Fine-Tuning Transformers For Genomic Tasks

Vlastimil Martinek [*][1] **David Cechak** [*][1] **Katarina Gresova** [1] **Panagiotis Alexiou** [1] **Petr Simecek** [1]

Abstract

Transformers are a type of neural network architecture that has been successfully used to achieve state-of-the-art performance in numerous natural language processing tasks. However, what about DNA, the language life written in the four-letter alphabet? In this paper, we review the current state of Transformers usage in genomics and molecular biology in general, introduce a collection of benchmark datasets for the classification of genomic sequences, and compare the performance of several model architectures on those benchmarks, including a BERT-like model for DNA sequences DNABERT as implemented in HuggingFace (armheb/DNA_bert_6 model). In particular, we explore the effect of pre-training on a large DNA corpus vs training from scratch (with randomized weights). The results presented here can be used for identification of functional elements in human and other genomes.

1. Introduction

In the past five years, Deep Learning methods for Natural Language Processing (NLP) came through a revolution that has been possible thanks to two key novel innovations: language models and transfer learning. With this approach, the model is first trained in an unsupervised fashion with unlabelled data and then fine-tuned to a specific downstream task with labelled data. (Howard & Ruder, 2018) trained the ULMFit model to predict the following word in English Wikipedia corpus and then fine-tuned it to six text classification tasks (outperforming the state-of-the-art methods at a time). While ULMFit architecture was still based on Long Short Term Memory networks (LSTMs), the novel model architecture based on Encoder / Decoder structure and self-attention was introduced at around the same time – Transformers (Vaswani et al., 2017) – and have dominated the NLP field since then. It was shown that some neurons and attention heads have a direct connection to text features like sentiment (Radford et al., 2017) or direct objects of verbs (Clark et al., 2019). While the original transformer models like BERT (Devlin et al., 2018) have just lower hundred millions of parameter, much larger language models have been recently introduced like GPT-3 with 175B parameters (Brown et al., 2020), Gopher with 280B parameters (Rae et al., 2021) and GLaM with more than 1.2T parameters (Du et al., 2021). Other recent changes include the unification of different tasks (Raffel et al., 2019) expansion of transformer architecture beyond traditional sequential models, e.g. Vision transformers (Dosovitskiy et al., 2020), (Dai et al., 2021) and/or 3D Point Cloud transformers (Zhao et al., 2020).

But what about DNA, the language life written in the four-letter alphabet? For the simplicity reasons, we restrict ourselves to the human genome in this paper. It consists of more than 3 billion base pairs organized into 22 paired chromosomes (autosomes) and the 23rd pair of sex chromosomes (XX for females, XY for males). The known successful deep learning applications for convolutional neural networks (CNNs) and recurrent LSTMs include identification/classification of genes from their sequence (Georgakilas et al., 2019) and identification of functional elements regulating gene expression, namely gene promoters (Umarov & Solovyev, 2017), enhancers (Liu et al., 2016), enhancer-promoter interactions (Zeng et al., 2018) and transcription factor binding sites (Shen et al., 2018).

Unfortunately, unlike in NLP, there are no widely recognized DNA benchmarks. To overcome this problem, we have started to work on a collection of genomic datasets and propose the first five of them in the Method section. The second issue is more serious, DNA is written rather in several languages than one original language. The $\sim 20,000$ protein coding gene sequences represent $\sim 1\%$ of the human genome. Approximately 50% of the human genome is made up of repetitive sequences, mostly transposons, but also microsatellites and minisatellites and even duplications of large segments (Haubold & Wiehe, 2006).

There are also not so many language models trained for DNA. (Hoarfrost et al., 2020) trained ULMFit-like model LookingGlass on microbial genomes. Karl Heyer published

[*]Equal contribution [1]Panagiotis Alexiou Research Group, Centre for Molecular Medicine, Central European Institute of Technology, Masaryk University, Brno, Czechia. Correspondence to: Petr Simecek <petr.simecek@ceitec.muni.cz>.

his experiments as GitHub repo GenomicULMFit (`https://github.com/kheyer/Genomic-ULMFiT`). Regarding transformer architecture, to the best of our knowledge, the only known language model is DNABert (Ji et al., 2020) trained on the human genome that we will utilise for our purposes. While it is not explicitly mentioned inDNABert paper, the model can be found in HuggingFace model repository (armheb/DNA_bert_6 model).

2. Methods

2.1. Datasets

Due to the lack of established genomic benchmarks, we have started to put together our own. The collection is based on a combination of existing datasets obtained from published papers and novel datasets constructed from public databases. The data are distributed as a Python package available at `https://github.com/ML-Bioinfo-CEITEC/genomic_benchmarks`, the minimalist version (compressed list of genomic coordinates) is stored on GitHub itself and full datasets (full DNA sequences) are cached on Google Drive.

For this paper, we have used the five datasets that have already been curated and will be part of the benchmark in the future. For testing we use $\sim 30\%$ of data points. All five datasets contains exactly two classes and are either balanced or (in case of human promoters) close to it. The summary table with number of sequences and their lengths are in Table 1.

2.1.1. Human Non-TATA Promoters

A promoter is a sequence of DNA that binds a protein initiating the gene transcription. Effectively, it turns gene expression on and off. It is usually located close (from -200 to 50bp) to the transcription splice site (TSS). This dataset has been adapted from the paper (Umarov & Solovyev, 2017).

2.1.2. Human Enhancers Cohn

An enhancer is a sequence of DNA that can bound specific proteins and therefore increase a change of transcription of a particular gene. Unlike promoters, enhancers do not need to be in a close proximity to TSS (might be several Mb away). This dataset has been adapted from (Cohn et al., 2018) paper.

2.1.3. Human Enhancers Ensembl

For this dataset of human enhancers, we have queried Ensembl database (Howe et al., 2021), release 100. The data are originally coming from VISTA Enhancer Browser project, (Visel et al., 2007). The Unlike the other datasets, this one has variable length of the sequences.

2.1.4. Coding Vs Intergenomic Regions

This dataset has been originally used for teaching purposes at ECCB2020 workshop. It consists of randomly generated 50,000 sequences (200bp long) from intergenomic regions and randomly generated 50,000 sequences from human transcripts.

2.1.5. Human Or Worm?

Randomly chosen DNA sequences (200bp long) either from the human genome or from the genome of C. elegans (worm).

2.2. Models & Training

We have trained and evaluated three models for each dataset: First, we fine-tuned DNABert model pre-trained on human DNA (Ji et al., 2020). Second, to assess the effect of pre-training, we trained the model initialized with random weights (no pre-training). Lastly, as a baseline we have then used CNN architecture previously successfully used to similar problems (Klimentova et al., 2020).

We have repeated each training five time to evaluate the variability of the results. As a loss function, we have used binary cross entropy. We have used early stopping and fallback to the model that achieved the lowest loss on the validation set.

The BERT models were trained with batch size of 48 and weight decay of 0.1. The learning rate was linearly increased to 0.0002 during the warmup period. AdamW was used as an optimizer. The CNN models were trained with the Adam optimizer, using learning rate of 0.001, no weight decay, and batch size of 32.

Code Repository

All code to derive results in this paper is available in a GitHub repository:

`https://github.com/ML-Bioinfo-CEITEC/genomic_benchmarks`

3. Results

To evaluate the performance of the model on a testing set, we will use the F1 metric:

$$F_1 = 2 \cdot \frac{\texttt{precision recall}}{\texttt{precision + recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)},$$

where TP is a number of true positives, FP a number of true positives and TN is a number of false negatives.

The running time has been 5-30 minutes for one run of

Table 1. Number of sequences and sequence length per dataset.

NAME	# OF SEQUENCES	MEDIAN LENGTH	STD. OF LENGTH
HUMAN_NONTATA_PROMOTERS	36131	251	0.0
HUMAN_OR_WORM	100000	200	0.0
HUMAN_ENHANCERS_ENSEMBL	154842	269	122.6
CODING_VS_INTERGENOMIC_SEQS	100000	200	0.0
HUMAN_ENHANCERS_COHN	27791	500	0.0

CNN model and 2-6 hours for transformer models (Google Clound Platform, n1-highmem-8 virtual machine, NVIDIA Tesla T4 GPU).

The performance of the models summarized in F1 metric on a testing set is reported in the Table 2. As you can see DNABert is superior in all five our benchmark datasets and the fine-tuned DNABert outperformed the model with the randomized weights in four out of five cases.

4. Discussion

In this paper, we have experimented with transformers applied to classification of DNA sequences. We have shown that the model pre-trained on human genome achieves better accuracy than the same model with randomized weights and a convolutional neural network model. While ML researchers in the genomic field currently uses rather simple architectures like LSTMs and CNNs, the HuggingFace implementation of DNABert should encourage wider adoption of transformers.

DNA sequences present a unique challenge for machine learning because of their length and complexity. Transformers provide a more effective way to model these sequences than traditional neural networks. However, with only one transformer model trained over DNA available, many questions remain open for further investigation. Would the bigger models achieve better performance as for natural language and also for protein sequences (Rives et al., 2021), (Elnaggar et al., 2020), (Xiao et al., 2021)? If one universal DNA language model sufficient or would it be better to train a separate language model for each model organism (human, mouse, zebrafish, . . .).

And finally, taking into account the heterogeneous nature of human genome, would it be better to train on corpus that would not be the whole genome but rather a handcrafted specific subsample, e.g. for promoters taking only segments close to transcription splice site? This should be investigated in future work.

Acknowledgement

P. Simecek's work on this paper has been supported by H2020 MSCA IF grant LangOfDNA (Grant ID 896172).

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are Few-Shot learners. May 2020.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. What does BERT look at? an analysis of BERT's attention. June 2019.

Cohn, D., Zuk, O., and Kaplan, T. Enhancer identification using transfer and adversarial deep learning of dna sequences. *BioRxiv*, pp. 264200, 2018.

Dai, Z., Liu, H., Le, Q. V., and Tan, M. CoAtNet: Marrying convolution and attention for all data sizes. June 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. October 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. October 2020.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., Zoph, B., Fedus, L., Bosma, M., Zhou, Z., Wang, T., Wang, Y. E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q. V., Wu, Y., Chen, Z., and Cui, C. GLaM: Efficient scaling of language models with Mixture-of-Experts. December 2021.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

Table 2. F1-score on testing sets, best model in bold font.

Evaluation: Data set	CNN	DNABERT (randomized weights)	DNABERT (pretrained)
HUMAN_NONTATA_PROMOTERS	83.9 ± 1.5	88.0 ± 0.7	**91.9 ± 0.9**
HUMAN_OR_WORM	83.3 ± 0.6	80.0 ± 0.6	**96.8 ± 0.2**
HUMAN_ENHANCERS_ENSEMBL	61.9 ± 1.4	83.8 ± 0.7	**86.9 ± 0.3**
CODING_VS_INTERGENOMIC_SEQS	74.8 ± 0.5	78.3 ± 0.6	**92.8 ± 0.6**
HUMAN_ENHANCERS_COHN	67.1 ± 0.6	**76.5 ± 0.5**	74.1 ± 0.4

Georgakilas, G. K., Grioni, A., Liakos, K. G., Malanikova, E., Plessas, F. C., and Alexiou, P. MuStARD: a deep learning method for intra- and inter- species scanning identification of small RNA molecules. March 2019.

Haubold, B. and Wiehe, T. How repetitive are genomes? *BMC Bioinformatics*, 7:541, December 2006.

Hoarfrost, A., Aptekmann, A., Farfañuk, G., and Bromberg, Y. Shedding light on microbial dark matter with a universal language of life. December 2020.

Howard, J. and Ruder, S. Universal language model finetuning for text classification. January 2018.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., et al. Ensembl 2021. *Nucleic acids research*, 49(D1):D884–D891, 2021.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. DNABERT: pretrained bidirectional encoder representations from transformers model for DNA-language in genome. September 2020.

Klimentova, E., Polacek, J., Simecek, P., and Alexiou, P. Penguinn: Precise exploration of nuclear g-quadruplexes using interpretable neural networks. *Frontiers in Genetics*, 11:1287, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, F., Li, H., Ren, C., Bo, X., and Shu, W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Sci. Rep.*, 6:28517, June 2016.

Radford, A., Jozefowicz, R., and Sutskever, I. Learning to generate reviews and discovering sentiment. April 2017.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., van den Driessche, G., Hendricks, L. A., Rauh, M., Huang, P.-S., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X. L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J.-B., Tsimpoukelli, M., Grigorev, N., Fritz, D., Sottiaux, T., Pajarskas, M., Pohlen, T., Gong, Z., Toyama, D., de Masson d'Autume, C., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., de Las Casas, D., Guy, A., Jones, C., Bradbury, J., Johnson, M., Hechtman, B., Weidinger, L., Gabriel, I., Isaac, W., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K., Stanway, J., Bennett, L., Hassabis, D., Kavukcuoglu, K., and Irving, G. Scaling language models: Methods, analysis & insights from training gopher. December 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified Text-to-Text transformer. October 2019.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Shen, Z., Bao, W., and Huang, D.-S. Recurrent neural network for predicting transcription factor binding sites. *Sci. Rep.*, 8(1):15270, October 2018.

Umarov, R. K. and Solovyev, V. V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One*, 12(2):e0171410, February 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. June 2017.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. Vista enhancer browser—a database of tissue-specific human enhancers. *Nucleic acids research*, 35(suppl_1): D88–D92, 2007.

Xiao, Y., Qiu, J., Li, Z., Hsieh, C.-Y., and Tang, J. Modeling protein using large-scale pretrain language model. *arXiv preprint arXiv:2108.07435*, 2021.

Zeng, W., Wu, M., and Jiang, R. Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics*, 19(Suppl 2):84, May 2018.

Zhao, H., Jiang, L., Jia, J., Torr, P., and Koltun, V. Point transformer. December 2020.