Everything, Everywhere, All at Once: How learning order shapes the embedding space.

SLLD exam - Mod 1 & 2
Lucia Domenichelli - National PhD in Artificial Intelligence

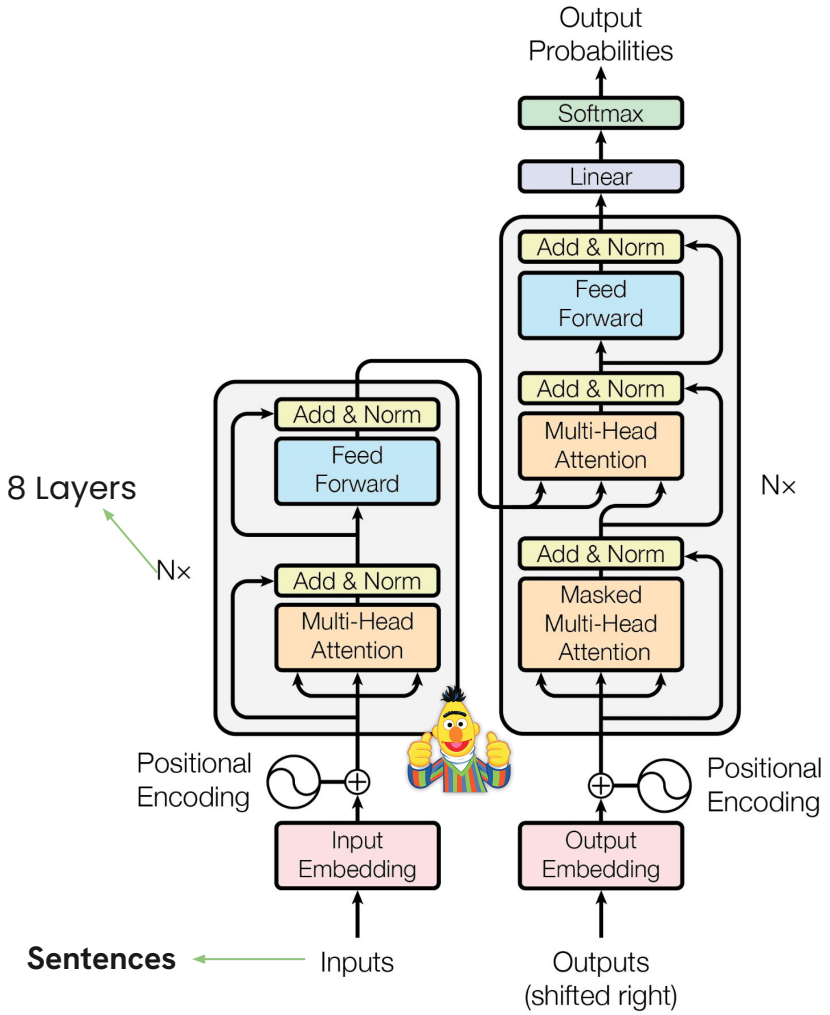
TABLE OF CONTENTS

HOW IT WORKS?

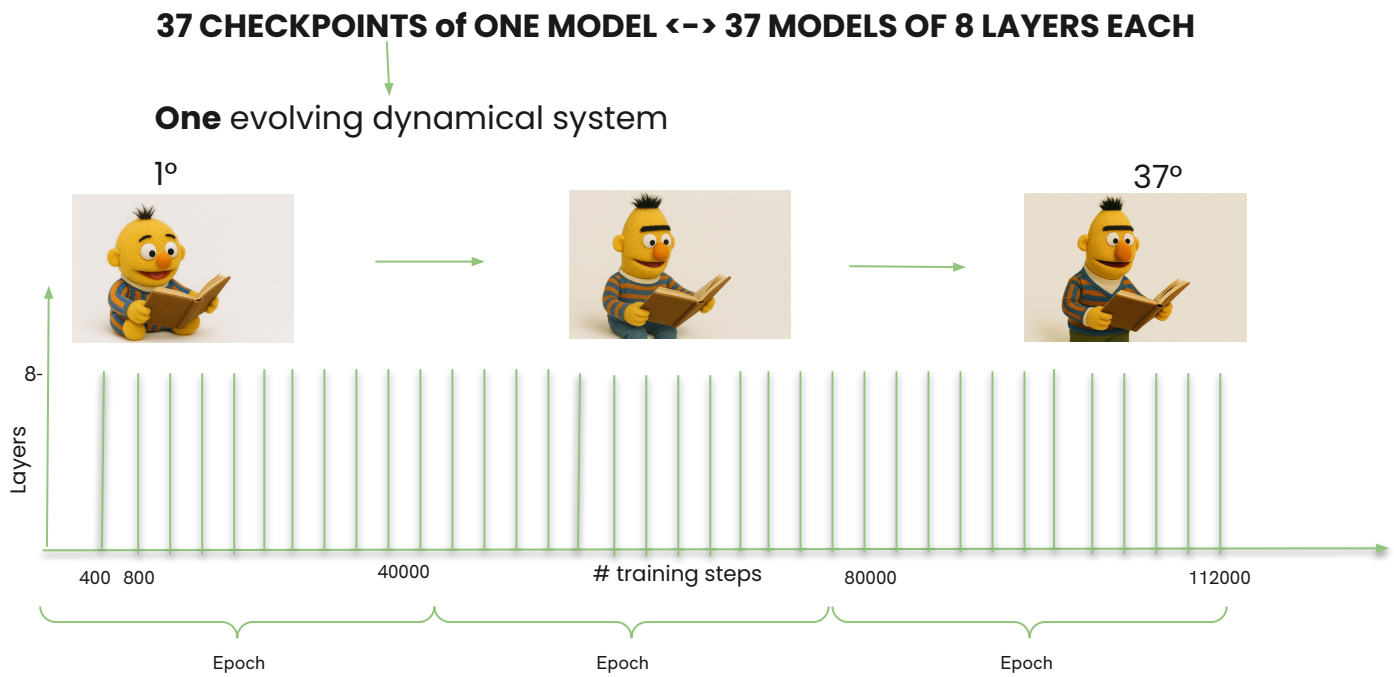
- Transformers Models have become ubiquitous in Natural Language Processing
- Pretraining + finetuning
- The task to solve in pretraining is called Language Modelling/Masked Language Modelling:

[CLS]Lucia likes to [MASK] climbing [SEP]

To predict



WE PRETRAINED BERT-MEDIUM (8 LAYERS - 512 FEATURES) ON AN ITALIAN DATASET!



02
Study
■ Objectives

CURRICULUM LEARNING

Gulpease = $89 + \frac{300 * (numero\ delle\ frasi) - 10 * (numero\ delle\ lettere)}{numero\ delle\ parole}$

ReadIT → makes use of linguistics annotations as well (lexical, morpho-syntactic and syntactic information.)

1) Does using a curriculum improve performance in **downstream tasks**?

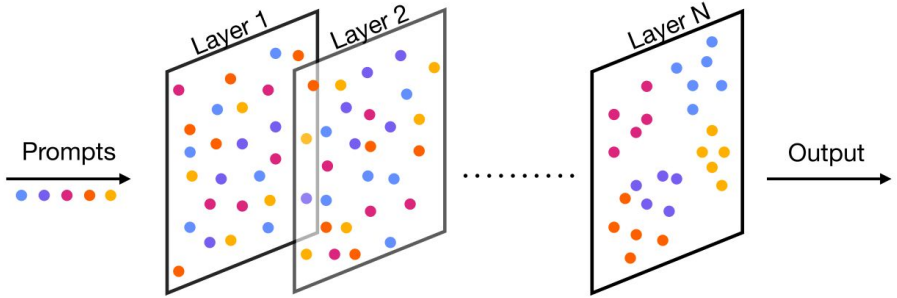
2) Does using a curriculum **enhance linguistic** knowledge of the model?

3) How does the geometry of the **representation space** change when pretraining using a curriculum?

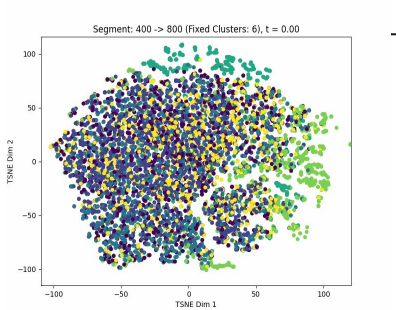
The strategies:

- ❏ Sentences' length
- ❏ Gulpease
- ❏ ReadIT
- ❏ 3 Random orderings

} Linguistically Motivated



THE DATASET



“Point cloud” of representations.

TRAINING

- 10020001 sentences longer than 6 tokens and up to 60 from half italian wikipedia.
- Avg length of a sentence 21.40 tokens
- 1000001 training +20000 test samples randomly selected

PROBING

- The UD-Treebank in Italian
- 8000 train + 2000 test sentences annotated with 136 linguistic features.
- Similar **data domain** but different **topic** and distribution of tokens from the training data.

EMBEDDINGS SPACE

- We extract representations of the 20000 test sentences
- We keep the **mean pooling of tokens** of a sentence as representation for each sentence.

SIMILAR RESULTS ON THE CLS TOKEN!

03

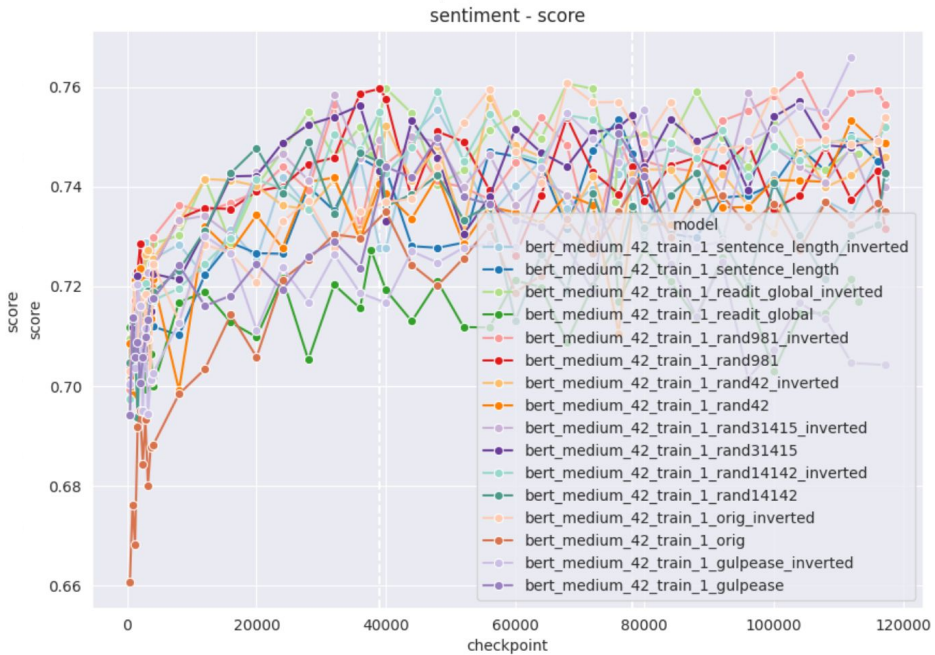
■ Our Work

FINETUNING

Sentiment analysis (*GLUE - sst2*, 2 classes)

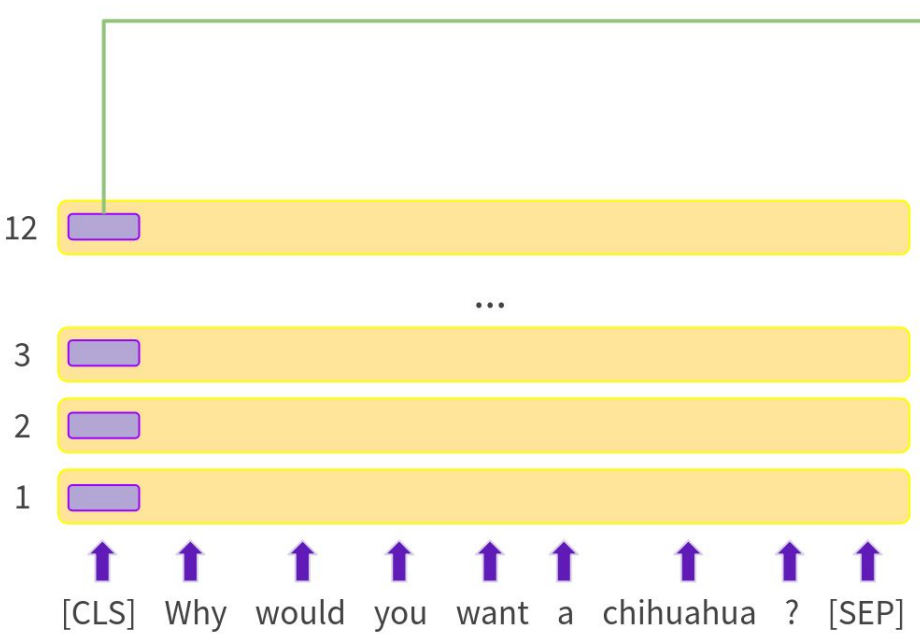
Complexity (*Human Complexity Judgment*, 7 classes)

POS tagging (235 classes)



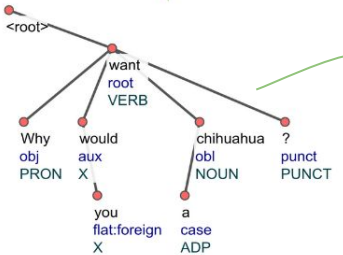
Not so promising!

PROBING TASKS



Linear Model

Which one?

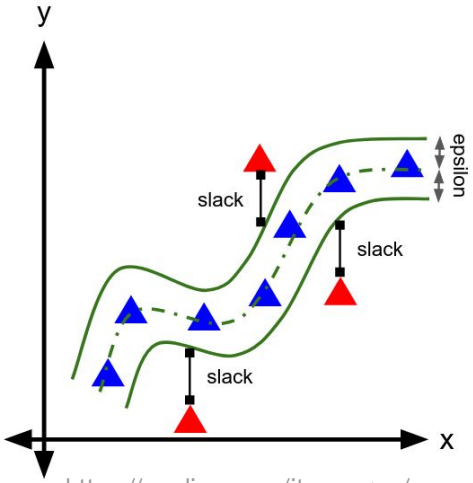


136 linguistic features

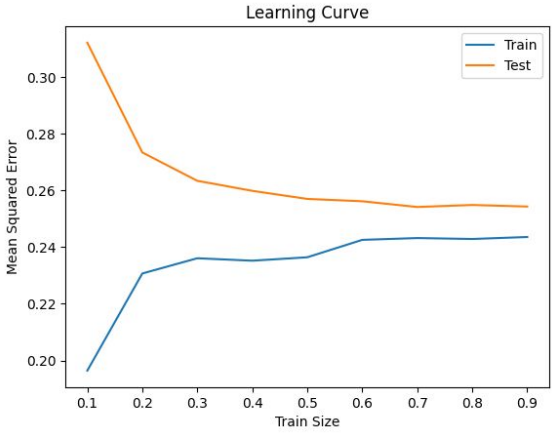
Probing Transformer Internal Representations by Alessio Miaschi.
Url:https://github.com/gsarti/lcl23-xnlm-lab/blob/main/notebooks/1.2_Probing.ipynb

SUPPORT VECTOR MACHINE

- A simple Support Vector Regressor (`LinearSVR` class from `scikit-learn`) with Linear Kernel was employed!
- Using "simple" linear models is usually the way to go in this context
- Prior to data processing, we normalize the data using a MinMax scaler to ensure features fall within the range [0, 1] for comparability.
- 5-Fold Cross Validation has been used to mitigate estimated variance

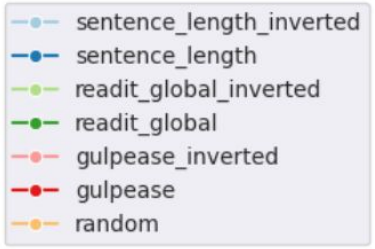


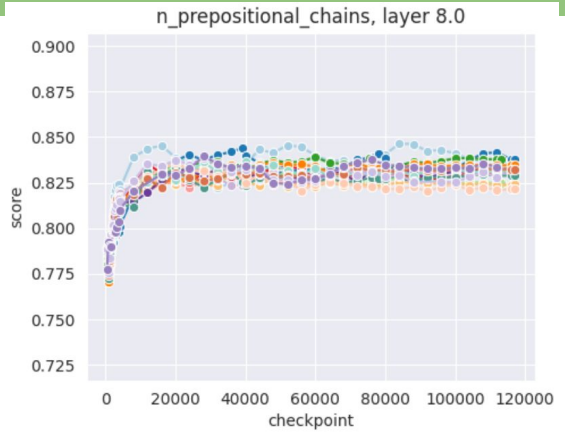
Source:https://medium.com/it-paragon/support-vector-machine-regression-cf65348b6345

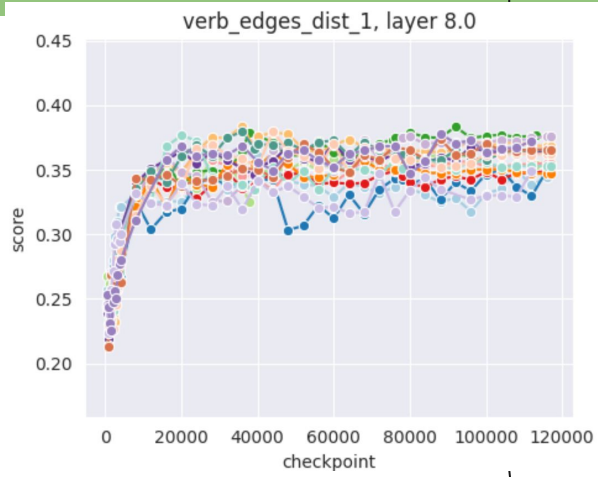


SOME RESULTS

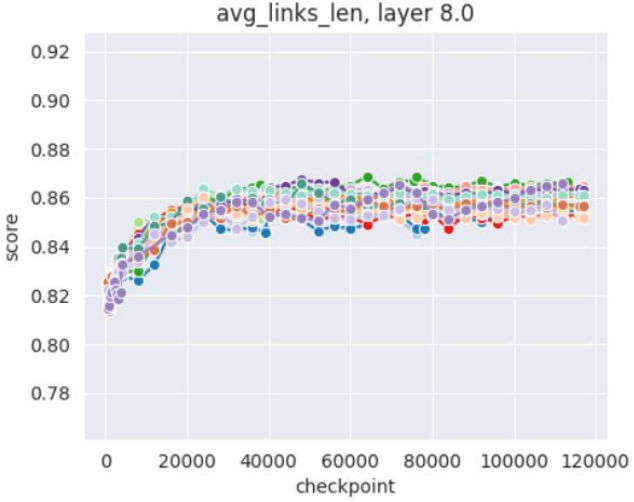
Not so much here...

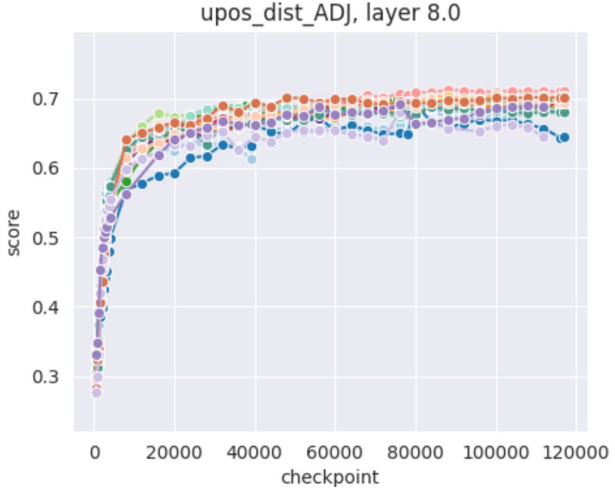


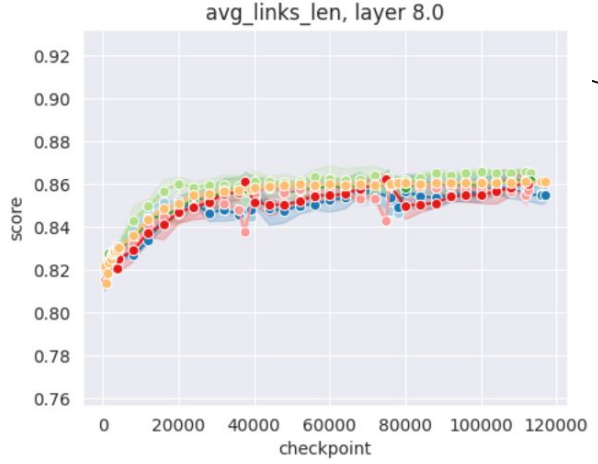




And many more!

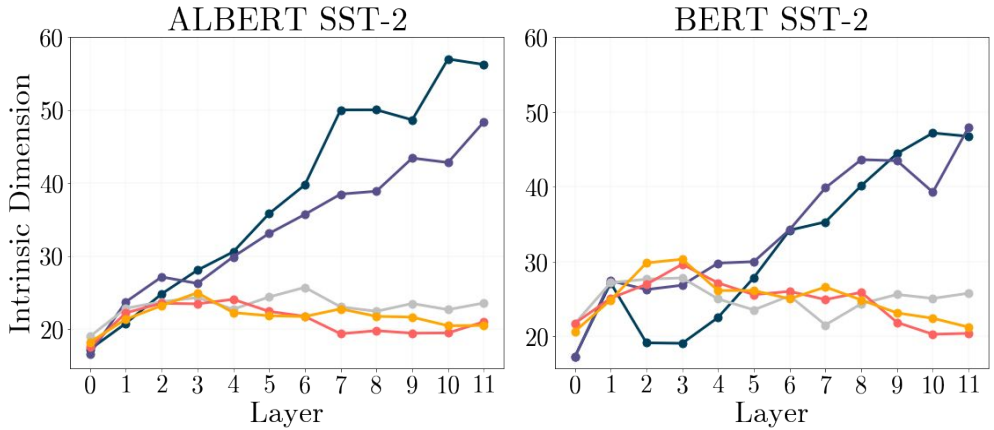






STUDYING THE GEOMETRY of LLMs

- **Narrow Cone Hypothesis (Ethayarajh 2019)** ->Models are not isotropic, i.e., they do not uniformly utilize the embedding space.
- **Manifold Hypothesis (Clayton, 2015)** -> High Dimensional data lies on a manifold of much lower dimensionality than the number of features

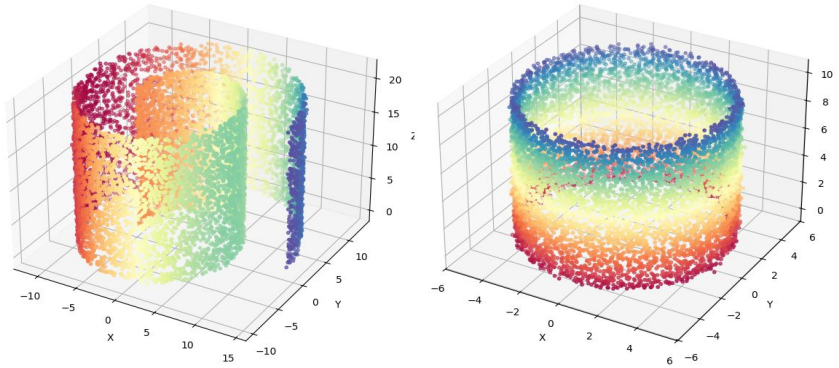


Intrinsic dimensionality estimate of ALBERT and BERT sentence embeddings from SST-2. (Rudman, 2023)

STUDYING THE GEOMETRY of LLMs

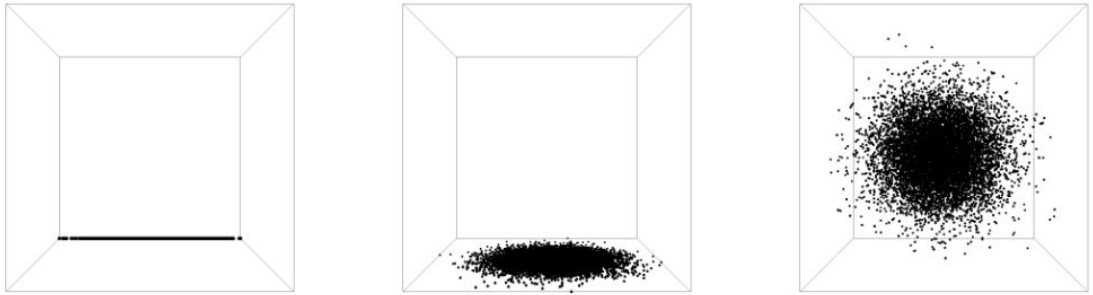
However...

- Curved and twisted hypersurfaces → Can we “iron” them?
- Complex topologies (no hyperplanes)



ISOTROPY

Isotropy: a distribution is *isotropic* if the variance of the data is uniformly distributed (i.e. the covariance matrix is proportional to the identity matrix).

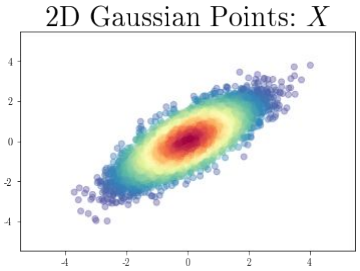


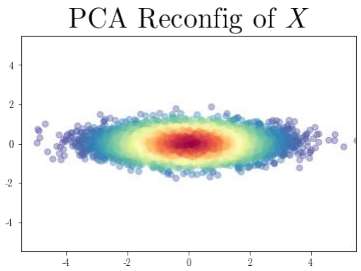
Left: line embedded in 3D space. **Middle**: circle embedded in 3D space. **Right**: Sphere in 3D space.

Why is it useful?

1. Low intrinsic dimensionality in later layers correlates to better model performance. (Recanatesi et al., 2019; Ansuini et al., 2019).

IsoScore





$$\begin{pmatrix} 1.80 & 0.00 \\ 0.00 & 0.20 \end{pmatrix}$$

1) Point cloud X in R^2.

2) Project X using PCA to get X^{PCA}.

3) Compute covariance of X^{PCA}.

$$\frac{\sqrt{2}}{||\begin{pmatrix} 1.80 & 0.20 \end{pmatrix}||} \cdot \begin{pmatrix} 1.80 & 0.20 \end{pmatrix} \rightarrow \frac{||\begin{pmatrix} 1.41 & 0.16 \end{pmatrix} - \begin{pmatrix} 1 & 1 \end{pmatrix}||}{\sqrt{2(2-\sqrt{2})}} \rightarrow 0.22$$

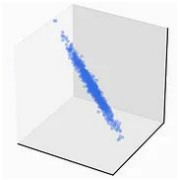
4) Normalize the diagonal of X^{PCA} to have the same norm as (1,1) to get V^{PCA}.

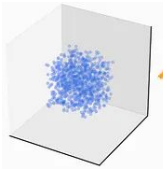
5) Calculate the Euclidean distance between V^{PCA} and (1,1) then normalize.

6) Linearly rescale to be in the interval [0,1].

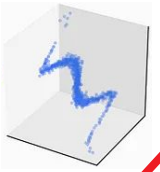
INTRINSIC DIMENSIONALITY

Linear





Nonlinear



We employ a simple measure derived from PCA.

How many **principal components** are needed to retain **99% of the variance** in the data.

TwoNN score

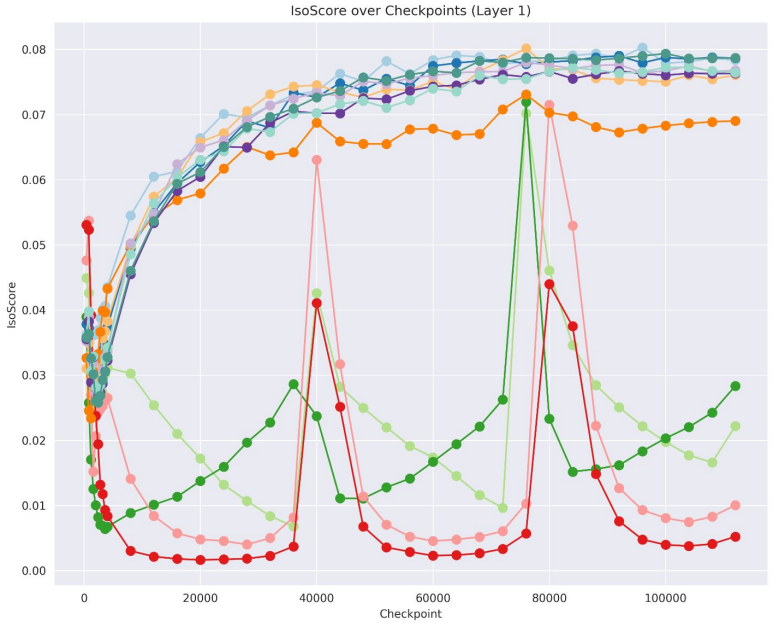
$$\Delta\nu_l = \omega_d(r_l^d - r_{l-1}^d),$$

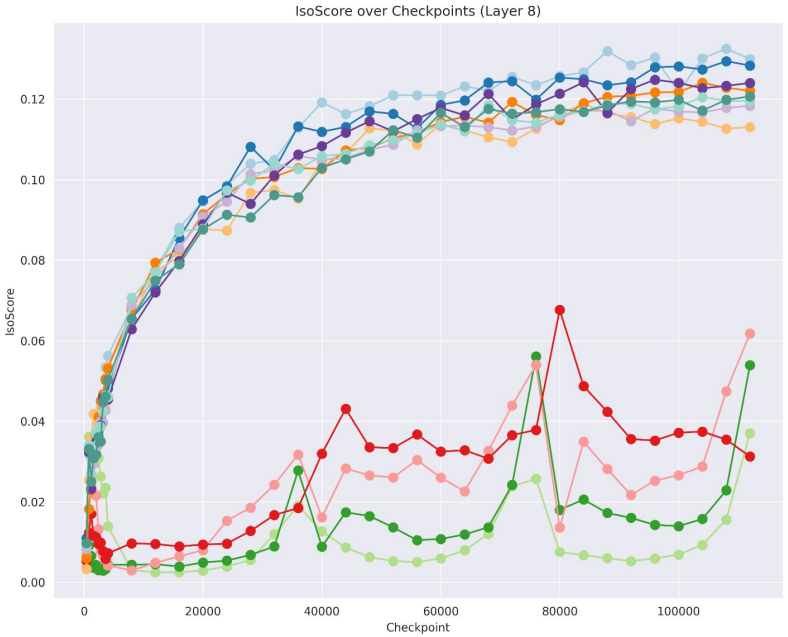
$$P(\Delta\nu_l \in [\nu, \nu + d\nu]) = \rho e^{-\rho\nu} d\nu.$$

$$P(R \in [\overline{R}, \overline{R} + d\overline{R}])$$
$$= \int_0^\infty d\nu_i \int_0^\infty d\nu_j \rho^2 e^{-\rho(\nu_i+\nu_j)} 1_{\left\{\frac{\nu_j}{\nu_i} \in [\overline{R}, \overline{R}+d\overline{R}]\right\}}$$
$$= d\overline{R} \frac{1}{(1+\overline{R})^2},$$

1. Compute the pairwise distances for each point in the dataset $i = 1, \ldots, N$.
2. For each point i find the two shortest distances r_1 and r_2.
3. For each point i compute $\mu_i = \frac{r_2}{r_1}$.
4. Compute the empirical cumulate $F^{emp}(\mu)$ by sorting the values of μ in an ascending order through a permutation σ, then define $F^{emp}(\mu_{\sigma(i)}) \doteq \frac{i}{N}$.
5. Fit the points of the plane given by coordinates $\{(log(\mu_i), -log(1 - F^{emp}(\mu_i)))|i = 1, \ldots, N\}$ with a straight line passing through the origin.

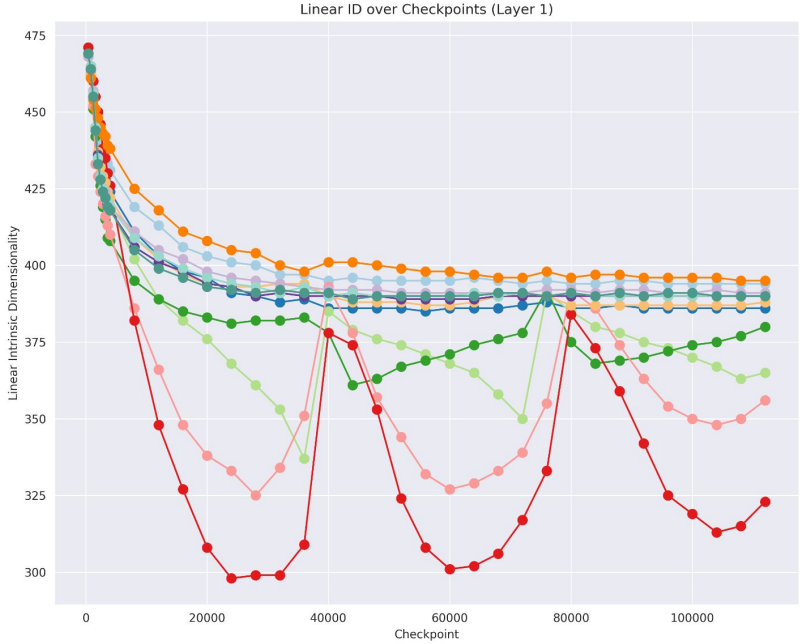
IsoScore

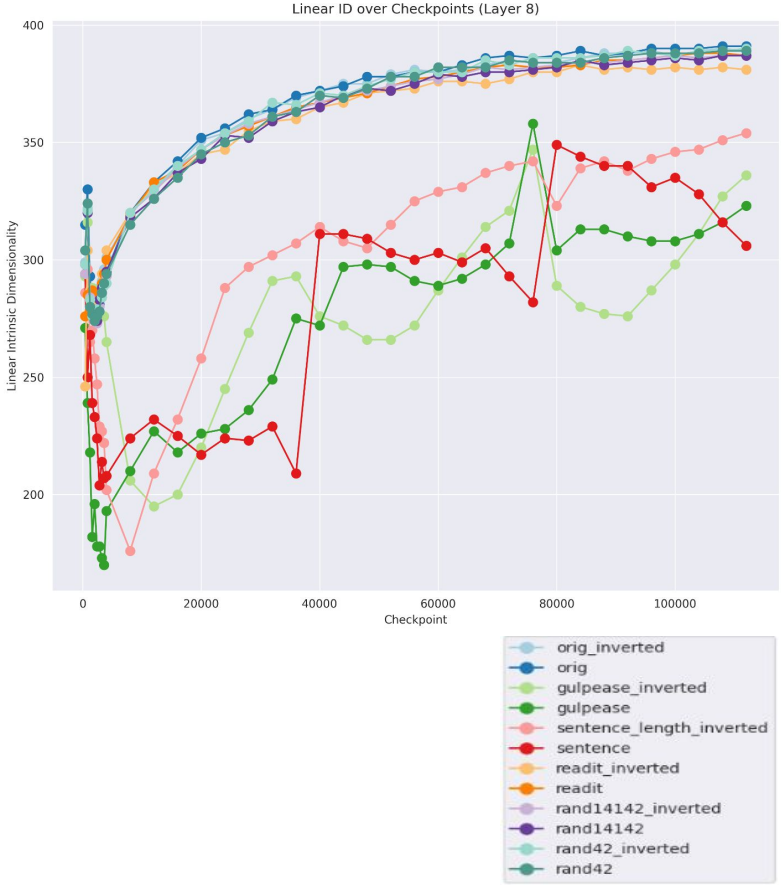

IsoScore over Checkpoints (Layer 1)
IsoScore
Checkpoint
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0
20000
40000
60000
80000
100000


IsoScore over Checkpoints (Layer 8)
IsoScore
Checkpoint
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
20000
40000
60000
80000
100000


orig_inverted
orig
gulpease_inverted
gulpease
sentence_length_inverted
sentence
readit_inverted
readit
rand14142_inverted
rand14142
rand42_inverted
rand42

Linear Intrinsic Dimensionality

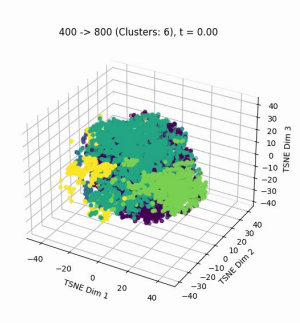




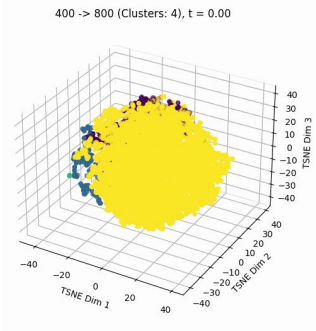
COHERENT CLUSTERS?

- ❏ Feature selection/reduction with PCA → **Cumulative Explained Variance** plot
- ❏ KMeans for Clustering → **Knee plot** for best ***k***
- ❏ **Visualization** with T-SNE

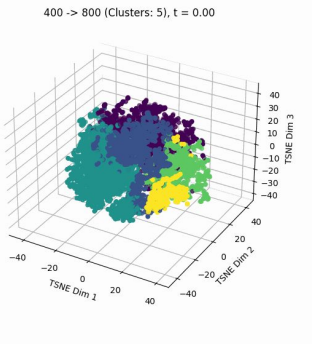
Nonlinear dimensionality reduction technique, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.



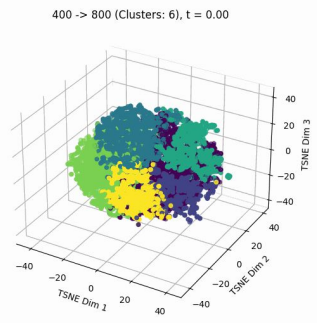
Sentence length



Gulpease

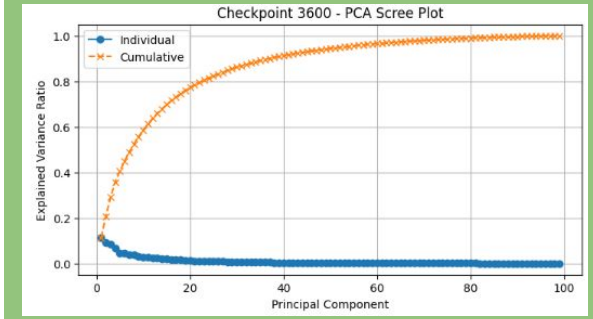


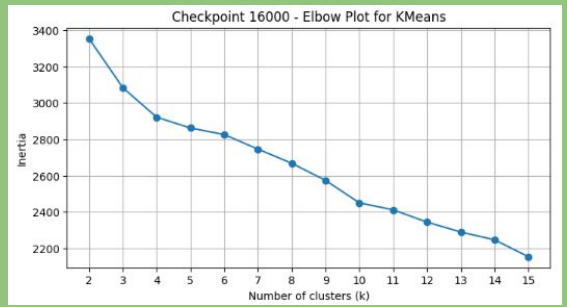
ReadIT



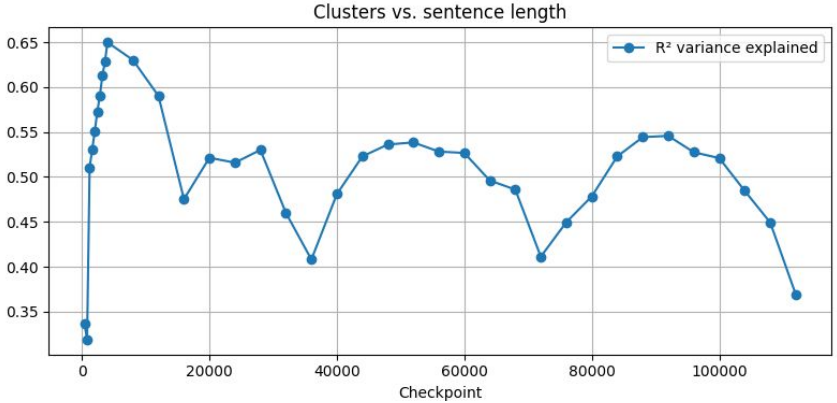
Random

Some example plots.

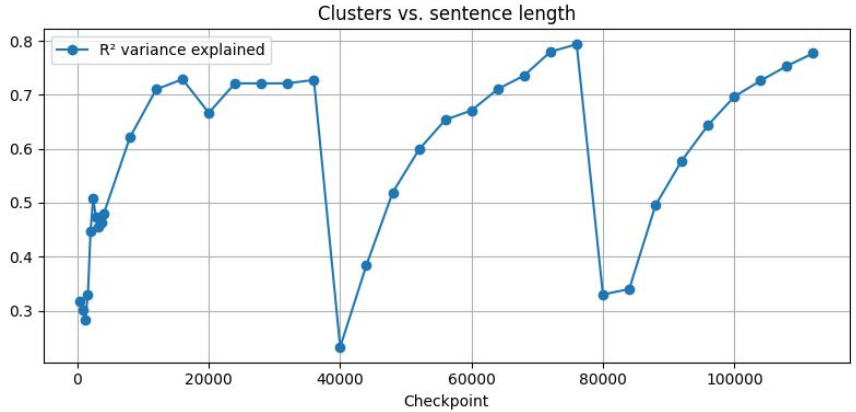




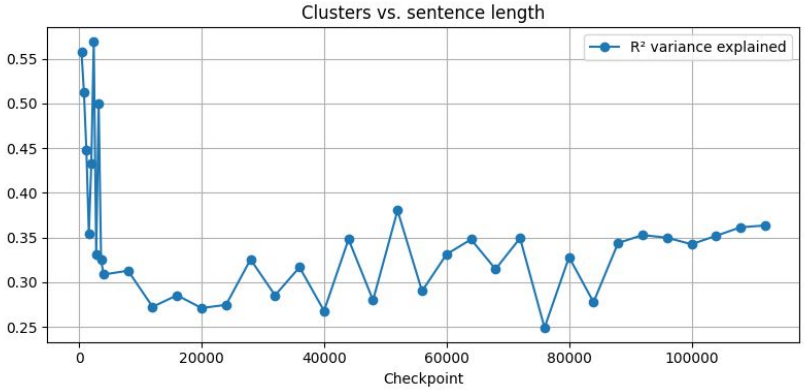
COHERENT CLUSTERS?



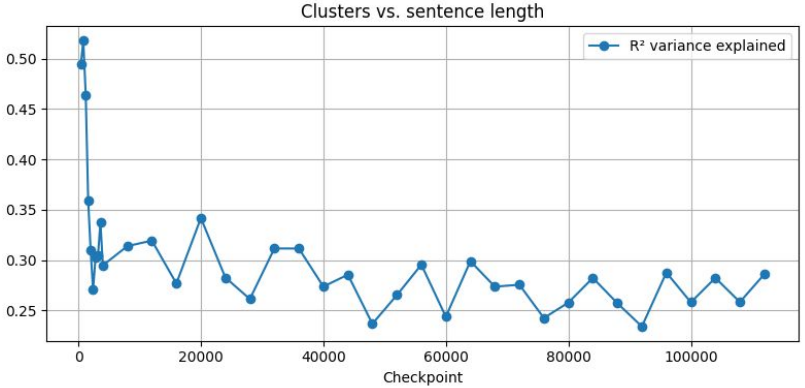
Gulpease



Sentence Length



ReadIT



Random

FUTURE WORK ...

EMPLOY NONLINEAR METRICS AND APPROACHES

PERFORM THE SAME ANALYSIS ON WORDS RATHER THAN SENTENCES

APPLY FUNCTIONAL ANALYSIS AND CONSIDER DEPENDENCE FROM “TIME”

DOES THIS MEAN REPRESENTATIONS CAN BE COMPRESSED ?

REFERENCES

- Curriculum Learning: A Survey - Petru Soviany, Radu Tudor Ionescu, Paolo Rota, Nicu Sebe
- Modeling Easiness for Training Transformers with Curriculum Learning - Leonardo Ranaldi, Giulia Pucci, Fabio Massimo Zanzotto
- Probing Linguistic Knowledge in Italian Neural Language Models across Language Varieties - Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, Giulia Venturi
- How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings - Kawin Ethayarajh
- Mechanics and Geometry of Solids and Surfaces - J. D. Clayton
- IsoScore: Measuring the Uniformity of Embedding Space Utilization - William Rudman, Nate Gillman, Taylor Rayne, Carsten Eickhoff
- Stable Anisotropic Regularization - William Rudman, Carsten Eickhoff