/mlst

Build Specialist LLMs Like It's 2019 (Randall Balestriero)

March 23, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: 10603272

Topic headings include retention indicators: ● above average, ● average, ● below average (vs. similar YouTube videos)

Interested in licensing ReScript for your podcast? Get in touch with Tim Scarfe.

Contents

Overview 2

Random Initialization Rivals Pre-Training 3

Is Next-Token Prediction Worth the Cost? 4

What Do These Models Actually Learn? 5

Build Specialist LLMs Like It Is 2019 7

The Fair Language Model Paradox 8

Benchmarks, Generation, and Understanding 9

The Birth of Self-Supervised Learning 11

Class Balance, VICReg, and Unifying Representation Learning 12

No Location Left Behind: Fairness in Earth Models 15

Policy, Accountability, and Crowdsourced Data Bias 17

References 20

Overview

Randall Balestriero (Meta AI) shares three recent results that each push back on conventional wisdom in ML.

First, the headline finding: if you take a 7-billion-parameter language model, initialize it randomly, and train it from scratch on just 20,000 labeled examples for a classification task like sentiment analysis, it works. Stable training curves, minimal overfitting, performance that matches LoRA-finetuned pre-trained models. The obvious question – is months of expensive pre-training on internet-scale data actually worth it? – gets a surprisingly qualified answer. For narrow discriminative tasks, random initialization is competitive. Pre-training still wins for generation and open-ended reasoning, but there is a whole spectrum between the two extremes that nobody is really exploring yet.

Second, a theoretical result with Yann LeCun proving that self-supervised and supervised learning objectives are mathematically equivalent up to how you define the label structure. SSL does not learn better representations because of its loss function; it learns them because it uses finer-grained pairwise relationships instead of collapsing all cars into "car." This equivalence lets you port decades of supervised learning theory – class imbalance corrections, neural collapse results, semi-supervised weighting – directly into SSL, and Randall walks through how VICReg falls out naturally from a least-squares supervised objective under this framework.

Third, a fairness audit of implicit neural representations used for earth/climate data. Models that look accurate on average turn out to be nearly random around islands and coastlines – exactly the places where policy decisions about climate adaptation matter most. The culprit is partly architectural: Fourier bases assume stationarity, and switching to wavelets recovers some of the lost localization. But the deeper problem is data bias, including the same geographic skew Mark Ibrahim documented in ImageNet, where most training images come from North America.

Random Initialization Rivals Pre-Training

Randall Balestriero 00:00:00

We just launched this experiment and then we are very surprised to see that the hugely overparameterized model not only train out of the box, like you have very nice training curves, but also they don't overfit aggressively at all. And what we found empirically is that we just out of the box use typical supervised training. We don't have to play with hyperparameter optimizer and you have very, very stable training. So this brings us to the question, is it worth it to spend so much money together on gigantic pre training data set, spend like months on many GPUs to produce those models? But at least for some applications, it seems to not be much better than random.[1][2]

Dr. Tim Scarfe 00:00:45

MLST is sponsored by Tufa AI Labs. Now they are the DeepSeek based in Switzerland. They have an amazing team. You've seen many of the folks on the team. They acquired Minds AI, of course. They did a lot of great work on ARC. They're now working on o1-style models and reasoning and thinking and test time computation. The reason you wanna work for them is you get loads of autonomy. You get visibility. You can publish your research. And, also, they are hiring as well as ML engineers. They're hiring a chief scientist. They really, really want to find the best possible person for this role, and they're prepared to pay top dollar as as a joining bonus. So if you're interested in working for them as an ML engineer or their chief scientist, get in touch with Benjamin Crouzier, go to TufaLabs.ai and see what happens.

[1] Is LLM Pre-Training by Next Token Prediction Worth the Cost? — paper The central paper discussed in the first half: showing randomly initialized 7B models match pre-trained LLMs on supervised classification tasks

[2] Tufa AI Labs — organization Episode sponsor: Swiss AI lab working on reasoning and test-time computation

Is Next-Token Prediction Worth the Cost?

Randall Balestriero 00:01:29

Originally, main motivation was to see, okay, how much information you gain by doing pre training, right? And is this next token prediction really making your network learn something about language and reasoning? So then we are saying, okay, 1 way to compare this least empirically is to just take a randomly initialized model, train it from scratch on a supervised task like sentiment prediction, sentiment analysis. And then in theory, because we have a very very small training data, let's say like 20,000 samples, and because those models have like 7,000,000,000 parameters, the pre trained 1 will perform very nicely with a little bit of LoRa fine tuning because it already knows how to reason about the world, right? So maybe you just adjust a little bit to the specific task that you want. But since you have so many prior knowledge, you will solve the task very easily. But the random 1 either will overfit completely because you have like 7,000,000,000 parameters and only 20,000 training samples or maybe it will not learn at all because training dynamics will be completely chaotic. And so we just launched this experiment and then we are very surprised to see that the 7,000,000,000 or like the hugely overparameterized model not only train out of the box, you have very nice training curves, like you train MNIST, but also they don't overfit aggressively at all. They overfit less than if you just train MLP on MNIST basically. And this is very surprising. And so basically from this we said, okay, actually maybe there is a deeper question which could be how much implicit bias you have in those language model? Because already we knew from computer vision that for example ImageNet you can have a 50,000,000 model on the 1,000,000 data sets. We have this fifty:one ratio and you have the implicit bias that prevents you from overfitting and just serving the task, right? But still it's 50 to 1, so this may sound a lot for statisticians, but now it's like 7,000,000,000 to 20,000, so the ratio is gigantic, right? And yeah, no, to me it was very surprising that the scale, like the size of this ratio still allows you to learn some things that does not overfit. This is very surprising because in vision, for example, transformers are known to overfit more easily than ResNet, so they seem at least in Vision to have actually less implicit bias or implicit regularization. But at least with this type of NEXT token causal architecture or LLM, yeah, you don't seem to fit easily to your data. So this was quite surprising.

Dr. Tim Scarfe 00:04:09

Yeah, and we should bring in the name. So this was your workshop paper, self-supervised learning workshop here at NeurIPS. Correct. For perception tasks, is LLM pre training by NEXT token prediction worth the cost? So this is absolutely fascinating. So we've been given this belief that we need to have these huge pre trained models that trained on all the data on the internet. And it turns out that certainly for discrimination tasks, so things like classification, modeling generation, actually you can just start from scratch with a fairly small model, and you get sometimes even better results.

Randall Balestriero 00:04:44

Yeah,

What Do These Models Actually Learn?

Randall Balestriero 00:04:44

And even small or even large model, like you just start from scratch, you do this very simple supervised classification task, right? Okay, given this prompt, is it good or bad sentiment? Or what type of job is the prompt describing? You know, this type of, I will not call it reasoning, but more semantic classification and turns out that you start from random. Even if you have a small training data set, you will have performances that are sometimes as good than a pre trained model. So this brings us to the question, is it worth it to spend so much money together on gigantic pre training data set, spend like months on many GPUs to produce those models? And for some cases, for generation, alright, there is no question, this is what you need to do. As you have your next token prediction, you learn how to generate samples, but at least for some application it seems to not be much better than random, so it's quite interesting.

Dr. Tim Scarfe 00:05:44

So what are the differences representations?[3]

Randall Balestriero 00:05:46

So that's something we do not really look at like low dimensional representation of what you learn. It's possible. So some work try to look at the attention entropy and the mechanistic interpretability viewpoint of LLMs. So it will be interesting

[3] Lottery Ticket Hypothesis — paper Referenced during discussion of what randomly initialized models learn

to see if you have this sort of neural collapse that happens. So even if you're like a 7,000,000,000 parameter, maybe you end up learning a very, very, very simple sub network that does the task, a bit like lottery ticket hypothesis as well, and that naturally emerged from the training dynamics. Or is it really exploiting all the parameters? I think that's 1 thing. So to extend the workshop paper to conference, we want to probe into more what are the useful parameters, what did they learn. Are each layer actually learning something? Or maybe the first layers don't really learn anything, just the last few ones are learning something. So yes, there is lots of open questions for this.

Dr. Tim Scarfe 00:06:42

What does it tell us about the nature of understanding, and maybe even intelligence? Because we we think that the reason why these things understand is because they're they they just have all of these representations to all of these different, you know, things in in in their experience. Yeah. And And now we can shortcut to know to want of a better word. What does that tell us?

Randall Balestriero 00:07:02

Yeah I think that's a good question. So in this case we must look at very specific classification tasks. For example you have a description of a job, what job it is, is it a good or bad sentiment. And this you are able to solve it good, but you are not able to go out of distribution to solve a new type of question. For example, for this job description, then you cannot answer, okay, is this job paying you more than this job? Because this was not present in the training data, right? So I think you get very good models cheaply, quickly from random initialization, but they will be very specialized. And I think the benefit of having the pre may come if you want to do more of like open ended classification or reasoning. So I think it really depends on the type of application you want to solve, what's your downstream task and how much you want to generalize to new scenarios. But at least now it shows that it's not just pre training with next to can prediction, it's better for everything.

Dr. Tim Scarfe 00:07:58

So

Build Specialist LLMs Like It Is 2019

Dr. Tim Scarfe 00:07:59

I mean going back 5 years, data scientists used to build specific classification models for doing everything. And now we're in this regime of we need these really big models and we do in context learning and maybe even some fine tuning and we get them to do fairly specific discriminative tasks. But now you're saying we should almost go back to where we were 5 years ago and start building specialised models again. Only now, rather than building classification models, we're actually we're still using the transformers and the LLMs, but we're making them do specific tasks.

Randall Balestriero 00:08:29

Yeah, yeah, exactly. I think if you only want to solve a few specific tasks, use this prior knowledge to have a nice architecture, supervised dataset for that and just do that from scratch. This is something that's gonna probably work much better. But again, you need to make sure that the downstream application will never go too out of distribution. So that's why it really depends on the application and the type of use cases that you have. But I think at least here it shows that there exists some task where next token prediction is not the answer. And in fact, it's not just not the answer, but it's not better than random initialization, which is really sort of the worst case scenario.

Dr. Tim Scarfe 00:09:11

Interesting. I mean, from a fairness and bias point of view, lot of people say that, you know, large language models are bad in a way because there's the dominance of North American cultures and so on. Yeah. But you could also argue the converse, which is that the good thing about them is that they do have some awareness of value. You know, so we can fine tune them to have guardrails and to sort of say the right thing and so on. Is that harder to do with this approach?

Randall Balestriero 00:09:34

Yes, so here because you are in a fully supervised setting, you don't have as much flexibility to, let's say, change the behavior of your model or to take the form of supervised fine tuning. But because you don't have a generative capability, it certainly restricts the type of interaction you have with the model and how you can improve it. Because output is just okay, it a good or bad sentiment? It's not something that gives you a full answer that then you can try to argue against and

generate a fine tuning data set from. It's just okay, good, bad and that's it.

Dr. Tim Scarfe 00:10:09

Another thing is training strategies. So you know like the big players building these LLMs, they have lots of internalized knowledge around, you know, even the order in which you train the language models. Everything is important, you know, certainly in the old days of basic models, you just stick a load of data in there, no 1 really cares. Now do people need to be sort of thinking about the specialized knowledge, maybe thinking about curriculum learning and all of this kind of stuff?

Randall Balestriero 00:10:31

Yeah,

The Fair Language Model Paradox

Randall Balestriero 00:10:31

so this is a good point. So we did a paper recently called the Fair Language Model Paradox, where we show that when you do this next token prediction, because you have some tokens that are very low frequency, it's very hard to train on them and it takes a very long training. So it's very wasteful, right? And the problem is that because you do this next token prediction, you need to really capture all your distribution of tokens and so you spend a lot of time. But in this case, if the low frequency tokens are not useful to solve your task, you actually don't need to capture it at all. So in terms of training dynamics, this is actually a much simpler problem in many cases and what we found empirically is that we just out of the box use typical supervised training, we don't have to play with hyperparameter optimizer and you have very, very stable training. So that's 1 thing that could be interesting for future work is to see is this something that is easier to optimize and maybe that's why those like 7,000,000,000 parameter model can learn and not overfit on like 10,000 samples. And then it's also bringing other things that maybe this on its own could be a better initialization for next token prediction as well. So this is like very open up in the air, but maybe you could think of a simpler supervised objective that would be a better pre training solution that then you can use for next token prediction if you wanted to, but at least this would be a better starting point from random. So you'll almost reverse the trend.

Dr. Tim Scarfe 00:12:02

So we've spoken about 2 extremes. So on the 1 extreme we have pre training and you can like use it for any downstream task. And on the other extreme we have, you know, you start from scratch just with 1 task. Is there an intermediate solution? So what if I did this new approach, but for multi task, let's say for 5 tasks?

Randall Balestriero 00:12:17

Yeah, yeah. So that's a great question. So if you really think about it in the limit, you could formulate next token prediction as a multi task where you want to each task expecting is the next token this 1 or not. So in the extreme case, you could just recover a next token prediction on 1 end and on the other end you have what we have here, so just 1 task, very coarse, high level, predict if it's a good or bad sentiment or whatever. So in between you have a huge spectrum that you can exploit and if you can find, as you said, maybe 5 very different representative tasks, this should be enough to or could be enough to learn the representation that is as general as possible and then you can use this for maybe new tasks that come on the go. So I think the research question is how to design the minimum amount of tasks so that you have as diverse representation as possible. And of course we don't want to go to the extreme of just doing again next token prediction. But this is a very, very nice research question because if you have this spectrum and you can control where you want to be, then you can really have a per use case choice. So it's not, okay, you're always here or always here, tell me what you want to do, how much new tasks you expect your model to be exposed to, and I tell you where you need to be in this spectrum. So this could be very interesting as well.

Dr. Tim Scarfe 00:13:38

Very

Benchmarks, Generation, and Understanding

Dr. Tim Scarfe 00:13:39

cool, very cool. It does make me think though that these models understand through naive statistical alignment. And is it possible that the benchmarks we use just don't cap, you know, the gap of understanding that we've lost from moving from the pre trained models isn't being captured?

Randall Balestriero 00:13:55

Yeah, think because especially in the recent years we focus a lot on generative decoder only methods, all the evaluation and the type of objectives we put on ourselves is really about good generation, right? Even if you want to answer a question, need to generate a good explanation, you need to understand what are the intermediate steps. And I think the fact that we focus on generative models means that we completely bias the evolution and the way we approach this thing. And maybe you could have still knowledge that is learned without being able to generate anything. So I think this is also something that could be interesting to look at, or at least keep in mind when we explore those models.

Dr. Tim Scarfe 00:14:38

But philosophically though, isn't generation analogous to thinking in some sense? Models that generate, aren't they smarter in some deep way? Probably what you want to

Randall Balestriero 00:14:47

do is maybe imagine what could be, but I don't think you want to do generation with very granular details like next token generation. Because if you think about it, even just in terms of like classification task, you have a lot of different uncertainty depending on the token. If I start the sentence, okay, I saw this movie 4 minutes, there is no way you can tell what was the next token after 4, right? So this means that you know a period it would be like a time component, right? Maybe it's like 1 hour, 10 minutes, 2 hours, but you really need to be able to generate the, I don't know, 52 minutes or whatever the answer was to actually understand that I was seeing a movie, therefore I was staying in a place for at least more than 5 seconds, right? So I think token is way too granular and if you had maybe like concept tokens that's where you could start seeing okay this is meaningful because that's closer to maybe what we do. But right now we are very, very, very low level because tokenization is a lossless compression, right? So this is too close to the raw data. And yet we have the life easy compared to computer vision because already you work in language which is very compressed representation of knowledge, but still token is probably too low level still.

The Birth of Self-Supervised Learning

Dr. Tim Scarfe 00:16:04

Well, was a fascinating paper. Let's move on to your next 1. So the birth of self supervised learning as supervised theory, and that was with Yann LeCun. Yes. And basically you said that the observed differences between self supervised learning and supervised learning are not due to the loss function themselves, but rather the labeling of the dataset used in training. Give us the elevator pitch.[4]

Randall Balestriero 00:16:22

Yeah, so basically what we show in this paper is that you can have a supervised objective, like let's say least squares to make it simple. So you have the inputs, you have your network's prediction and you have the labels and you can turn this objective which tries to predict sample Xn to prediction Yn into a self supervised learning objective which tries to compare samples with each other. So basically you go from saying okay this image is a car or a dog to saying are those 2 images the same or not which is like the self supervised type of joint embedding world. And so you can show that if you have labels or you have knowledge of this pairwise relationship, they are actually learning the same representation up to some symmetries that is irrelevant if you do linear probing. So the loss function in itself, the SSL 1 or the supervised 1, try to do the same thing. They just operate on a different view of the labeling whether this image is that or are those 2 images or 2 samples representing the same thing. So given that, then the next question is how OCOM self supervised learning is able to generalize better than supervised. And from this perspective what you can say is that it's because it's as if they were solving a supervised task where the labels are not about predicting all the cars to cars but are very, very, very fine grained label where in the limit each image is its own class basically. So if you think about supervised learning in this extreme setting, you also don't overfit to the task because you don't collapse any image to another 1 and so theoretically speaking you can solve many downstream tasks as you want. So this equivalence of loss at least brings a slight new perspective on the fact that it's not really about the objective, it's more about how you design the SSL pipeline or you say okay, sample is related to this sample, but it's not the objective that makes you learn a better representation.

[4] The Birth of Self-Supervised Learning: A Supervised Theory — paper Paper with Yann LeCun proving mathematical equivalence between SSL and supervised learning losses

Dr. Tim Scarfe 00:18:16

Okay, and in the paper you were talking about how SSL can maximise the worst case downstream task performance. Can you sketch that?

Randall Balestriero 00:18:22

Yeah, so basically if you think about all the possible realisation of downstream tasks, you could have some very core scale ones. We have maybe different pictures of cars and buses and you just want to circuit the car or a bus, so no details need to be encoded to solve this. But then you can have downstream tasks where you want to say, okay, which brand of car is it or which color of car is it? So you have a distribution of downstream tasks, right? So the point now is that you want to learn the representation so that if you look at the distribution of downstream task performance, you are able to be as good as possible on most of them, right? So you don't want to be very good on some and then in the tail you are very bad on the majority of them. And so then from this you can try to say okay, what will be the labeling that tries to make your worst case as good as possible And from this you can say, okay, this is actually the labelling that self supervised is actually implicitly doing.

Dr. Tim Scarfe 00:19:14

How

Class Balance, VICReg, and Unifying Representation Learning

Dr. Tim Scarfe 00:19:14

does the class balance affect the difference in the losses?

Randall Balestriero 00:19:18

Oh yeah, so this is a very good point actually. In a follow-up paper we are doing right now, we show that current SSL objectives assume class balanceness. And this is something we already highlighted quickly in this SSL supervised learning as a uniform cluster prior paper we did a couple of years ago. And we show that current SSL objectives assume balanced representation of classes or concepts. And this means that if you train on ImageNet, things work out very well because concepts are sort of equally represented. But then if you go to other data sets like iNaturalist, which are very heavy tail, then you have a huge bias in your

representation. So until now, people do not really know how to solve this and so 1 way people approach this is through data curation and they say okay I'm just going to remove the oversampled concepts to try to make it more uniform and then I do self supervised learning on this. But because now we have this theoretical formulation and this equivalence of losses, we can use the exact same setting that people use in supervised learning to re weight depending on the frequency of classes. We can use that to come up with a new self supervised learning loss that takes this imbalance into account. So this type of thing is enabled from this mathematical formulation and it's principled. So the way we do this weighting, you can prove that it's the right way to do it from this supervised theory. And so this is really nice because suddenly from this seemingly naive connection, you can now come up with new generation of self supervised learning models where you can actually match what the real world data distribution is like. Non uniform distribution of classes, maybe even if you have some samples that are more noisy than others, you can include that information as part of the assessment objective as well. So suddenly you have a whole new world of possibilities that comes and because there is this connection you can actually prove okay this is the right way to do it at least from this supervised theory viewpoint. You also pointed out a connection to VCREG. Exactly, so basically what we do in the paper is that we show if you have a least square supervised type of objective and you turn it into a SSL 1, what you obtain is basically VCREG. So then you have a few variation, it could be VCREG or W-MSE depending on how you do this from supervised to SSL, But you can show that depending on the type of supervised loss you recover different type of SSL ones. If you look maybe more at cross entropy, supervised learning is going to be more like SimCLR type of loss. But you have this 1 to 1 correspondence. And this is also very nice because in supervised learning at least you know when 1 loss may be preferred compared to another 1. And this has been studied for a long time, right, because supervised learning has been around forever. And so now we can reuse those insights for self supervised learning. So this to me is also a very, very strong benefit of this thing is that suddenly all the theory and like the thousands of papers that have been done in supervised learning, we can just take it and apply it in SSL. Another example is a neural collapse, for example, that has been proven in supervised settings. Now it applies like in 5 lines in a SSL setting as well. So this connection is really beyond just trying to say okay, it's not the objective that makes SSL better, it's really tying those 2 huge communities together towards a goal where you have a single unified objective to learn representation. And this is nice too because if you speak to people they will think okay you have supervised learning on 1 side, SSL on the other side and basically you are either in 1 camp or the other. But now what we show is that actually SSL is pretty much everything in representation learning and supervised is just 1 realization of SSL. Then VCREAG without labels is another 1. Then this 1 is another 1. So you really have a better understanding of this relationship and

what representation learning is trying to do.[5][6]

Dr. Tim Scarfe 00:23:11

Galaxy brain question incoming. Could you combine SSL and super supervised objectives in some way to improve generalization?

Randall Balestriero 00:23:17

Yes, yes. So there is 1 paper which is supervised contrastive learning. So the way they do it is that they use the labels within a SimCLR framework to try to basically do fully supervised learning but with a SimCLR objective. So first of all, we can show that indeed this makes sense and that basically we can explain the empirical result that they had, but actually we can do a little bit more than that. So if you're in a semi supervised setting, for example, it may not be clear how to combine those 2 losses anymore. Or maybe you could say okay, have the 2 and I have a coefficient to weight them, but then you need to do cross validation and so on. But now from this perspective, you can combine them in a very principled way and you can understand which weighting makes sense depending on how much sample you have in 1 or the other. And you can use all the literature again from supervised learning for this setting as well. This is something you can do very easily with this formulation as well.

Dr. Tim Scarfe 00:24:12

Okay, so if SSL and supervise are 2 sides of the same coin, I mean of course we can use this theoretical framework to design new forms of SSL framework. Does it, you know, is the distinction relevant if they are the same thing?

Randall Balestriero 00:24:24

I think it's not just 2 sides of the same coin. SSL is more general than supervised Right. It's really SSL could be the more general objective to learn representation. The more prior knowledge you have, the more you know about your downstream tasks, more you know about your labels and then SSL slowly becomes supervised learning through the labels that you use for the SSL objective. But then because as you said, you have this hierarchy now, it does not really make sense to say you have either supervised learning or SSL. Rather what makes sense is to say okay, what is this relation matrix, what is pairwise matrix? If you build it from labels, it's supervised learning. If you build

[5] SimCLR: A Simple Framework for Contrastive Learning — paper SSL method recovered from cross-entropy supervised loss in the SSL-supervised equivalence

[6] Supervised Contrastive Learning — paper Combining supervised labels with SimCLR framework, explainable via the SSL-supervised theory

it from other a priori knowledge, for example 2 consecutive frames in a video basically of the same class, then you are more in an unsupervised SSL setting. But it's all about how do you build this per wise relation matrix? That's the main question.

No Location Left Behind: Fairness in Earth Models

Dr. Tim Scarfe 00:25:18

Very cool. Right, let's move on to our next paper. So No Location Left Behind: Measuring and Improving the Fairness of Implicit Representations for Earth Data. So there's loads and loads of modeling frameworks now that do these implicit neural representations of geospatial earth data. So things like climate modeling, resource allocation, environmental I was actually interviewing Johannes from NXAI yesterday. I don't know if you know him. He's working on similar stuff. The problem is you've studied this and you found that there's loads of biases and fairness problems.

Randall Balestriero 00:25:45

Yeah, exactly. So basically what we show is that when you want to model, for example, let's say temperature or precipitation to make it simple, and you want to learn for example implicit neural representation, it means that you want a model so that if you give a location and a date for example, it can predict what was the temperature there. So if you have this type of implicit neural representation it's very good because if you learn a nice model then you can actually interpolate those values, so maybe estimate what the temperature was in this part of the globe where you did not have a sensor. But you can also do extrapolation as well. If you assume you really learn the true physical model of the world, you could start saying okay, what the temperature will be 2 years from now, right? So this is very nice to have this type of model for all sorts of applications. The thing is that when you do this nowadays, depending on the architecture and the different design choices that you do, you will maybe have a very good prediction on average. So when you look at the rate performance or around the whole globe, but actually if you look for example around islands or coastal area, your prediction is going to be very bad, almost random. So this is something that can be very concerning because if you use this type of model to decide about a policy that will affect a specific island, using this model prediction is as good as using like random guesses. So it can be very detrimental and people need to be aware of those biases. So what we found is that for example for this type of climate data, islands often disregard that coastal area,

basically region where you have a big gradient in the type of data that you try to model.

Dr. Tim Scarfe 00:27:21

How much of a responsibility do modelers have to detect these kinds of biases in the data?

Randall Balestriero 00:27:27

So I think there is like 2 components as you said. So 1 could be that just the dynamic of the data you are trying to model is harder, near islands or maybe it's even unpredictable because you don't have enough observations to do that. So you have some uncertainties that probably you can never recover from good design. But still what we found here is that a lot of the biases now comes from the architecture and all you want to do to encode those positions, type of basis you use to do the prediction. So right now it seems that a big chunk of the bias come from the architecture but I totally agree that I don't think we can remove the bias entirely because there is maybe just different type of uncertainty at different parts of the planet as well.

Dr. Tim Scarfe 00:28:11

I mean, the world is a very, very complicated place. I mean, realistically, to what extent can we mathematically model it?

Randall Balestriero 00:28:16

Yeah, so that's a good question. So I think it depends on type of horizon that you have and the type of data that you want to model. If you have a system that is much more chaotic or can vary very quickly without much changes in the past observations, that's something that current models are having a very hard time with. If you want to predict something else, for example, temperature in North America, not near the coastal areas or really inland, maybe that's why you have less gradient dynamics, things are a bit more stationary, especially in through time, so then it can become much better. But I think at this point we don't have an architecture that is really able to understand that you have different physics, different dynamics models at different parts of the globe. And so because of this, just see what's the best on average and it means you miss out a lot of details.

Dr. Tim Scarfe 00:29:06

Can you tell us about some of the technical framework?

Randall Balestriero 00:29:08

So 1 thing we showed for example, at least for this type of GLOBE data representation, is that people use a Fourier basis to model the prediction. And this is something that is better than not using any basis at all, but what it means that you imply the type of signal you're predicting is very stationary and not localized at all. And this is a very strong prior, right? So this may be true for some things, but for other things like precipitation or temperature where you have localized very high gradients, then it's a strong bias. And if you come from signal processing community, you know very well that to have better localization you go from Fourier to wavelets and so that's 1 thing we did in this paper and we showed that using wavelet basis to encode those data allows you to have better localization and this removes some of the biases. And here it's more of a proof of concept that different design choices give you a different type of bias trade off. Wavelet is not the answer to everything, right? But I think the next step is to really be able to encode less and less a priori which basis to use and let the model learn from the data on its own. And we are not yet at this point, at least for this type of climate data.

Policy, Accountability, and Crowdsourced Data Bias

Dr. Tim Scarfe 00:30:24

How could it handle noisy or missing data?

Randall Balestriero 00:30:26

This depends really on the type of model you use. So for example, if you have INR, then you will not use the missing data as part of your training pipeline and that's 1 of the benefits of them. So if 1 of your sensors stopped recording during some years, you just don't use that as part of your training data because you really control where do you have the data and when you have it, what the prediction should be. So these

Dr. Tim Scarfe 00:30:48

earth models, they are now informing policy around the Who should we hold accountable? Mean is it the technology? Is it the scientists who design the models? Is it the policymakers who interpret the results?

Randall Balestriero 00:31:00

I think it's very hard for the person who designs the model to know clearly what it's going to be used for. So I think it's more downstream when you know clearly what you want to do with it. You should first set up a nice evaluation pipeline to make sure that it's something you can actually use to make those decisions and then you can report any type of failure modes you observe for people to improve on the design. But a priori, it's very hard to imagine what this model will be used for. So in an ideal setting, you wish that there would be no bias at all, but in practice, the world of possibilities being so large, it needs to be more of a feedback loop. And then iterate until you have something that you can really trust and then you can act on it.

Dr. Tim Scarfe 00:31:46

Earth modeling data is very anthropocentric, right? So you know, focus on human populations and so on. Should we also focus on, you know, like just ecosystems and places that have got nothing to do with humans?

Randall Balestriero 00:31:56

Oh yeah, that's a great question. And in fact, that's 1 of the big issue with a lot of the data set which is crowdsourced, because by definition the amount of data that you get is proportional to the number of users you have depending on the location. And this means you have a huge bias in what your model is learning and what your model is focusing on which means you miss out on a lot of things. So I think that's also 1 thing that okay crowdsourcing can give you a lot of data quickly, but it's very biased data. So then the question is how much of this biased data versus maybe paying a lot more and capturing other parts of the globe? How much of the tool you should have? And maybe you could be able to show that under some specific condition, just having 10% of the data which is high quality, uniformly sampled and then 90% which is crowdsourced, you can try to use those 10% to anchor your representation and then use all that data together. But there is a huge amount of research question in that because that's a very big source of bias.

Dr. Tim Scarfe 00:32:56

And a bit of a policy question, but we are using these things, you know, to do resource allocation, right? Know, so giving more resources to some populations might be taking it away from others. And then there's the fairness over time thing as well, which is what is fair like now might not be fair in 100 years' time. So how should we think about it?

Randall Balestriero 00:33:14

Well, yeah, that's a good question. I think this is also very application specific. So for example, if you want to predict where to build a house to solve some specific problem, Maybe you don't really mind having bad prediction where there is no population anyway because you're not going to build the house there. So in this case maybe the crowd sourcing type of data is actually good but this could really be dependent on the type of application. And just 1 thing I will say regarding the point you made before, this type of bias actually is something that you have in computer vision. So there is like a very nice paper done by Mark Ibrahim. Basically they showed that most of the data we have from ImageNet is from North America. So maybe you reach like 90% state of the art performance to predict for example type of chairs, cars, but only for North American models. And when you start looking at type of cars or chairs in Central Africa or East Asia, so there is a model performance is extremely bad. So this type of problem is something you have across modalities and that's something that's a very big big issue.[7]

Dr. Tim Scarfe 00:34:17

Randall, it's always a pleasure and an honor to have you on the show.

Randall Balestriero 00:34:20

Thank you so much. Likewise. Likewise. Thank you so much.

[7] Geographic bias in large visual models — paper ImageNet data is predominantly North American causing model performance degradation elsewhere

References

[1] **Is LLM Pre-Training by Next Token Prediction Worth the Cost?**
https://sslneurips24.github.io/
paper The central paper discussed in the first half: showing randomly initialized 7B models match pre-trained LLMs on supervised classification tasks

[2] **Tufa AI Labs**
https://tufalabs.ai
organization Episode sponsor: Swiss AI lab working on reasoning and test-time computation

[3] **Lottery Ticket Hypothesis**
https://arxiv.org/abs/1803.03635
paper Referenced during discussion of what randomly initialized models learn

[4] **The Fair Language Model Paradox**
https://arxiv.org/abs/2410.11985
paper Discussed around 10:31 showing that low-frequency tokens create wasteful training dynamics in next-token prediction

[5] **The Birth of Self-Supervised Learning: A Supervised Theory**
https://openreview.net/forum?id=NhYAjAAdQT
paper Paper with Yann LeCun proving mathematical equivalence between SSL and supervised learning losses

[6] **VICReg: Variance-Invariance-Covariance Regularization**
https://arxiv.org/abs/2105.04906
paper Recovered from the supervised-to-SSL equivalence when starting from least-squares supervised objective

[7] **SimCLR: A Simple Framework for Contrastive Learning**
https://arxiv.org/abs/2002.05709
paper SSL method recovered from cross-entropy supervised loss in the SSL-supervised equivalence

[8] **Supervised Contrastive Learning**
https://arxiv.org/abs/2004.11362
paper Combining supervised labels with SimCLR framework, explainable via the SSL-supervised theory

[9] **No Location Left Behind**
https://arxiv.org/abs/2502.06831
paper Third paper: geographic bias in implicit neural representations for climate data

[10] **Geographic bias in large visual models**
https://arxiv.org/abs/2304.12210
paper ImageNet data is predominantly North American causing model performance degradation elsewhere