EVOLUTIONARY PERSPECTIVE ON MODEL FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Be it in natural language generation or in the image generation, massive performances gains have been achieved in the last years. While a substantial part of these advances can be attributed to improvement in machine learning architectures, an important role has also been played by the ever-increasing parameter number of machine learning models, which made from-scratch retraining of the models prohibitively expensive for a large number of users. In response to that, Transfer Learning (TL) - starting with an already good model and further training it on the data relevant to a new, related problem, gained in popularity. TL is formally similar to the natural evolution of genetic codes in response to shifting environment. Our core contribution, presented in this paper, is to define a class of evolutionary algorithms - Gillespie-Orr EA (GO-EA) and prove that they are equivalent in the limit to stochastic gradient descent (SGD). Based on this equivalence we present a number of tricks used y naturally evolving organisms to accelerate their adaptation, applicable to TL, as well as a set of hypotheses as to properties of artificial neural networks trained with SGD and GO-EA, resulting from such equivalence.

1 INTRODUCTION

Evolution-inspired algorithms are all but new in machine learning. Introduced in 1966, Simulated Evolution took the core components of the evolutionary processes as understood at the time - mutation and selection - and attempted simulate them in order to generate an artificial intelligence (Fogel et al., 1966). Appearing a mere 15 years after Robbins and Monro introduced the Stochastic Gradient Descent (SGD) (Robbins & Monro, 1951), Evolutionary Algorithms (EA) captured the imagination of the AI and ML communities and were refined to more closely follow the understanding of natural evolution, such as chromosomes and crossover (Schwefel, 1981), resulting in perhaps the most well-known evolutionary optimization and search algorithm of the family - the Genetic Algorithm (Goldberg, 1989).

However, this fascination slowly came to an end after in the late 70s and early 80, as several groups discovered independently that in most cases, SGD on the artificial neural networks (AANs) could achieve same, if not better, results as EA but with significantly lower computational expenses (Werbos, 1974; Parker, 1985; LeCun, 1985; Rumelhart et al., 1986). Shortly after, theoretical investigations in ANNs suggested that even simple architectures, such as multilayer feedforwards perceptrons, could work as universal approximators and be trained to approximate any measurable function, provided they had a sufficient number of layers, nodes per layer and non-linear activation functions (Hornik et al., 1989). A rapid flurry of innovations showing that ANNs armed with SGD could be scaled up to perform better and could efficiently deal with a number of problems considered hard until then (LeCun et al., 2015), switched the attention away from EA, until the superhuman performance of ImageNet on image classification tasks (Krizhevsky et al., 2012) made the ANN/SGD approach ubiquitous for computer vision, whereas the Transformer and its derivatives (Vaswani et al., 2017; Brown et al., 2020; Fedus et al., 2021) made ANN and SGD with inertia ubiquitous in Natural Language Processing.

However, despite its impressive success, the SGD has a flaw. By it's nature, it requires the manifold on which learning occurs to be differentiable. A limitation to which the Evolutionary Algorithms are not subject to, and one that is frequently encountered in conditions where multiple agents need to interact within a real or simulated environment (as for instance described by Majumdar et al.

(2020)). This advantage, combined with the advent of massively parallel computing that made the simultaneous evaluation of entire populations of models viable and led to a steady progress in EA application to ANNs, with the introduction and refinement of methods such as ESP (Gomez & Miikkulainen, 1997), CMA-ES (Hansen & Ostermeier, 2001), CoSynNE (Gomez et al., 2008), NES (Wierstra et al., 2008), reviewed in (Hansen et al., 2015). In addition to that, neuroevolution has been used to optimize the topology and hyperparametrs of ANNs before training them with SGD, as introduced by Floreano et al. (2008), or in combination with deep reinforcement learning (RL) (Mnih et al., 2015) - another approach to optimization in non-differentiable manifolds, as presented for instance by Fernando et al. (2017). Overall, the intersection of approaches trying to combine EA with ANN is a rapidly expanding area, reviewed more in depth by Galván & Mooney (2020).

However, compared to SGD, our understanding of how well EA are able to perform optimization and explore the parameter space of the machine learning model based on the data is limited. Specifically, it is unclear how prone they are to being trapped in local minima (Nguyen et al., 2021; Nguyen, 2019; Jacot et al., 2018), finding flat minima (Hochreiter & Schmidhuber, 1994; 1997; Goodfellow & Vinyals, 2015; Li et al., 2018), generalizing (Dinh et al., 2017; Zhang et al., 2020), memorizing (Arpit et al., 2017; Yun et al., 2019; Zhang et al., 2017; 2021) or interacting with ANNs architectures (Balduzzi et al., 2017; Li et al., 2018).

While the question in itself would likely be harder to approach than in the case of SGD, just because of the sheer diversity of EAs, we approach the topic specifically in the context of model fine-tuning, or transfer learning (Yosinski et al., 2014; Oquab et al., 2014; Bengio, 2012; Bengio et al., 2011; Caruana, 1995). The fine-tuning and transfer learning have grown in the importance over the last five years, given the exponential growth in the ANN models size, (Sanh et al., 2019; Brock et al., 2019; Fedus et al., 2021), going hand in hand with the training dataset sizes and computing power required to train them Brown et al. (2020). In this context, the from-scratch model training is prohibitively expensive for the vast majority of users, leading to the proliferation of model fine-tunes or transfers to closely related domains. As of the time of submission of this article, HuggingFace NLP model repository (HuggingFace, 2020) counts over 1300 model fine-tunes, of which around 700 of BERT model alone (Devlin et al., 2019).

Within the context of fine-tuning models or Tranfer Learning (Bozinovski & Fulgosi, 1976; Pratt, 1992; Pan & Yang, 2010), we show that thanks to very general results on the limit distributions derived by Gnedenko et al. (1968) and more generally Fréchet (1927); Fisher & Tippett (1928); Mises (1936); Gnedenko (1943), we can establish a direct equivalence between the SGD and the Fisher Geometric Model of Evolution (FGM) (Fisher, 1930) in the Orr-Gillespie formalization - a fairly large family of EA commonly used in theory of evolution (Tenaillon, 2014; Joyce et al., 2008; Orr, 2005).

By building upon this direct equivalence as well as some heuristics with regards to how much the FGM in the Orr-Gillespie formalization can be generalized outside its formal description (Joyce et al., 2008; Rokyta et al., 2008), we suggest additional insights on some features of ANNs trained with SGD, notably with regards to generalization and minima flatness, as well as implications for the EA family with which we drew the equivalence. We then offer a number of hypotheses applicable to ANNs trained with SGD, based on behaviors observed in biological system undergoing natural evolution.

1.1 Fisher Geometric Model in the Orr-Gillespie formalization

The Fisher Geometric Model was introduced by sir Ronald Fisher as a cornerstone in his effort to unify the theory of evolution and reconcile Biometricists, convinced of the gradual evolution as presented by Darwin, and Geneticists, convinced of combinatorial genetic inheritance presented by Mendel (for a detailed review see Orr (2005)). The elegance of the model consisted in suggesting that the fitness, rather than being absolute, was a result of a mach between the organism and its environment, mediated by a set of traits that cold be adjusted to enable such an adaptation - aka a phenotypic space. Within this space, environment allowed the organism to achieve a maximal fitness in a single point, with the fitness gradually decaying as the organism moved further from the optimum. Gene alleles could either have a drastic effect on the phenotypic space, in which case their inheritance was Mendelian, or small, in which mixing of alleles from parents provided an illusion of gradual change (Fisher, 1930).

The resulting visual aid, presenting a stochastic walk towards the optimum on the fitness landscape, driven by the successive mutations and selections, is perhaps the most iconic aspect of the Fisher Geometric model and was declined for their own needs by fields with their own version of greedy exploration of fitness (or loss) landscapes.

However simplicity did not seem to bode well with the biological reality. The slow and gradual adaptation suggested by the model was at odds with fossil records (Gould & Lewontin, 1979) and couldn't be easily mapped to the recently discovered biological reality of the almost-binary-tape DNA sequence controlling the biological machinery of organisms and radically differences in phenotypes, back then almost always tracked to modifications in a single gene. Most importantly, the general and statistical nature of the FGM provided little insight into which genes were specifically under selection in organisms for different phenotypes

These shortcomings of the classical Fisher model, associated with a general change of perspective on the evolutionary genetics, led to a transition towards sequence substitution - oriented evolution. One of the first post-FGM models was the purely neutral drift model, where all the genetic and phenotype variation was a product of randomness and a number of successive bottleneck effects (Kimura & Crow, 1964; Kimura, 1968). With hard experimental evidence that some mutations were indeed advantageous and a large number was indeed deleterious, the model was refined to the near-neutral theory of evolution, where the vast majority of mutations had no effect on the fitness of the organism, a significant part had a deleterious effect, of which only a small fraction had a significant impact, and only a vanishingly small number of mutations were beneficial (Ohta, 1992).

However, the real breakthrough that allowed the FGM to be resurrected occurred through a paradigm shift in theoretical biology. First, a realization that fitness is environment-dependent. In other terms, organism that are highly adapted to their environment are subject to no evolutionary pressure and can remain unchangend for hundrends of millions of years, like the customary example of the horseshoe crab. An evolution will start only if a shift in fitness peak occurs , usually accompagniated by the initial population contraction allowing a rapid adaptive burst to occur (Lande, 1986), sometimes by leveraging pre-existing diversity in the genetic code space accumulated by neutral drift (Kauffman, 1969; Kauffman & Johnsen, 1991). Second, by leveraging recent results from limit distribution - specifically the Fisher-Tippet-Gnedenko (Fréchet, 1927; Fisher & Tippett, 1928; Mises, 1936; Gnedenko, 1943), closely realed to the generalization of the Central limit theorem by Gnedenko et al. (1968), Gillespie (1983; 1984) established that if the adaptative burst occurred under high selection/low mutation condition, starting from one already fit organism, the iterative steps in the genetic code space towards the most adapted genetic code would increase the fitness according to the Gumbel limit distribution. Consequently, Orr (2002; 2006) showed that the distribution of fitness of steps in the phenotypic space in the FGM model belonged to the same bassin of attraction as the distribution of fitness of steps in the genetic code space. In turn, it meant that the FGM could be used as a convenient visual model of evolutionary processes occurring in the genetic codes space, with concepts such as phenotypic dimension or fitness peak flatness becoming directly interpretable. Follow-up work by Joyce et al. (2008) showed that the conclusions reached for the distributions in the Gumbel limit distribution bassin of attraction still mostly held in the vicinity of it, in the bassins of attraction of Weibull and Fréchnet distributions.

Remarkably, the Gillespie-Fisher model of evolution is in no way linked to biological genetic codes specifically. It is applicable to any finite codes for which a fitness (or conversely loss) is defined and that are being iterative modified by a greedy search algorithm, starting from a code with already a high fitness value.

Fine-tuning real-world ANNs through SGDs is an instance of such code space search.

2 CENTRAL RESULTS

2.1 GILLESPIE-ORR EVOLUTIONARY ALGORITHM

In their formulation of the evolutionary process, Gillespie and Orr make three fundamental assumptions to make their model analytically tractable. Specifically:

- Haploid populations (single code evaluated for fitness)

- Under high selection ($Ns \gg 1$)
- In the low mutation limit ($N\mu < 1$)

Where N is the population size, s is a typical selection coefficient and μ is the per-site mutation rate.

The main purpose of those assumptions is to ensure that a new advantageous mutation swipes [1] through the population entirely upon appearance, before a next advantageous mutation can emerge and ensuring that a deleterious mutation never appears at the same time as an advantageous one.

While covering an important class of biological questions, such as drug resistance in cancers and bacteria Tenaillon (2014), this formulation is restrictive from the population genetics point of view. It is, however, perfectly adapted for the ANN neuroevolution, whenever after a mutation round only the highest fitness (conversely lowest loss) parameter $\boldsymbol{\theta}$ is retained and no parameter mixing occurs. We will hence refer to such an Evolutionary Algorithm as *Gillespie-Orr Evolutionary Algorithm* (GO-EA)

2.2 SGD EQUIVALENCE TO THE GILLESPIE-ORR EA FOR A SUFFICIENTLY LOW LEARNING RATE

Let $f_{\boldsymbol{\theta}}(\cdot)$ be a neural network parametrized by $\boldsymbol{\theta}$, that maps inputs $\mathbf{X}$ of the form $\{\mathbf{x_i}\}_{\mathbf{i=1}}^{\mathbf{M}} \in \mathbb{Z}_{\mathbf{2}}^{\mathbf{n_x} \times \mathbf{d_x} \times \mathbf{M}}$ to outputs $\mathbf{Y}$ of the form $\{\mathbf{y_i}\}_{\mathbf{i=1}}^{\mathbf{M}} \in \mathbb{Z}_{\mathbf{2}}^{\mathbf{n_y} \times \mathbf{d_y} \times \mathbf{M}}$, where $\mathbb{Z}_2 = \{0, 1\}$, d_x and d_y are respectively the dimensionalitity of $\boldsymbol{x}$ and $\boldsymbol{y}$, n_x and n_y respectively the binary code length required to describe a single component of the vectors of $\boldsymbol{x}$ and $\boldsymbol{y}$ and M the maximum number of inputs the network can encounter, with potentially $M = \inf$.

Let $\mathcal{L}_{\boldsymbol{\theta}}$ be the fitness function associated to $f_{\boldsymbol{\theta}}$ on the $\mathbf{X}$ and $\mathbf{Y}$. A priory, $\mathcal{L}$ is inaccessible, given it requires an evaluation on all the possible input-output pairs. However, it can estimated with a finite sample of inputs and outputs $\mathbf{X_{samp}}$, $\mathbf{Y_{samp}}$, giving us an $\hat{\mathcal{L}}_{\boldsymbol{\theta}}|_{\mathbf{X}_{samp}, \mathbf{Y}_{samp}}$.

Let $\mathcal{O}$ be a greedy optimization process, such that $\mathcal{O}(\boldsymbol{\theta}) = \boldsymbol{\theta}'$, with a rewrite capacity d, such that $||\boldsymbol{\theta}' - \boldsymbol{\theta}||_p < d$, where $p \in \mathbb{N}$ and $\hat{\mathcal{L}}_{\boldsymbol{\theta}'}|_{\mathbf{X}_{samp}, \mathbf{Y}_{samp}} \geq \hat{\mathcal{L}}_{\boldsymbol{\theta}''}|_{\mathbf{X}_{samp}, \mathbf{Y}_{samp}}$ for any $\boldsymbol{\theta}''$ such that $||\boldsymbol{\theta}'' - \boldsymbol{\theta}||_p < d$.

Greedy optimization processes SGD and GO-EA are almost surely equivalent in the limit of SGD learning rate $l \to 0$ and GO-EA neighbourhood sample population $N \to \inf$ with GO-EA rewrite capacity $d = l$, up to a saddle point.

Since the SGD is applicable, we can perform the Taylor expansion in the neighbourhood of $\boldsymbol{\theta}$ of $\hat{\mathcal{L}}_{\boldsymbol{\theta}''}|_{\mathbf{X}_{samp}, \mathbf{Y}_{samp}}$. As $d \to 0$, only the first order terms of the expansion remain, meaning that a gradient descent of the loss function ($-\hat{\mathcal{L}}$) with a step of l will lead to the optimum within the d ball around $\boldsymbol{\theta}$. Given GO-EA has an infinite population, it will find the same optimum within the ball d. In case all the first order terms are nill, either the greedy optimization algorithm acheived a local minimum, in which case both SGD and GO-EA will stay put, or it is located in a local saddle point, in which case the SGD will acheive no movement and GO-EA will move to the hightest fitness point within d of $\boldsymbol{\theta}$. In real conditions, the noise level provided by the sampling on the fitness function would lead the probability of the saddle point to vanish, if the samples on which $\hat{\mathcal{L}}$ is evaluated are different.

In order to achieve this result, we had to introduce the rewrite distance on the model parameters $\boldsymbol{\theta}$, which operates on a norm and seemingly is incompatible with the single-mutation edits with which GO-EA operates in the context of genetic codes in biological organism. This is not the case. The parametrization $\boldsymbol{\theta}$ combined with the ANN architecture is just one of the many possible ways to encode the model $f_{\boldsymbol{\theta}}(\cdot)$, and it is certain that more compact and efficient codings exist, where the transition to a local minimum would correspond to single code character change.

[1] In population genetics and theory of evolution, an allele is said to swipe through the population when it prevalence increases until every single individual in the population has it. At this point, it is said to have been fixated in the population

2.3 PROBABILITY OF FINDING BETTER PARAMETERS DURING FINE-TUNING WITH GILLESPIE-ORR EA

In the context of fine tuning, we expect to start off a with a model $f_{\boldsymbol{\theta}_0}(\cdot)$ parametrized so that it already performs well on all the sample tests drawn from the distribution it was used to train with - aka $\forall(\mathbf{X}_{samp}, \mathbf{Y}_{samp}) \subset \mathbf{X} \times \mathbf{Y}, \mathbb{P}(\hat{\mathcal{L}}_{\boldsymbol{\theta}_0}|_{\mathbf{X}_{samp},\mathbf{Y}_{samp}} \sim \max_{\boldsymbol{\theta}} \hat{\mathcal{L}}_{\boldsymbol{\theta}}|_{\mathbf{X}_{samp},\mathbf{Y}_{samp}}) \sim 1$

Formally, fine-tuning consists in finding a new transfer parametrization $\boldsymbol{\theta}_T$, so that $\forall(\mathbf{X}_{samp}, \mathbf{Y}_{samp}) \subset \mathbf{X} \cup \mathbf{X}' \times \mathbf{Y} \cup \mathbf{Y}', \mathbb{P}(\hat{\mathcal{L}}_{\boldsymbol{\theta}_T}|_{\mathbf{X}_{samp},\mathbf{Y}_{samp}} \sim \max_{\boldsymbol{\theta}} \hat{\mathcal{L}}_{\boldsymbol{\theta}}|_{\mathbf{X}_{samp},\mathbf{Y}_{samp}}) \sim 1$, where the $\mathbf{X}'$ and $\mathbf{Y}'$ are new domains application of the model.

Assuming $|\boldsymbol{X}| >> |\boldsymbol{X}'|$ and $|\boldsymbol{Y}| >> |\boldsymbol{Y}'|$ (otherwise fine-tuning would be equivalent to model re-training), the model is already performing well on the fine-tuned model and the vast majority of the parameters within rewrite capacity d of $\boldsymbol{\theta}_0$ would be deleterious or neutral, meaning that the parametrizations offering improvement would be distributed according to the generalized Pareto distribution (Pickands, 1975; Joyce et al., 2008), which in the case of Gumbel domain of attraction would result in an exponential distribution of fitnesses $\boldsymbol{s} = (s_1, ..., s_{i-1})$ where $s_j = \hat{\mathcal{L}}_{\boldsymbol{\theta}_j}|_{\mathbf{X}_{samp},\mathbf{Y}_{samp}}$, the j^{th} best parametrization of of better parametrizations and a probability to reach the better parametrization $\boldsymbol{\theta}_j$ of rank j in the neighbourhood from a parametrization $\boldsymbol{\theta}_i$ of the rank i of $\mathbb{P}_{i,j}(\boldsymbol{s}) = \frac{s_j}{\sum_{k=1}^{i-1} s_k}$

In other terms, with finite populations, GO-EA sampling the parametrization neighborhood of the current optimum $\boldsymbol{\theta}_i$ will find advantageous model code rewrites with the probability that's in reverse exponential probability of the difference between the loss associated to $\boldsymbol{\theta}_i$ and smallest possible loss within the edit distance budget.

While formally proven for the distributions in the Gumbel domain of attraction, this results has been shown to hold as well in the adjacent domains of attraction of the Weibull and Frechnet limit distributions, although the behavior further away from the Gumbel domain might differ radically, leading to all ranks for fitness being equally likely to be picked up in the limit case of Weibull bassin of attraction and only the lowest rank distribution being reachable by neighborhood sampling in the limit case of Frechnet bassin of attraction (reviewed in depth by Joyce et al. (2008)).

Unfortunately, the specific size of the sampling population N needed to sample at least one advantageous parameter within the rewrite capacity is directly connected to the effective latent dimension of the model - by analogy with the phenotypic space of the FGM. While we will discuss potential strategies to estimate it in the section 3, it is not directly accessible.

3 IMPLICATIONS OF CENTRAL RESULTS

3.1 DIRECT IMPLICATIONS

3.1.1 ROBUSTNESS OF THE GILLESPIE-ORR EVOLUTIONARY ALGORITHM

A substantial amount of research has been invested to better understand how SGD interacts with ANNs architecture, managing to robustly find parametrizations for the ANNs that avoid local minima and provide reasonable noise-resistance and generalization capabilities.

To our knowledge, such results were entirely absent for the Evolutionary Algorithms until now. Thanks to the results in the 2.2, we can now claim that in case of a differentiable loss landscape, Gillespie-Orr Evolutionary Algorithm is equivalent to SGD in the limit of low learning rate and high sampling population size. This means that all the results previously shown for SGD are valid for GO-EA whenever this approximation holds.

3.1.2 COMPUTATIONAL ADVANTAGE OF SGD OVER EAS

Given these small learning rate and large sampling population approximation needed for the proof of 2.2, we can also fairly confidently say that whenever applicable, SGD is also more computationally efficient than GO-EA, given that the derivation and back-propagation are not computationally more expensive than new parameters sampling by more than the expected latent dimension of the model.

Given the simplicity of the GO-EA, we expect this computational efficiency relation to hold for other evolutionary algorithms, given that GO-EA is one of the possibly simplest ones.

3.1.3 MOST GENERALIZEABLE MODEL SELECTION AND EFFECTIVE EMBEDDING DIMENSION DETECTION

Prior work on diverse populations of biological systems evolving in a manner compatible with the Gillespie-Orr formalization has shown that it was possible to both estimate the effective phenotypic dimension, which for ANNs correspond to the latent dimension of the model, as well as to select within the population the sub-population that was the best at general performance (Kucharavy et al., 2018).

Specifically, in order to perform the latent dimension extraction, this work leveraged the Gnedenko-Kolmogorov formulation of the Central limit theorem, Gnedenko et al. (1968) that has proved that if a complex system can be altered in a large number of random ways, the resulting deviation from the base state converges to a Gaussian. Specifically, such a deviation would be visible along each axis relevant to adaptation to the environment, with the total deviation being characterized by the Chi-n function, where n is the effective phenotypic dimension of the organism/environment match, up to a renormalization. By subjecting this population to a number of diverse environments, it was possible to leverage the relationship between the mean and standard deviation of the fitnesses of the heterogeneous population among different population to both calculate the effective phenotypic dimension in which the population could move to match the environment, as well as the population that were the best at dealing with all the environments.

In the context of model fine-tuning, this could mean that in presence of a sufficiently diverse set of validation datasets and heterogeneous models, differing either by their architecture, initialization, hyperparameters or training dataset, it would be possible to both evaluate the effective latent dimension of the model family and the problem and to determine the model that is most likely to be a good starting point for fine-tunes that could cover all the datasets. Alternatively, the model ability to retain generality across validation datasets could also be used to identify models within the family least prone to catastrophic forgetting during the transfer learning.

3.2 HYPOTHESES

Whereas the previous subsection was dedicated to the direct implications of our central results, here we present several hypotheses based on the conceptual framework of evolutionary algorithms

3.2.1 MINIMA FLATNESS AS ERROR CORRECTION REDUNDANCY

SGD converging to flat minima is one of the conditions on the architecture of the ANN models for their training to be stable (Li et al., 2018).

Minima flatness was considered to the generalization abilities of the model through its presumed relationship to the minimal coding length of the model (Hochreiter & Schmidhuber, 1997; Goodfellow & Vinyals, 2015), although recently evidence to the contrary emerged (Dinh et al., 2017; Zhang et al., 2020; Mulayoff & Michaeli, 2020).

Within the theory of evolution, the flatness of the fitness peak is commonly associated to the tolerance to the neutral drift - aka error correction capabilities. By using this analogy, we suggest that just like in the context of the evolution, the flatness of the loss function minimum in ANNs optimized through SGD is determined by the redundnancy of the features used by the trained ANN to recognize patterns in the target data.

This intuition seems to be consistent with empirical observations about the loss function minima flatness. Architectures that provide the model with means to encode redundant features, such as with extremely large hidden layers or with skip-forwards connections in deep Convolutional Feed-Forwards ANNs, contribute to making the loss landscape minima more flat, as demonstrated by Li et al. (2018). Similarly, drop-out regularization Srivastava et al. (2014), forcing the ANNs to learn redundant, error-correcting codings seem to flatten minima as well, along with the smaller batches, which can contain a large proportion of samples that defy the heuristics that the ANN has learnt until now (Goodfellow & Vinyals, 2015)

From this perspective, we do not expect flatter minima to lead to better generalization, but rather to allow for more robust and less noise-sensitive models, which seems to be confirmed by the numerical experiments showing that ANNs with architectures, regularizations and training modes known to lead to flatter minima also tend to memorize less (Arpit et al., 2017).

3.2.2 FLAT MINIMA AND TRANSFER LEARNING

Building on top of the hypothesis presented above, if the minima flatness is indeed related to the classification robustness and error correction, we expect models that learnt a variety of error-correcting representations of training data to not be able to transfer those representations without training onto new data presenting similar features.

Intuitively, they rely on a simultaneous redundant subpaths through their ANN layers detecting redundant relevant features present in the training dataset. With only some of those features present in the dataset on which the transfer task is performed, their error correction property is likely to interfere with the the output of a corresponding output without an expected degree of redundant detection.

If this hypothesis is correct, a particular attention need to be payed when training ANNs that need to be both robustly map inputs to outputs in a noise-resistant manner and can be exposed to rare inputs presenting only some of the features on which the action need to be taken. We expect this problem to be separate from the adversarial examples one and more closely related to the generalization one, given that its goal is to recognize partial features.

4 HEURISTICS FROM THEORY OF EVOLUTION TO ACCELERATE THE FINE-TUNING PROCESS

Here we present a number of heuristics that are thought to be critical for the acceleration of the natural evolution, that we expect to be transferable to the training of well-formed ANNs through SGD.

4.1 MODEL MIXING IS NOT NECESSARY, ALTHOUGH CAN BE BENEFICIAL.

One of the prominent features of the later Genetic Programming compared to the early evolutionary algorithms was the emphasis put on mixing the models through ”chromosome” ”recombination”. Directly mapping to the importance of the sexual reproduction in the context of natural evolution, it is neither necessary nor applicable in the context of GO-EA or SGD. Unlike in biological systems, there is no spurious mutation accumulation, so there is no need for the purifying selection to prevent Muller's ratchet from eliminating the population of ANNs (Lynch et al., 1993). Similarly, studies of model parameter interpolation between two good solutions indicate that the intermediate parameters tend to perform uniformly poorly without dedicated regularization (Goodfellow & Vinyals, 2015).

In case the model mixability is desireable however, for instance for rapid model aggregation attempts, or in the case the models independently drift away from the optimum, it is possible to develop regularization schemes that would preserve mixability (Livnat et al., 2008). Another reason such a mixability might be desirable, is that it might promote the diversity and redundance of feauture extraction subnets in the ANNs, further improving training stability and resilience to memorization.

4.2 RAPID EXPLORATION OF THE LATENT FEATURE SPACE

One of biological systems that are most consistent with the assumptions of the Gillespie-Orr model of natural evolution are pathogens developing resistance to treatments. While the means to achieve such resistance differ between the pathogens and treatments, they nonetheless use a common trick to accelerate their adaptation before going extinct. Specifically, rapidly generate random variation in the phenotypic space, in hopes that at least one of such variant populations would be more fit than the original one Kucharavy et al. (2018). This trick seems to be highly efficient, allowing the pathogens to adapt to new stressful conditions in a matter of a dozen of generations as opposed to thousands that would be expected in case of a gradual evolution. Given that the underlying model

for theoretical work of Kucharavy et al. (2018) is FGM with asexual reproduction, the results are fully compatible with GO-EA and hence with models fine-tuned with SGD.

To summarize the results from (Kucharavy et al., 2018), the dimensionality detection algorithms relies on a family of related models (for instance members of a population with random perturbation to parameters compared to a reference model), each evaluated on a heterogeneous benchmark. Based on the correlation of the average performance of models on each test task in the benchmark compared to the standard deviation of the models performance on each test, Kucharavy et al. (2018) shows that there is an expected correlation between the two and that a direct regression with two orthogonal parameters is possible to evaluate the underlying dimension of the problem. It is important to note that the dimension of the problem is not intrinsic to the problem at hand but also involves the architecture and parameter values of the ANN that has been trained to solve it, corresponding to the amount of independent "axes" along which the ANN can move to better adapt to task.

The mechanism we expect to limit the catastrophic forgetting requires a cross-evaluation of a family of models on a heterogeneous benchmark. Unlike the problem dimension evaluation, it evaluates the average performance of a model, as well as how uniformly it performs relative to other models on different tasks in the heterogeneous benchmark. Specifically, by calculating the Gini index of the model performance, Kucharavy et al. (2018) predicts that it will be inversely correlated with model performance and that the model with the lowest inequality of performance across tasks would also have reasonable performance. Given that it have a good average performance and perform well across most tasks in the benchmark, we expect that this model has not undergone catastrophic forgetting. It can be seen as a regularization that leverages intrinsic parallelizability of the GO-EA class of algorithms and ensures that as transfer learning is performed, at each exploration-selection cycle the loss of performance on other tasks in the benchmark is minimal while the new task is learnt. This regularization can also be applied in case of parallel model-fine tuning with SGD, resulting in a family of related models, from which the least "forgetfull" model is selected.

By analogy between the GO-EA and SGD adapted to fine-tuning the models, we expect that if a random variation was injected into the ANNs before starting the fine-tuning process, at the level where they would be getting close to leaving the flat minimum in their original loss space, their fine-tuning could be significantly accelerated, even if the least performing models are eliminated after the first couple of fine-tuning training epochs performed in parallel.

5 Discussion

In this paper we establish a formal equivalence between the Gillespie-Orr model of evolution and SGD applied to ANNs. Build on top of strong limit distribution convergence results established by the Fisher-Tippet-Gnedenko theorem, GO-EA leverages the representation of ANNs as learnt codes that are greedily improved through neighbourhood exploration to provide an insight into how SGD might work and how it can be improved, as well as to what can be expected from parameter space search with GO-EA.

While we expect GO-EA algorithms to be more computationally expensive than SGD on differentiable manifolds, it does have two important application domains. First, in the cases where the evaluation of gradient and backpropagation of gradients are significantly more computationally expensive than the evaluation of the model performance (on the order of magnitude of the number of model parameters). In this circumstances, parallel computational capabilities and tricks allowing a more efficient communication of model updates, such as introduced in Such et al. (2017) can allow a faster and more computationally efficient model training. Second, in cases where the loss surface can be assumed to be smooth, but cannot be directly differentiated, such as in behavior strategies learning (as for instance described by Majumdar et al. (2020)).

While on the surface that last case seems to be rather limited, research in the context of SGD (Li et al., 2018) have shown that the smoothness of the loss landscape is dependent on the ANN architecture and parameter values rather than the problem alone. This suggests that there are classes of problems that area currently assumed to be difficult due to non-smooth loss landscapes that can be made more approachable by new ANN architectures and hence be efficiently explored by GO-EA algorithms.

Based on that formal equivalence we offer a new perspective on the minima flatness, which we link to the redundancy of the compressed codes representing the learnt model, rather than the minimal code length, as suggested previously, showing that is consistent with the experimentally observed results linking loss landscape flatness with skip connections, hidden layers width and the use of dropout regularization.

We further build on this insight in order to hypothesize that it is possible to identify the number of latent dimensions used by a model family to learn to map a training dataset inputs to corresponding outputs, as well as the most generalizeable model in a population, provided a sufficiently diverse set of secondary validation datasets is available.

Finally, we suggest a number of heuristics we expect would accelerate model fine-tuning or EA application to ANNs in general.

While the paper could greatly benefit from the experimental validation of hypotheses and heuristics presented here, the multiple model training restarts to collect statistics, hyperparameter space exploration and model population sizes needed for the EA methods mean that it is a task better suited for an entity with large computational capabilities and could represent a work in its own right.

In fact, ResNet-56 on CIFAR-10 used in (Li et al., 2018) to evaluate minima flatness requires 18 hours of training time on top-of-the line consumer hardware. An initial training run combined with fine-tuning is likely to multiply this time by a factor of magnitued, whereas orthogonal filter space search on a grid with filter normalization on the pre-fine-tuned model and fine-tuned models with intermediate steps would require similar a similar of magnitude of compute time. Combined with multiple restarts from different random seeds in order to obtain statistics on results, this means single experiment run times in the 70-80 days range of GPU time, assuming no other bottlenecks. While those experiments are trivially parallelizable, they require access to a cluster with sufficient compute budget to run them.

Moreover, our paper is consistent with prior experimental results, such as presented in Such et al. (2017). There, authors define a "genetic algorithm" that is fully elitist and does not proceed to any recombination and is hence an algorithm in the GO-EA class. By using this algorithm, they observe a number of features initially discovered and theoretically explored in the context of SGD, such as for instance local density of good solutions near a random initialized vector in case of sufficient model over-parametrization (Jacot et al., 2018). However, even a single training run of their model to train their 4 million parameter ANN for a single Atari game required 720 CPU core-hours for a single run - or 4-8 days of wall time on consumer-grade CPUs with 4-8 cores. Multiple restarts would be required to collect representative and comparable results, leading to simulation run times in the 20-40 days on consumer hardware.

We hope that our work provides novel perspective on the SGD convergence in the context of ANNs, as well as Evolutionary Algorithm application more generally.

REFERENCES

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017. URL http://proceedings.mlr.press/v70/arpit17a.html.

David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 342–350. PMLR, 2017. URL http://proceedings.mlr.press/v70/balduzzi17b.html.

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. *Unsupervised and Transfer Learning Challenges in Machine Learning*, 7:19, 2012.

Yoshua Bengio, Frédéric Bastien, Arnaud Bergeron, Nicolas Boulanger-Lewandowski, Thomas M. Breuel, Youssouf Chherawala, Moustapha Cissé, Myriam Côté, Dumitru Erhan, Jeremy Eustache, Xavier Glorot, Xavier Muller, Sylvain Pannetier Lebeuf, Razvan Pascanu, Salah Rifai, François Savard, and Guillaume Sicard. Deep learners benefit more from out-of-distribution examples. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík (eds.), *AISTATS*, volume 15 of *JMLR Proceedings*, pp. 164–172. JMLR.org, 2011. URL http://dblp.uni-trier.de/db/journals/jmlr/jmlrp15.html#BengioBBBBCCCEEGMLPRSS11.

S Bozinovski and A Fulgosi. The influence of pattern similarity and transfer learning upon training of a base perceptron b2. In *Proceedings of Symposium Informatica*, pp. 3–121, 1976.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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, 2020. URL http://arxiv.org/abs/2005.14165. cite arxiv:2005.14165Comment: 40+32 pages.

Rich Caruana. Learning many related tasks at the same time with backpropagation. In G. Tesauro, D. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems*, volume 7, pp. 657–664. The MIT Press, 1995.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL http://proceedings.mlr.press/v70/dinh17b.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *CoRR*, abs/1701.08734, 2017. URL http://dblp.uni-trier.de/db/journals/corr/corr1701.html#FernandoBBZHRPW17.

R. A. Fisher. *The genetical theory of natural selection*. Oxford University Press, Oxford, 1930. ISBN 0-19-850440-3. URL http://www.archive.org/details/geneticaltheoryo031631mbp.

Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pp. 180–190. Cambridge University Press, 1928.

Dario Floreano, Peter Dürr, and Claudio Mattiussi. Neuroevolution: from architectures to learning. *Evol. Intell.*, 1(1):47–62, 2008. URL http://dblp.uni-trier.de/db/journals/evi/evi1.html#FloreanoDM08.

L.J. Fogel, A.J. Owens, and M.J. Walsh. *Artificial intelligence through simulated evolution*. Wiley, Chichester, WS, UK, 1966.

Maurice Fréchet. Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Math. Polon.*, 6:93–116, 1927.

Edgar Galván and Peter Mooney. Neuroevolution in deep neural networks: Current trends and future challenges. *CoRR*, abs/2006.05415, 2020. URL http://dblp.uni-trier.de/db/journals/corr/corr2006.html#abs-2006-05415.

John H Gillespie. A simple stochastic gene substitution model. *Theoretical population biology*, 23 (2):202–215, 1983.

John H Gillespie. Molecular evolution over the mutational landscape. *Evolution*, pp. 1116–1129, 1984.

Boris Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, pp. 423–453, 1943.

B.V. Gnedenko, A.N. Kolmogorov, and K.L. Chung. *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley Mathematical Series. Addison-Wesley, 1968. URL https://books.google.ch/books?id=rYsZAQAAIAAJ.

David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA, 1989.

Faustino J. Gomez and Risto Miikkulainen. Incremental evolution of complex general behavior. *Adapt. Behav.*, 5(3-4):317–342, 1997. doi: 10.1177/105971239700500305. URL https://doi.org/10.1177/105971239700500305.

Faustino J. Gomez, Jürgen Schmidhuber, and Risto Miikkulainen. Accelerated neural evolution through cooperatively coevolved synapses. *J. Mach. Learn. Res.*, 9:937–965, 2008. URL https://dl.acm.org/citation.cfm?id=1390712.

Ian J. Goodfellow and Oriol Vinyals. Qualitatively characterizing neural network optimization problems. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6544.

Stephen Jay Gould and Richard C Lewontin. The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the royal society of London. Series B. Biological Sciences*, 205(1161):581–598, 1979.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evol. Comput.*, 9(2):159–195, 2001. doi: 10.1162/106365601750190398. URL https://doi.org/10.1162/106365601750190398.

Nikolaus Hansen, Dirk V. Arnold, and Anne Auger. Evolution strategies. In Janusz Kacprzyk and Witold Pedrycz (eds.), *Springer Handbook of Computational Intelligence*, Springer Handbooks, pp. 871–898. Springer, 2015. doi: 10.1007/978-3-662-43505-2_44. URL https://doi.org/10.1007/978-3-662-43505-2_44.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In Gerald Tesauro, David S. Touretzky, and Todd K. Leen (eds.), *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, pp. 529–536. MIT Press, 1994. URL http://papers.nips.cc/paper/899-simplifying-neural-nets-by-discovering-flat-minima.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Comput.*, 9(1):1–42, 1997. doi: 10.1162/neco.1997.9.1.1. URL https://doi.org/10.1162/neco.1997.9.1.1.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989. ISSN 0893-6080. doi: http://dx.doi.org/10.1016/0893-6080(89)90020-8. URL http://www.sciencedirect.com/science/article/pii/0893608089900208.

HuggingFace. HuggingFace Model repository. https://huggingface.co/models, 2020. [Online; accessed 05-October-2021].

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8580–8589, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Paul Joyce, Darin R Rokyta, Craig J Beisel, and H Allen Orr. A general extreme value theory model for the adaptation of dna sequences under strong selection and weak mutation. *Genetics*, 180(3): 1627–1643, 2008.

Stuart A Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3):437–467, 1969.

Stuart A Kauffman and Sonke Johnsen. Coevolution to the edge of chaos: coupled fitness landscapes, poised states, and coevolutionary avalanches. *Journal of theoretical biology*, 149(4): 467–505, 1991.

Motoo Kimura. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetics research*, 11(3):247–270, 1968.

Motoo Kimura and James F Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725, 1964.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

Andrei Kucharavy, Boris Rubinstein, Jin Zhu, and Rong Li. Robustness and evolvability of heterogeneous cell populations. *Molecular biology of the cell*, 29(11):1400–1409, 2018.

Russell Lande. The dynamics of peak shifts and the pattern of morphological evolution. *Paleobiology*, 12(4):343–354, 1986.

Yann LeCun. Une procedure d'apprentissage ponr reseau a seuil asymetrique. *Proceedings of Cognitiva 85*, pp. 599–604, 1985.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6391–6401, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html.

Adi Livnat, Christos Papadimitriou, Jonathan Dushoff, and Marcus W Feldman. A mixability theory for the role of sex in evolution. *Proceedings of the National Academy of Sciences*, 105(50):19803–19808, 2008.

Michael Lynch, Reinhard Bürger, D Butcher, and Wilfried Gabriel. The mutational meltdown in asexual populations. *Journal of Heredity*, 84(5):339–344, 1993.

Somdeb Majumdar, Shauharda Khadka, Santiago Miret, Stephen Mcaleer, and Kagan Tumer. Evolutionary reinforcement learning for sample-efficient multiagent coordination. In *International Conference on Machine Learning*, pp. 6651–6660. PMLR, 2020.

R von Mises. La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique*, 1: 141–160, 1936.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nat.*, 518(7540):529–533, 2015. doi: 10.1038/nature14236. URL https://doi.org/10.1038/nature14236.

Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7108–7118. PMLR, 2020. URL http://proceedings.mlr.press/v119/mulayoff20a.html.

Quynh Nguyen. On connected sublevel sets in deep learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4790–4799. PMLR, 2019. URL http://proceedings.mlr.press/v97/nguyen19a.html.

Quynh Nguyen, Pierre Bréchet, and Marco Mondelli. On connectivity of solutions in deep learning: The role of over-parameterization and feature quality. *CoRR*, abs/2102.09671, 2021. URL https://arxiv.org/abs/2102.09671.

Tomoko Ohta. The nearly neutral theory of molecular evolution. *Annual review of ecology and systematics*, 23(1):263–286, 1992.

Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pp. 1717–1724. IEEE Computer Society, 2014. ISBN 978-1-4799-5118-5. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2014.html#OquabBLS14.

H Allen Orr. The population genetics of adaptation: the adaptation of dna sequences. *Evolution*, 56 (7):1317–1330, 2002.

H Allen Orr. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, 6(2): 119–127, 2005.

H Allen Orr. The distribution of fitness effects among beneficial mutations in fisher's geometric model of adaptation. *Journal of theoretical biology*, 238(2):279–285, 2006.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22 (10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191. URL https://doi.org/10.1109/TKDE.2009.191.

D. B. Parker. Learning-logic. Technical Report TR-47, Center for Comp. Research in Economics and Management Sci., MIT, 1985.

James III Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1): 119–131, 1975. ISSN 00905364. doi: 10.1214/aos/1176343003. URL http://www.jstor.org/stable/2958083.

Lorien Y. Pratt. Discriminability-based transfer between neural networks. In Stephen Jose Hanson, Jack D. Cowan, and C. Lee Giles (eds.), *Advances in Neural Information Processing Systems 5, [NIPS Conference, Denver, Colorado, USA, November 30 - December 3, 1992]*, pp. 204–211. Morgan Kaufmann, 1992. URL http://papers.nips.cc/paper/641-discriminability-based-transfer-between-neural-networks.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214%2Faoms%2F1177729586.

Darin R Rokyta, Craig J Beisel, Paul Joyce, Martin T Ferris, Christina L Burch, and Holly A Wichman. Beneficial fitness effects are not exponential for two viruses. *Journal of molecular evolution*, 67(4):368, 2008.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. URL http://dx.doi.org/10.1038/323533a0.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

HP. Schwefel. *Numerical optimization of computer models*. Wiley, Chichester, WS, UK, 1981.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958, 2014. URL http://dl.acm.org/citation.cfm?id=2670313.

Felipe Petroski Such, Vashisht Madhavan, Edoardo Conti, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. *CoRR*, abs/1712.06567, 2017. URL http://arxiv.org/abs/1712.06567.

Olivier Tenaillon. The utility of fisher's geometric model in evolutionary genetics. *Annual review of ecology, evolution, and systematics*, 45:179–201, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

P. J. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974.

Daan Wierstra, Tom Schaul, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2008, June 1-6, 2008, Hong Kong, China*, pp. 3381–3387. IEEE, 2008. doi: 10.1109/CEC.2008.4631255. URL https://doi.org/10.1109/CEC.2008.4631255.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3320–3328, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small relu networks are powerful memorizers: a tight analysis of memorization capacity. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 15532–15543, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/dbea3d0e2a17c170c412c74273778159-Abstract.html.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Michael C. Mozer, and Yoram Singer. Identity crisis: Memorization and generalization under extreme overparameterization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1l6y0VFPr.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776.