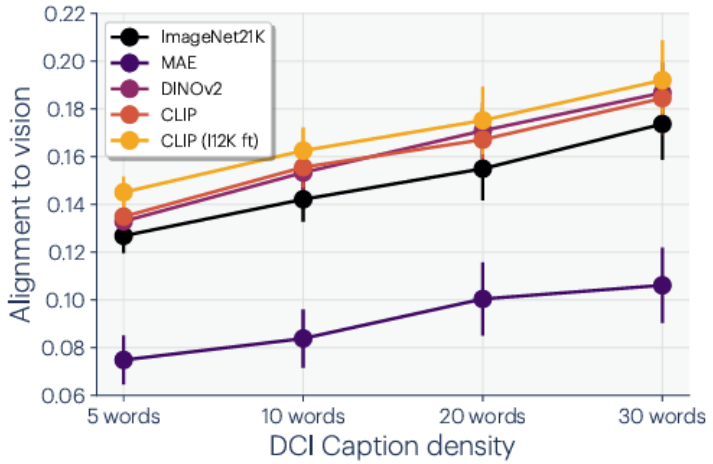
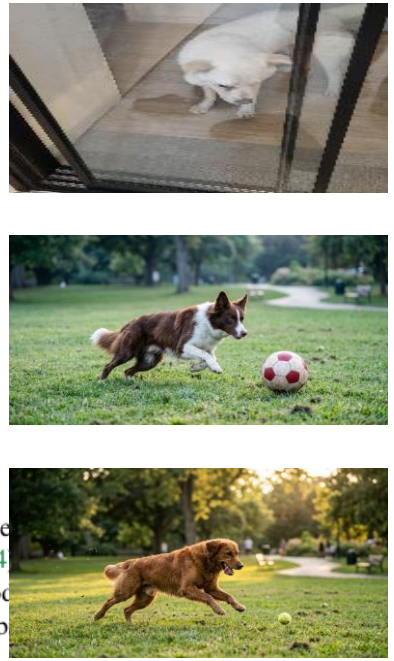


Figure 12. **Increasing caption density improves alignment:** We vary caption length using the Dense dataset (Urbanek et al., 2023). Starting from a dense caption, we used LLaMA3-8B-Instruct (Meta, 2024 coarse-grained captions. We compute the average alignment score across all vision and language mod measured over the language models we evaluated. With denser captions, the mapping may become more b language-vision alignment scores.



"강아지."

↓

“갈색 강아지가 공원에서 공을 쫓는다."

↓

"늦은 오후의 공원 잔디밭에서, 갈색 리트리버가 노란 테니스공을 쫓아 전속력으로 달린다."

Why are representations converging?

$$f^* = \underset{f \in \mathcal{F}}{argmin}\, \mathbb{E}_{x \sim dataset}[\mathcal{L}(f, x)] + \mathcal{R}(f)$$

Hypothesis space

모델이 표현할 수 있는
함수들의 공간

손실 함수

주어진 데이터셋 위에서
모델이 풀어야 하는 과제

규제항

같은 손실이라면 어떤 함수 f를
고를지 정하는 항

논문 주장: 이 세 요소가, 각각 표현을 수렴시키는 압력이 된다.

Why are representations converging?

Convergence via task generality

$$f^* = \underset{f \in \mathcal{F}}{argmin}\, \mathbb{E}_{x \sim dataset}[\mathcal{L}(f, x)] + \mathcal{R}(f)$$

❖ 모델이 해결해야 하는 과제가 많아질수록 가질 수 있는 표현 공간은 어떻게 될까?

- 과제마다 그 과제를 푸는 표현들의 집합이 있는데, 과제가 늘수록 전부를 만족하는 교집합은 줄어듬
- 모델들은 점점 하나의 목적함수만 가지고도 다양한 과제를 풀 수 있도록 만들어지고 있음
  - ➢ e.g., next token prediction, masked reconstruction, contrastive learning

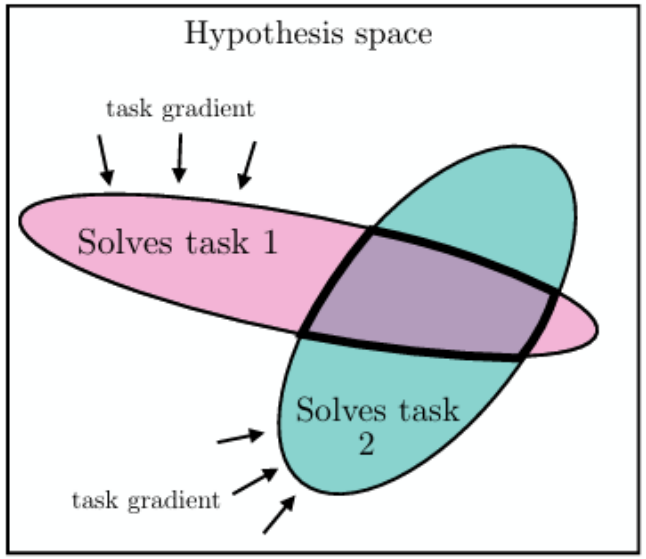
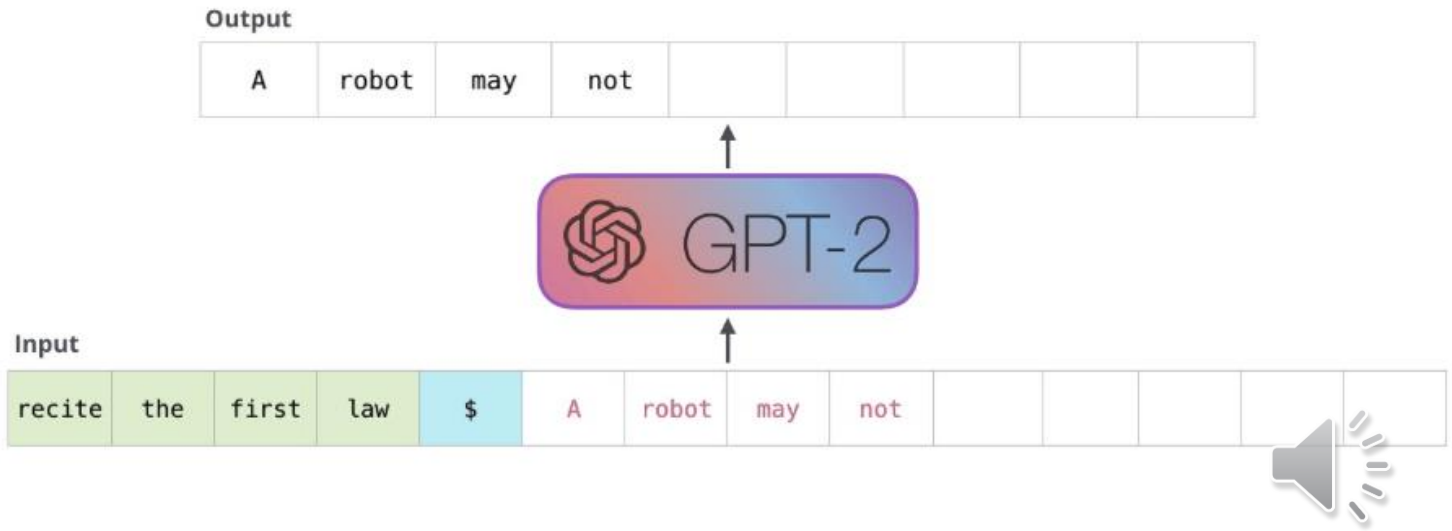


Figure 6. **The Multitask Scaling Hypothesis:** Models trained with an increasing number of tasks are subjected to pressure to learn a representation that can solve all the tasks.

목적 함수: next token prediction



Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Why are representations converging?

$$f^* = \underset{f \in \mathcal{F}}{argmin}\, \mathbb{E}_{x \sim dataset}[\mathcal{L}(f, x)] + \mathcal{R}(f)$$

Convergence via task generality

❖ 모델이 해결해야 하는 과제가 많아질수록 가질 수 있는 표현 공간은 어떻게 될까?

- 과제마다 그 과제를 푸는 표현들의 집합이 있는데, 과제가 늘수록 전부를 만족하는 교집합은 줄어듬
- 모델들은 점점 하나의 목적함수만 가지고도 다양한 과제를 풀 수 있도록 만들어지고 있음
  - ➢ e.g., next token prediction, masked reconstruction, contrastive learning

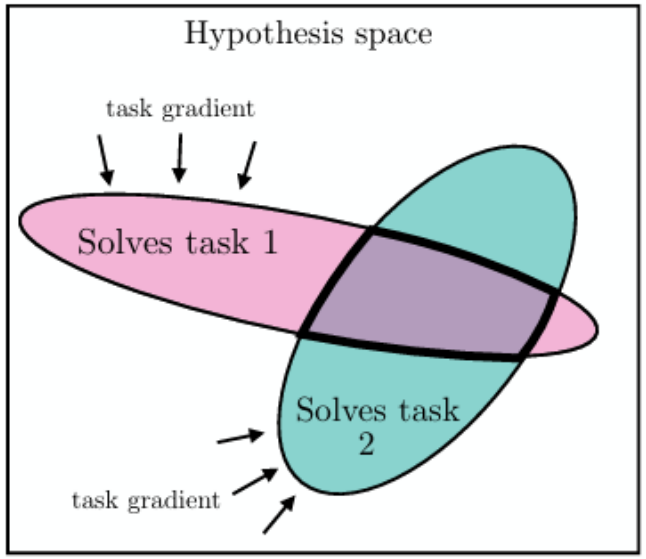
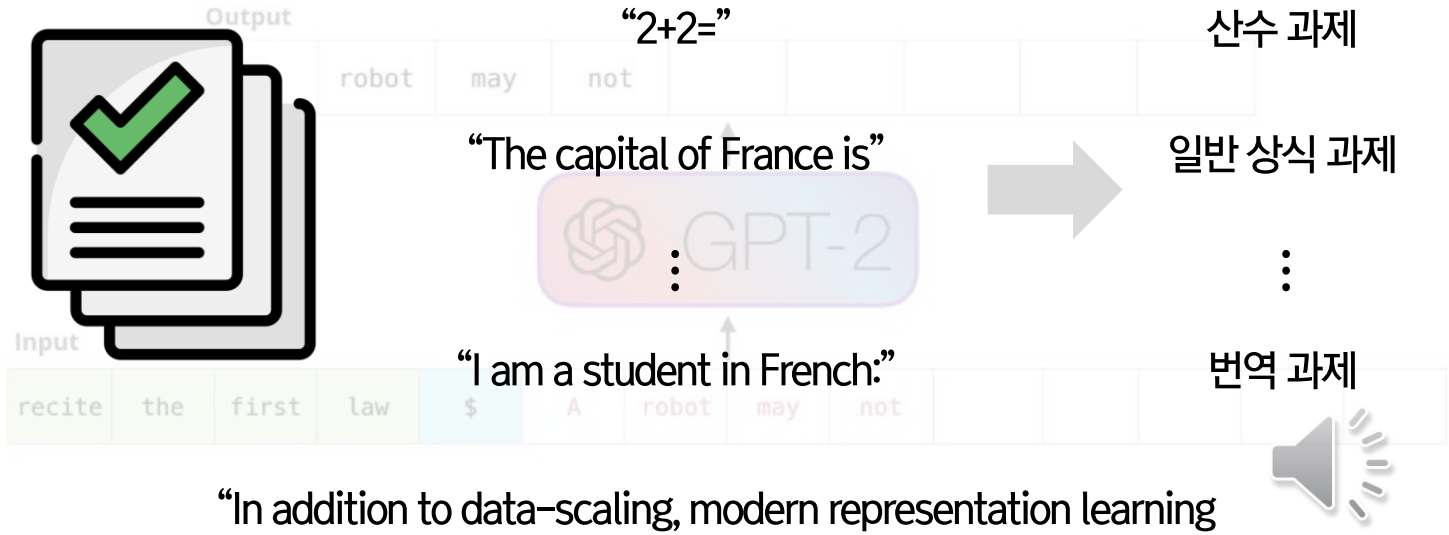


Figure 6. **The Multitask Scaling Hypothesis:** Models trained with an increasing number of tasks are subjected to pressure to learn a representation that can solve all the tasks.

목적 함수: next token prediction



"In addition to data-scaling, modern representation learning objectives directly optimize for multi-task solving"

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Why are representations converging?

Convergence via model capacity

$$f^* = \underset{f \in \mathcal{F}}{argmin}\, \mathbb{E}_{x \sim dataset}[\mathcal{L}(f, x)] + \mathcal{R}(f)$$

❖ 각 모델들은 다양한 과제를 푸느라 표현할 수 있는 공간이 줄어들고 있는데도 서로 같은 표현을 말할 수 있을까?

- 어떤 학습 목적에 대해서 전역적 최적의 표현이 있다고 가정 해본다면,
  - ➢ 모델 용량이 작을 때는 가설 공간 자체가 전역 최적 표현을 담지 못해서, 모델마다 서로 다른 표현을 하고 있음
  - ➢ 용량이 커져서 가설 공간에 전역 최적 표현이 포함되면, 모델들은 같은 표현을 뱉게 됨
- DINOv2 및 언어 모델들 용량이 커질수록, 서로의 가설 공간이 달라도 점점 표현 방식이 닮아가는 것을 확인할 수 있음

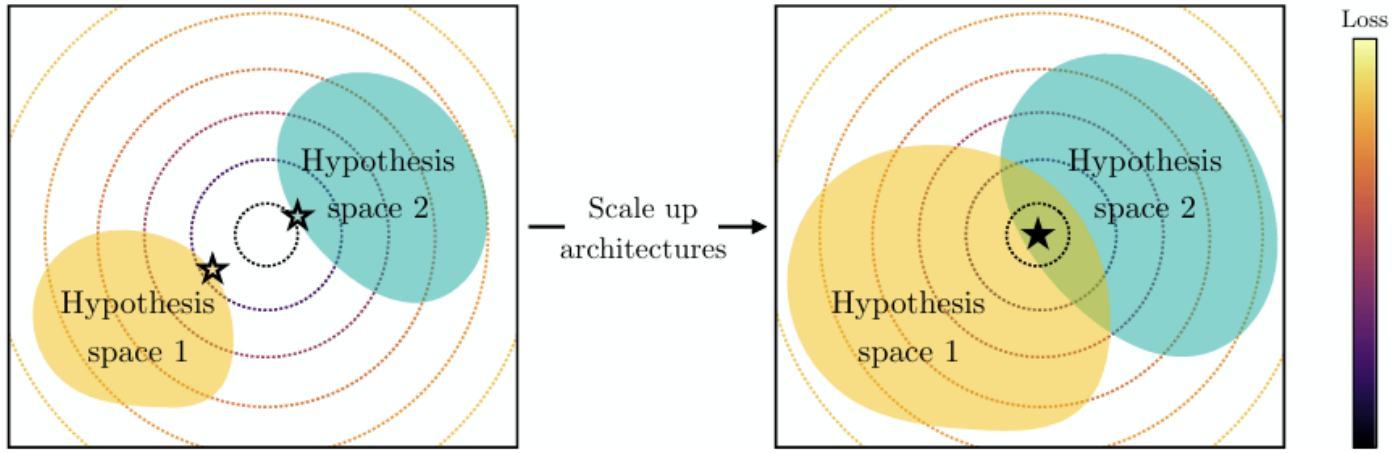
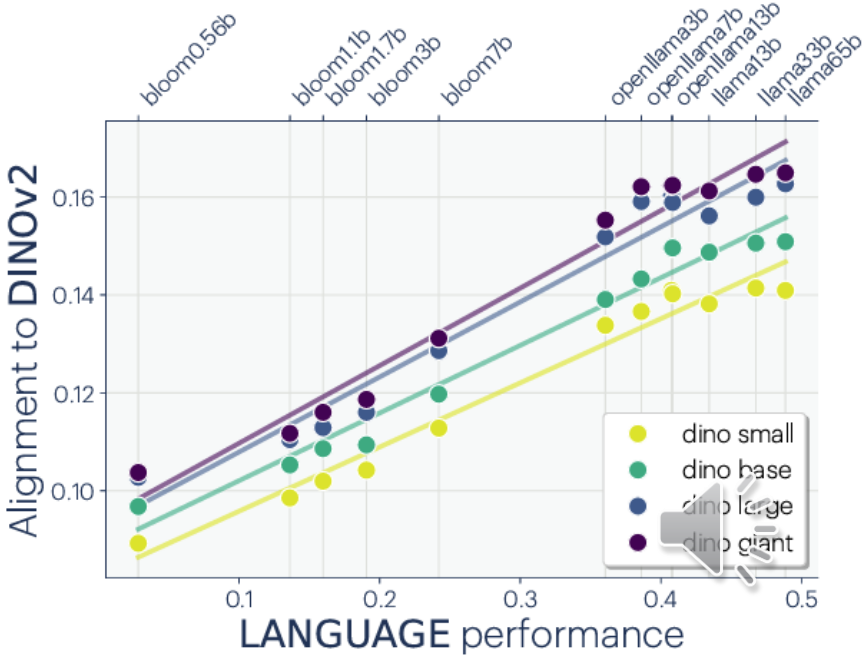


Figure 5. **The Capacity Hypothesis:** If an optimal representation exists in function space, larger hypothesis spaces are more likely to cover it. **LEFT:** Two small models might not cover the optimum and thus find *different* solutions (marked by outlined ☆). **RIGHT:** As the models become larger, they cover the optimum and converge to the same solution (marked by filled ★).



Why are representations converging?

$$f^* = \underset{f \in \mathcal{F}}{argmin}\, \mathbb{E}_{x \sim dataset}[\mathcal{L}(f, x)] + \mathcal{R}(f)$$

Convergence via simplicity bias

❖ 손실을 작게 만드는 표현은 여러 개일 수 있는데, 여전히 서로 다른 모델들의 표현들은 닮아갈 수 있을까?

- 딥러닝 모델에서 사용되는 규제항(weight decay, dropout 등) R(f)은 단순한 표현을 갖도록 만듦
- 경사 하강법을 사용해서 학습하는 딥러닝 모델들은 같은 손실을 갖는 표현들 중 단순한 표현으로 가려고 하는 경향을 보임
  - ➢ e.g., MNIST-CIFAR 데이터셋 실험

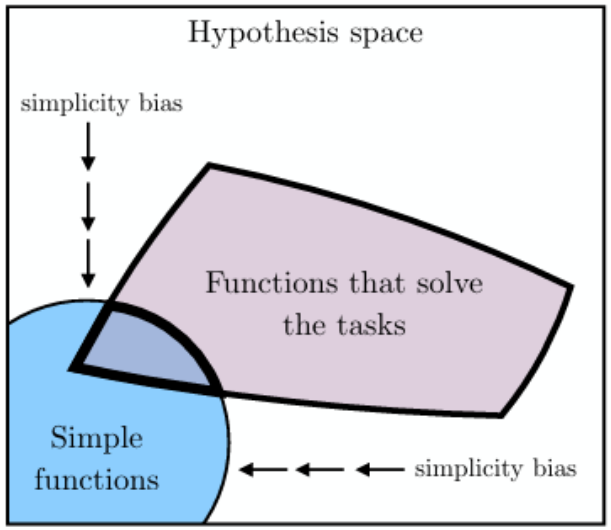
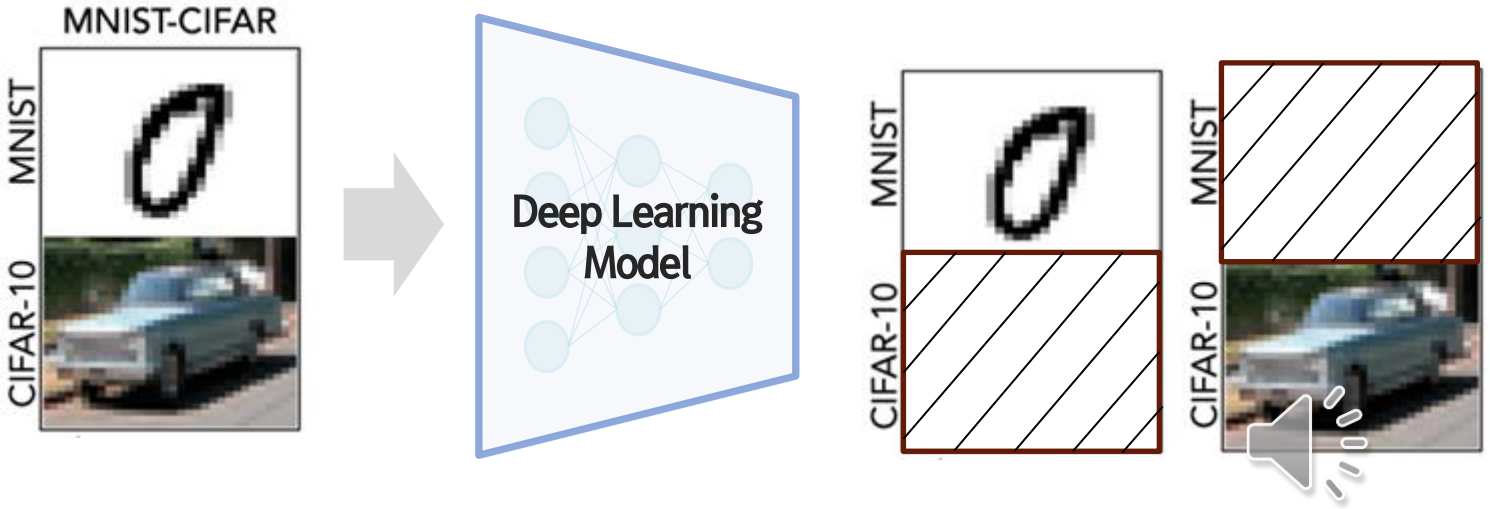


Figure 7. **The Simplicity Bias Hypothesis:** Larger models have larger coverage of all possible ways to fit the same data. However, the implicit simplicity biases of deep networks encourage larger models to find the simplest of these solutions.

모델은 MNIST와 CIFAR 중에 무엇을 배울까?



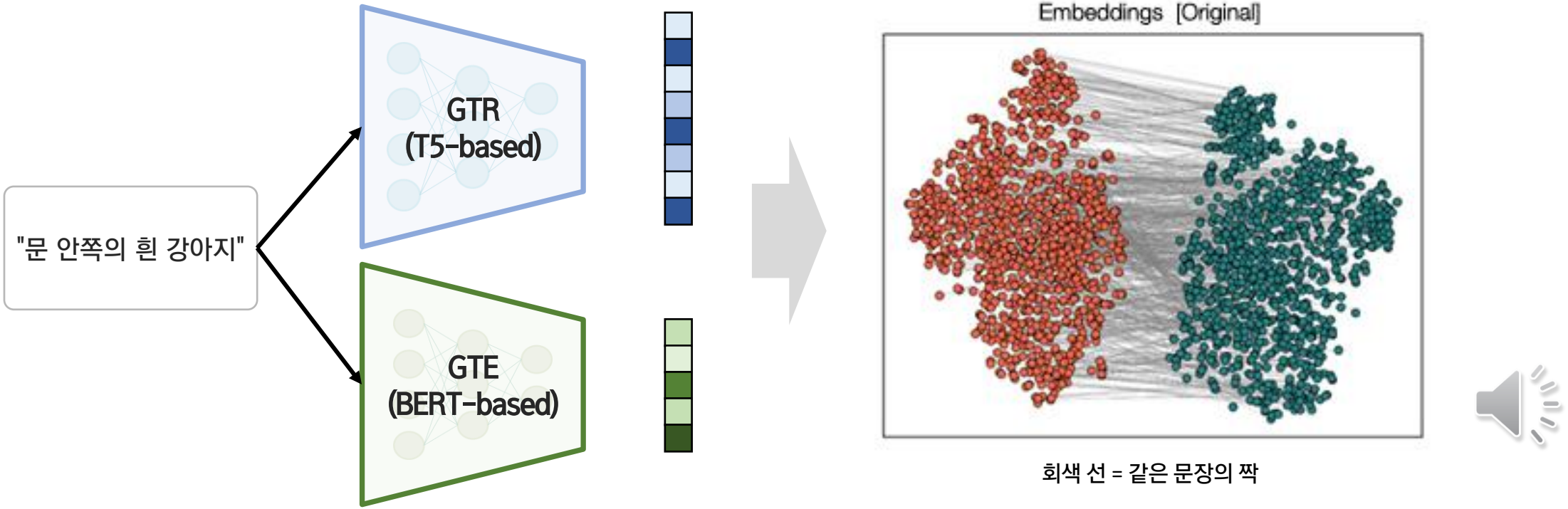
Shah, H., Tamuly, K., Raghunathan, A., Jain, P., & Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. Advances in neural information processing systems, 33, 9573-9585.

What are the implications of convergence?

Universal geometry enables translation between embedding spaces

❖ 같은 문장을 넣어도 텍스트 임베딩 모델마다 전혀 다른 벡터가 나옴

- 텍스트 임베딩 모델은 문장을 벡터로 바꿔서 의미가 비슷한 문장을 가까이 놓음
  - ➢ 의미 기반 검색 및 추천의 핵심(e.g., 벡터 DB)
- 같은 문장에서 출발했더라도 차원과 좌표계가 달라서 벡터들을 직접 비교를 할 수 없음



회색 선 = 같은 문장의 짝

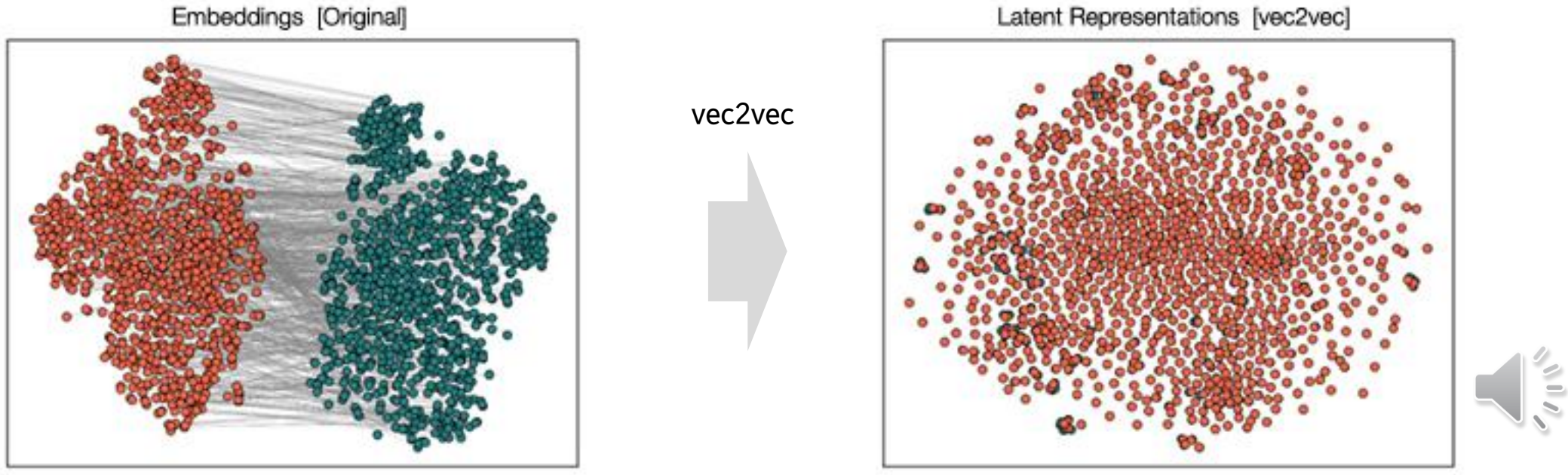
Jha, R., Zhang, C., Shmatikov, V., & Morris, J. (2026). Harnessing the universal geometry of embeddings. Advances in Neural Information Processing Systems, 38, 45963-45987.

What are the implications of convergence?

Universal geometry enables translation between embedding spaces

❖ 전혀 다른 위치에 있는 벡터들을 정렬시킬 수 있을까?

- Platonic representation hypothesis 핵심: 표현에 담긴 것은 좌표가 아니라 “무엇이 무엇과 가까운가”의 관계 구조
- 이러한 관계 구조는 모델이 서로 다르고, 표현 벡터가 서로 달라도 이미 닮아 있음
- vec2vec은 이 공유된 관계 구조만을 이용해서 두 벡터들을 어떤 하나의 잠재 공간으로 정렬시킴
  - ➢ 논문 시사점: 벡터 DB가 유출되면 임베딩 벡터만으로 원문 정보까지 역으로 추출할 수 있음



Jha, R., Zhang, C., Shmatikov, V., & Morris, J. (2026). Harnessing the universal geometry of embeddings. Advances in Neural Information Processing Systems, 38, 45963-45987.

Counterexamples and limitations

❖ 1. 근본적으로 다른 정보를 가진 두 모달리티는 같은 표현으로 수렴할 수 없음

- 언어에만 있는 정보, 이미지에만 있는 정보가 있는 한, 완전한 수렴은 없음

개기일식을 봤을 때의 경이로움을 언어로 온전히 전할 수 있을까?

"나는 표현의 자유를 믿는다."

이미지로 이 글을 온전히 담아낼 수 있을까?

❖ 2. 특수 목적을 갖는 분야에서는 표현 수렴이 힘듬

- 단백질 구조 예측(AlphaFold), 칩 배치, 물류 경로 최적화처럼 목적이 하나인 시스템

고맙습니다