Evolution Is All You Need: Phylogenetic Augmentation for Contrastive Learning

Amy X. Lu[1,2], Alex X. Lu[1], Alan Moses[1,2]

[1]Department of Computer Science, University of Toronto [2]Department of Cell and Systems Biology, University of Toronto

Introduction

- Current methods for learning self-supervised sequence embeddings in biology rely on large language models for NLP, leaving open the question of how best to design self-supervised methods which align with biological principles.
- In this perspectives piece, we illustrate how maximizing information across phylogenetic "noisy channels" is more biologically-motivated than current language models for protein representations, and theoretically desirable.

Background Works: Contrastive Learning for Mutual Information Maximization

- **InfoMax optimization principle**: Find a mapping g such that the Shannon mutual information between the input and output is maximized [6]:

$$\max_{g \in \mathcal{G}} I(X; g(X)) \tag{1}$$

- **InfoMax for Representation Learning:** Recently, works capture this intuition to train deep encoders for g, and yield empirically desirable representations for images, text, and audio, often following this general form [9]:

$$\max_{g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2} I'(g_1(v_1); g_2(v_2)) \tag{2}$$

  - Given input x, and transformations t_1 and t_2, define $v_1 = t_1(x)$ and $v_2 = t_2(x)$ as two different "views" of x, and encoder(s) and latent representations $z_1 = g_1(v_1)$ and $z_2 = g_2(v_1)$, respectively.
  - The goal is to find encoder mappings which maximize the mutual information between the outputs.
  - Maximizing Equation 2 is equivalent to maximizing a lower bound on true InfoMax objective [7].
- **InfoNCE Loss for Mutual Information Estimation:** The InfoNCE estimator [7] estimates $I(v_1, v_2)$ by minimizing the loss:

$$\mathcal{L}_{NCE} := \mathbb{E}_{v_1, v_2}\left[\log \frac{\exp(f(v_1^+, v_2))}{\exp(f(v_1^+, v_2)) + \sum_{j=1}^{N-1} \exp(f(v_1^-{}_j, v_2))}\right], \tag{3}$$

  - $v_1^+, v_2 \sim p(v_1, v_2)$ denote a "similar" pair drawn from the empirical joint distribution of the two views, and $v_1^-, v_2 \sim p(v_1)p(v_2)$ is a "dissimilar" pair of views.
  - Equation 3 is a cross-entropy which identifies the similar pair from the dissimilar pair of views; losses which fall into this general form are termed "contrastive learning" [1].
  - Augmentations are often-used strategy to generate views [2].

References

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[3] Lila L Gatlin et al. *Information theory and the living system*. Columbia University Press, 1972.

[4] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[5] Motoo Kimura. Natural selection as the process of accumulating genetic information in adaptive evolution. *Genetics Research*, 2(1):127–140, 1961.

[6] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

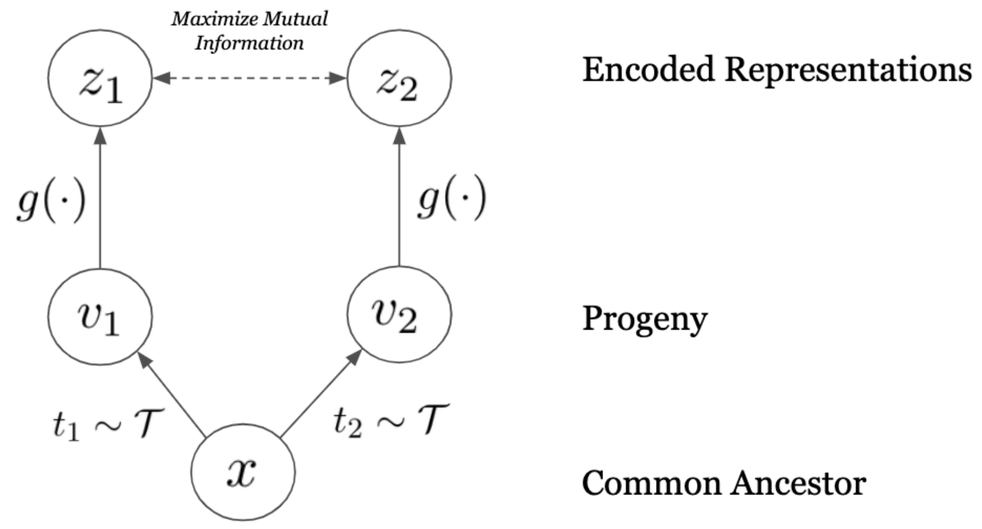
[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[8] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

[9] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Evolution as Sequence Augmentation

Here, we take SimCLR [2] as an illustrative example of a contrastive learning framework which can be directly adapted to use evolutionary as an augmentation strategy to create views.



- **Original SimCLR [2]:** x is an input image, and two image augmentation methods, t and t', are sampled from a set of image augmentation methods $\mathcal{T}$, to produce image augmentation v_1 and v_2, which are then passed into a trainable encoder $g(\cdot)$.
- **SimCLR augmentations, re-cast as a phylogenetic tree:** x can be viewed as a (hypothetical) common ancestor, while $\mathcal{T}$ is a set of sequence families, and t, t' are two families sampled from this database. One can use real or simulated sequences to create $\mathcal{T}$.

Why Evolution as Biological Sequence Augmentation for Contrastive Learning?

- **Invariant Representations Across Evolutionary "Noisy-Channels" Mirrors Comparative Genomics**
  - By using phylogenetic relationships to create views, the contrastive objective encourages agreement between important features across homologous sequences.
  - This directly captures the central philosophy of sequence conservation in comparative genomics.
- **Molecular evolution and the genotype-to-phenotype relationship has a clear analogy to information transmission**
  - The analogy between molecular evolution and noisy-channel coding is well-rooted in prior work [3]: DNA dictates information transmission across generations, which must be transferred through a noisy "mutation and drift channel".
  - Since the genotype-to-phenotype manifestation is information transfer, and genomic information is transferred by heredity, we may view functional phenotypes as "decoded" information that was transmitted from a common ancestor via molecular evolution [5].
  - This is thus a good proxy for maximizing structure and function, the central desiderata for pretrained biological sequence embeddings.
- **Evolutionary Augmentation Theoretically Accords with the InfoMin Principle [8] for Choosing Good Views**
  - **"InfoMin" principle for selecting optimal views:** Good views should have minimize their shared MI $I(v_1, v_2)$ while maximizing task-relevant information for downstream uses between input and embedding.
  - Sampling evolutionary trajectories $t_1, t_2 \sim \mathcal{T}$ to create $v_1 = t_1(x)$ and $v_2 = t_2(x)$ provide a simple way to reduce $I(v_1, v_2)$ by selecting paired views with a greater phylogenetic distance between them.
  - Conservation serves a semantic proxy for downstream labels, and thus implicitly performs *supervised* contrastive learning [4] while circumventing expensive experimental label gathering.

Conclusion

Existing methods for protein embeddings rely heavily on the analogy between protein sequences and natural language, leaving open the space for more elegant methods which better accord with traditionally successful philosophies in bioinformatics. We illustrate one such possibility, by viewing evolution as a means to generate views in recent advances in contrastive pretraining.