Shaping the representations space:

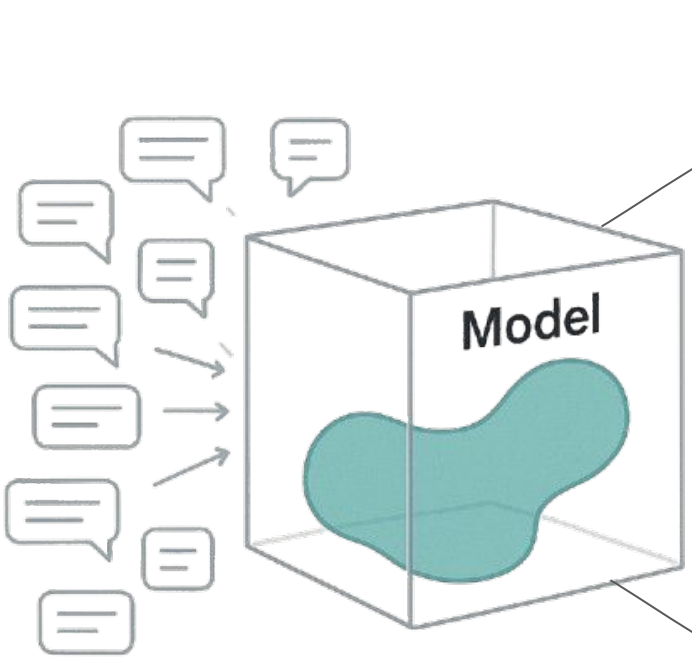
Geometric Diagnostics of Representations in Transformers

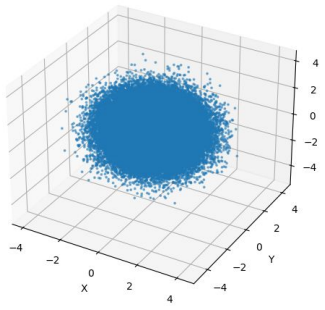
Lucia Domenichelli

Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC-CNR)
ItaliaNLP Lab – www.italianlp.it

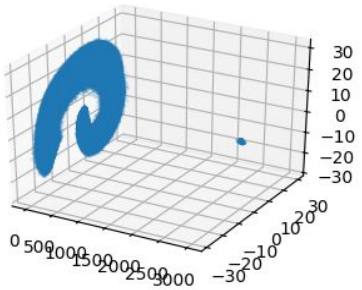
Lectures on Computational Linguistics 2025 - Milano, Università Cattolica del Sacro Cuore

The idea

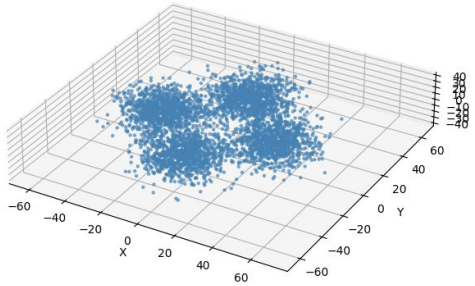




actually…

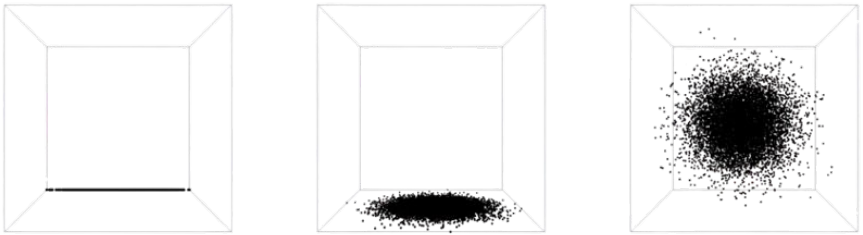


GPT-2



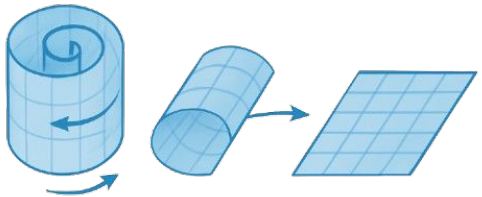
The metrics

Isotropy



A distribution is isotropic if its variance is uniformly distributed across all dimensions.

Intrinsic dimensionality



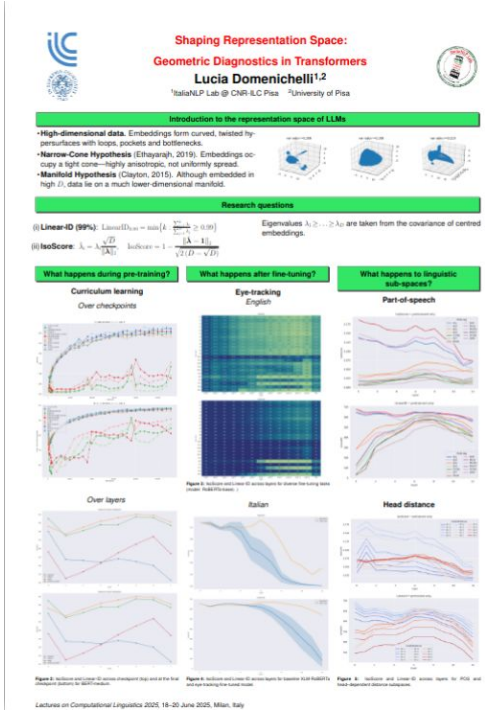
The dimension of a manifold, or *intrinsic dimension*ality, is the minimal number of degrees of freedom needed to describe it without information loss.

Research questions

Given some selected sentences and words representations:

- How do these spaces change between internal layers of Transformers?
- How do these spaces change during pretraining?
- How do these spaces change after fine-tuning?
- What about the individual subspaces of linguistic features of text?

Want to know more?



Shaping Representation Space:
Geometric Diagnostics in Transformers
Lucia Domenichelli[1,2]
[1]ItaliaNLP Lab @ CNR-ILC Pisa [2]University of Pisa

Introduction to the representation space of LLMs

- **High-dimensional data.** Embeddings form curved, twisted hypersurfaces with loops, pockets and bottlenecks.
- **Narrow-Cone Hypothesis** (Ethayarajh, 2019). Embeddings occupy a tight cone—highly anisotropic, not uniformly spread.
- **Manifold Hypothesis** (Clayton, 2015). Although embedded in high D, data lie on a much lower-dimensional manifold.

Research questions

(i) **Linear-ID (99%):** $\text{LinearID}_{99} = \min\left\{k : \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{D}\lambda_i} \geq 0.99\right\}$

(ii) **IsoScore:** $\lambda_i = \lambda\frac{\sqrt{D}}{\|\lambda\|_1}$ $\quad \text{IsoScore} = 1 - \frac{\|\lambda - \mathbf{1}\|_2}{\sqrt{2}\,(D - \sqrt{D})}$

Eigenvalues $\lambda_1 \geq \ldots \geq \lambda_D$ are taken from the covariance of centred embeddings.

What happens during pre-training? | What happens after fine-tuning? | What happens to linguistic sub-spaces?

Curriculum learning — *Over checkpoints*
Eye-tracking — *English*
Part-of-speech
Over layers
Italian
Head distance

Lectures on Computational Linguistics 2025, 18–20 June 2025, Milan, Italy

Want to share your opinion?

Please come to my poster!