References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online, August 2021. Association for Computational Linguistics.

[2] Mohammed AlQuraishi. ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, 20, 2019.

[3] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[5] T. Blevins, O. Levy, and L. Zettlemoyer. Deep rnns encode soft hierarchical syntax. *Proceedings of the 56th Association for Computational Linguistics (ACL)*, 2018.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[7] John-Marc Chandonia, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K Fox, and Steven E Brenner. SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research*, 50(D1):D553–D559, 12 2021.

[8] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[9] Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):1–13, 2020.

[10] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 11 2020.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] N.S. Detlefsen, S. Hauberg, and W. Boomsma. Learning meaningful representations of protein sequences. *Nature Communications*, 13, 2022.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171–4186, 2019.

[14] Diego Doimo, Aldo Glielmo, Alessio Ansuini, and Alessandro Laio. Hierarchical nucleation in deep neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[16] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[17] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7, 2017.

[18] Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Dadapy: Distance-based analysis of data-manifolds in python. *Patterns*, 3(10):100589, 2022.

[19] John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[20] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021.

[21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[22] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[23] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations*, 18–24 Jul 2021.

[24] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[25] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[26] Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and Sueyeon Chung. Emergence of separable manifolds in deep language representations. In *International Conference on Machine Learning*, pages 6713–6723. PMLR, 2020.

[27] Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Science*, 117(48):30046–30054, December 2020.

[28] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, jun 1993.

[29] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34, 2021.

[30] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*, 2005.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jun 2022.

[34] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. *Proceedings of the 38th International Conference on Machine Learning*, 139:8844–8856, 18–24 Jul 2021.

[35] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[36] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[37] B. Scholkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning series. MIT Press, 2018.

[38] Konstantin Schütze, Michael Heinzinger, Martin Steinegger, and Burkhard Rost. Nearest neighbor search on embeddings rapidly identifies distant protein relations. *Frontiers in Bioinformatics*, 2, 2022.

[39] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022.

[40] Johannes Söding and Michael Remmert. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. *Current opinion in structural biology*, 21 3:404–411, 2011.

[41] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

A Appendix

Table 1: Characteristics of the models employed for extracting representations.

Model	#Blocks	Emb. dim.	#Heads	#Params	Dataset	Reference
ProtBert	30	1024	16	420M	UR100	[16]
ProtT5-XL-U50	24	1024	32	3B	UR50 | BFD	[16]
ESM-1b	33	1280	20	650M	UR50/D	[35]
ESM-1v	33	1280	20	650M	UR90	[29]
ESM-2(8M)	6	320	20	8M	UR50/D	[24]
ESM-2(35M)	12	480	20	35M	UR50/D	[24]
ESM-2(150M)	30	640	20	150M	UR50/D	[24]
ESM-2(650M)	33	1280	20	650M	UR50/D	[24]
ESM-2(3B)	36	2560	40	3B	UR50/D	[24]
ESM-2(15B)	48	5120	40	15B	UR50/D	[24]
iGPT-S(76M)	24	512	8	76M	ImageNet	[8]
iGPT-M(455M)	36	1024	8	455M	ImageNet	[8]
iGPT-L(1.4B)	48	1536	16	1.4B	ImageNet	[8]

A.1 Experimental setup

A.1.1 Hardware

All experiments were performed on a machine with 2 Intel(R) Xeon(R) Gold 6226 with a total of 48 threads, 256GB RAM equipped with 2 Nvidia V100 GPUs with 32GB memory. The GPUs were used to generate embeddings and to compute nearest neighbors.

A.2 Experiments

A.2.1 Two Nearest Neighbors ID estimator

To estimate the intrinsic dimension of hidden representations, we use the Two-Nearest Neighbors-Based (TwoNN) ID estimator [17]. The algorithm is based on a simple analytical result: under the hypothesis of a uniform density of points in $\mathbb{R}^d$, the cumulative probability distribution of the random variable $\mu = \frac{r_2}{r_1}$, where r_1, r_2 are respectively the distance to the first and the second neighbor of a given point, is given by $F(\mu) = 1 - \mu^{-d}$. Therefore, for a given dataset whose points are indexed by $i = 1, \ldots, N$ in $\mathbb{R}^D$ (with $D >> d$ in interesting cases), we compute for each point the ratios μ_i, sort them in ascending order with a permutation σ, and, by defining the empirical cumulative distribution $F^{emp}(\mu_{\sigma(i)}) := \frac{i}{N}$, we can obtain an estimate of d as the slope given by a linear regression (passing through the origin) of the following variables: $(\log(\mu_i), -\log(1 - F^{emp}(\mu_i)))|i = 1, ..., N$. The TwoNN algorithm requires minimal information: the distances to each point's first and second nearest neighbor; therefore, the strong hypothesis of a uniform density used to obtain the main result can be relaxed to a weak assumption of *local* uniformity. We estimate the ID and its reliability through a progressive, random decimation process that allows testing the stability of the result with respect to a change in spatial scale. Since the estimate is approximately scale-invariant, we take the ID estimate as the mean over the values collected during the decimation.

A.2.2 GPU kNN search

The nearest neighbor searches for the calculation of the neighborhood overlap as in [18] were carried out by means of the Python interface of the Facebook AI Similarity Search library [21], version 1.7.2. The library is particularly suited for large datasets embedded in high dimensions since it is based on a reliable approximate and extremely fast similarity search procedure.

A.3 Further results

A.3.1 The ID shape for different pLMs architectures

The latest developments in the application of pre-trained pLMs for the solution of diverse biological tasks have been fuelled by two families of models: Prot-Trans [16] and Evolutionary Scale Modelling (ESM) [35, 34, 29, 24]. During the last years pLMs with different architectures, number of parameters, and embedding sizes have been trained on several datasets obtained starting from the UniProt [10] database. In Fig. 5 we complement our analysis in Section 3.1 including several models whose architectural details and training strategies are described in Table 1. Despite the significant differences of the pLMs considered in the analysis, the consistency of the three-phased behavior of the ID curve is remarkable: an initial peak is followed by a plateau where the ID assumes low values, and the ID grows again to values close to the one measured after the positional embedding.

A.3.2 Nearest neighbor search in plateau layers improves identification of protein relations

It was recently shown in [38] that first nearest neighbor searches for remote homologous protein domains based on the last hidden layer representations of the ProtT5-XL-U50 pLM outperform state-of-the-art methods based on sequence similarity. Adapting the approach in 3.2, we mimic the experiment performed in Section 2 of [38] by 1) considering protein domains in SCOPe belonging to a super-family with at least 2 sequences, 2) setting the number of neighbors to $k = 1$. Considering representations in the plateau layer improves the accuracy of the 1-kNN homology search. In particular, in Fig. 5 [Right] we observe an improvement of $\sim 6\%$ performing the search on a plateau layer instead of the last layer before the output. It is important to notice that the performance gain of $\sim 6\%$ is obtained without any further training.

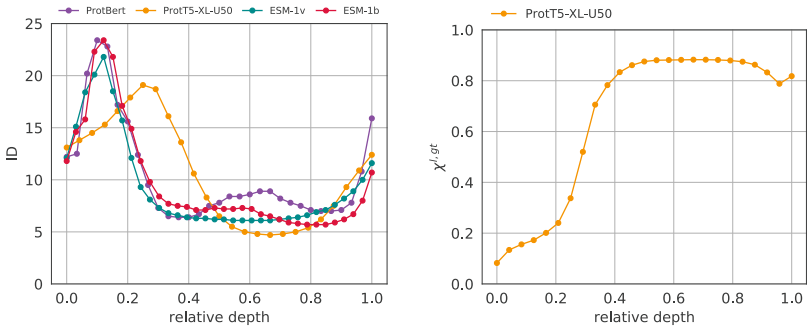

Figure 5: Further experiments. [Left] The ID curves for different pLMs trained on different datasets consistently show the three-phased behavior consisting of a peak, a plateau, and a final ascent. [Right] First nearest neighbor SCOPe super-family retrieval accuracy of Prot-T5-XL-U50 is higher in plateau layers.

A.3.3 NO curves are robust w.r.t. the number of neighbors

It was shown in [14], Fig. A.1 (a), that the trend of the neighborhood overlap (NO) curve is robust with respect to the choice of the hyperparameter k. We verify this also for pLMs and iGPT analyzing the NO curves of ESM-2(650M) and iGPT-L for different choices of number of neighbors k. The results of this analysis, reported in Fig. 6, show that the qualitative behavior of the NO curves is independent of k. As expected, the alignment of the neighbor composition with ground truth classes $\chi_k^{l,gt}$ decreases when k becomes larger. When considering the ESM-2(650M) model, due to the possibility of certain superfamilies having fewer than 50 elements, it is expected to observe significantly lower values of $\chi_k^{l,l+1}$ and $\chi_k^{l,gt}$ when $k = 50$.

A.3.4 Self-supervised pre-training is crucial for emergence of three-phased behavior

Different models pre-trained on different datasets present a similar ID shape characterized by a three-phased structure, and the global picture is shared across models and datasets. In particular,

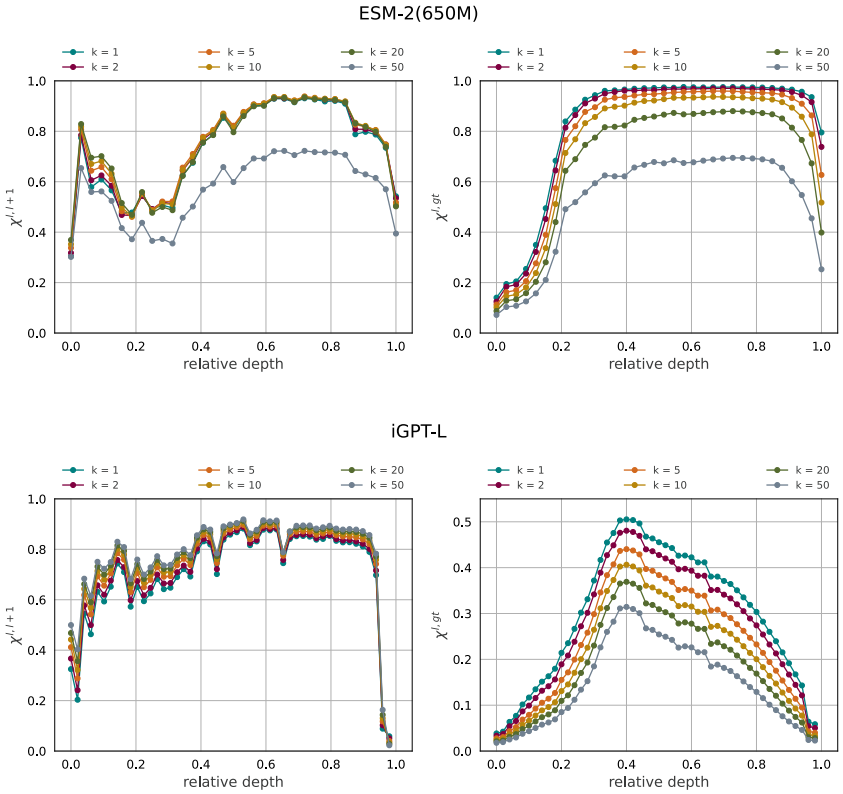

Figure 6: Neighborhood Overlap curves for varying hyperparameter k. [Top Left] and [Bottom Left]. The NO curves describing the overlap $\chi_k^{l,l+1}$ between successive layers are essentially unchanged for $k < 50$. [Top Right] and [Bottom Right] The NO curve $\chi_k^{l,gt}$ showing the alignment of neighborhood composition and classification have the same qualitative behavior.

Figure 5(a) shows that the ID shape of pLMs is affected only by slight modifications when the pre-trained dataset passes from UniRef50 to Uniref90 (ESM-1v) or to a combination of UniRef and BFD (ProtT5-XL-U50). In order to inspect further the role of pre-training on the behavior of the ID curve we perform an experiment whose results are reported in Figure 7. We consider a Vision Transformer (ViT) model [15], which has a very similar architecture to iGPT, with weights obtained through the weakly-supervised pretraining protocol by [39] followed by fine-tuning on ImageNet-1k. One can observe that in this setting the ID curve changes towards matching the hunchback shape that has been observed by [3] in the context of convolutional neural networks trained on Imagenet-1k classification, even if on a different scale of ID values. This highlights the crucial role of self-supervised pre-training for the emergence of the three-phased behavior.

A.3.5 ID curve of transformers for Natural Language Processing

The complexity of language data is extraordinarily high, requiring extremely heterogeneous tasks and probes to fully capture it. In addition, there is another substantial discriminant that separates pLMs (and vision transformers) from transformers applied in the NLP domain: for pLM models, we already reached an overparameterization regime on the UniRef dataset, as observed by [24], while this is far from true in the context of language, where large LMs that are exponentially increasing in size are still far from saturation. Furthermore, there is scarce consensus in the literature on which is the most appropriate method to construct sentence-level representations (CLS token, token concatenation, averages across tokens, etc.). For all these reasons, the experiments we report in this Section should be intended as an initial experiment on the geometry of representations of language transformers that will require a more in-depth analysis in future work.

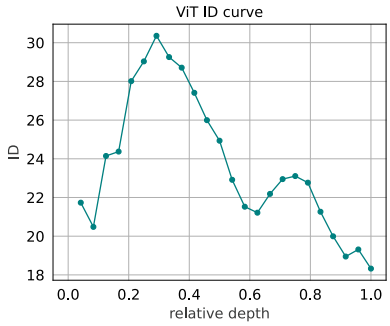

Figure 7: ID of representations of the Vision Transformer model plotted against relative depth. The ID profile of ViT in the fist half of the self-attention blocks is characterized by a prominent peak. In the second part, it presents a much less pronounced peak wrt iGPT followed by a progressive descent of the ID. The fine-tuning procedure pushes the ViT ID curve towards the hunchback shape observed in [3].

We analyze representations extracted from GPT-2 XL [32] with 1.5B parameters trained by next-token prediction on WebText, a specifically curated dataset selection of internet scraping from 2017. In particular, we report in Fig. 8 the ID curves obtained performing inference on two datasets: the English Penn Treebank [28] containing 38.219 sequences collected and annotated for evaluations of syntactic and semantic sequence-labelling tasks (Fig. 8 [Left]); the Stanford Sentiment Treebank v.2 (SST-2) [30] consisting of 43.296 sequences from movie reviews constructed as a benchmark for a complete analysis of the compositional effects of sentiment in language (Fig. 8, [Right]).

In both cases, the ID values at the last layers are very close to the initial ones. The most prominent feature of the ID profiles is their symmetry, which is consistent with what is observed for iGPT. In particular, the GPT-2 ID curve presents a single ID peak approximately in the middle of the network, with two small minima immediately after the input and before the output. It is important to notice that the ID spans a totally different range when considering different language datasets: the ID varies around the value 4 for the SST-2 dataset, and around 31 for the Penn Treebank dataset. This difference is particularly remarkable given the fact that we are considering representations of the same model; once again, this dissimilarity is a trace of the complexity and heterogeneity of language datasets.

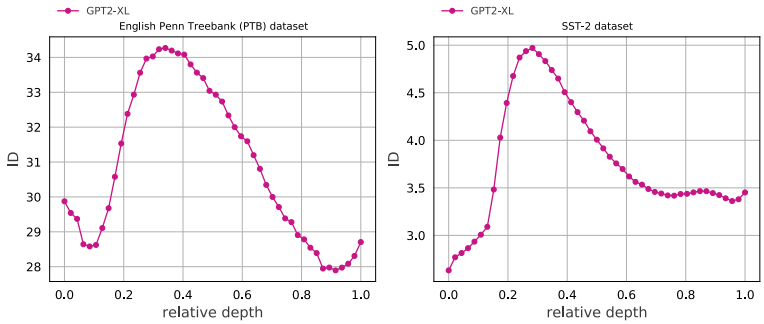

Figure 8: ID of GPT-2 XL representations plotted against relative depth for the Penn Treebank dataset [Left], and Stanford Sentiment Treebank [Right]. The ID curves have a single ID peak approximately in the middle of the network. The ID spans remerkably different values for the two datasets.