**Björn:** 0:00

Our perception is a much more active process. A camera just takes these individual pixels and processes that to a certain a part, but the camera is not really seeing something, at most perceiving individual pixels. It's not seeing objects in there, it does not care about them and it does not get that together. The brain does so and it is, as I would argue, an active process that's happening there. It's not a perception just coming over you without your brain having done that.

**Craig:** 0:27

Hi, I'm Craig Smith and this is Eye on AI. In this episode, we sit down with Björn Ommer, a visionary in the field of AI, to discuss the advancements and implications of generative models in technology and society. Björn gives us a fascinating look into the inner workings of diffusion models, the importance of creating AI that operates on consumer hardware, and the significant strides made by his team's creation, stable diffusion. As we tackle the complexities of AI's role in the model world and the steps toward democratizing AI technology, Björn shares his expert views on the balance between open-source innovation and proprietary development. I hope you find this conversation as captivating as I did. Why don't you start by introducing yourself, where you got your education, how you got into AI, and then we'll start talking about, obviously, stable diffusion and then everything that's going on.

**Björn:** 1:38

My name is Björn Ommer. I'm holding a chair for AI at the University of Munich for two and a half years now. Beforehand, I had my chair, my professorship, at University of Heidelberg. Before that, I did my postdoc in Berkeley in California. I got my education, my PhD, from ETH Zurich in Switzerland before that.

**Björn:** 2:00

Now, how did I get into AI? I was looking for challenges, for open frontiers that are out there. I very early on figured out that probably people have been traveling to the moon, been to the deep ocean and everywhere else on our planet. But the biggest frontier that I thought we are still having and I love to explore open frontiers is actually our mind, our brain, understanding things and being able to comprehend the world around it, which is like super complicated. Out of all the ways that you could think of our brains making sense of the world, it was really our eyes, vision being the one that I found most attractive in order to study that because of that being such a complicated window out to the world sort of capacity that we have in making sense of what arrives there at our eyes.

**Craig:** 3:03

And was that? You're pretty young? So was that before convolutional neural nets?

**Björn:** 3:08

That was way before deep learning. Yeah, so I've seen the world beforehand, I've seen it afterwards and guess what? When I'm teaching classes, it's always like these days, students totally get hung up on deep learning, but I try to hear, or they're also tell them lessons from what we had beforehand, because it's not all outdated, that we did what we did before.

**Craig:** 3:28

Yeah, and then how did you move into the stable diffusion project?

**Björn:** 3:34

Yeah, so we've for a long time done retrieval tasks, address that, recognition tasks and so on, and that has quite a bit of value to it. But eventually it comes to this point where you feel like, how are we making progress there? Right, I mean, eventually you might find pixels that are correlated with class labels in your data sets. And this only goes so far. And I was wondering, like, isn't there another way that we could actually figure out whether we're making progress or not? Isn't there a way that I could more carefully diagnose whether the model has actually learned something about the world or not? And to that extent generative models actually became super appealing to me and probably also to a lot of other people in the community, because there you can turn that process around right. You can say, like, with a discriminative model, a model that's trained on recognizing a particular object, you can say like, yeah, recognize all dogs and separate them from cats, but it might be that the model actually figures out that a bone is much more indicative and easier to learn to figure out what a dog is holding a bone in its mouth than anything else. But if you then turn the process around to do what a generative model does essentially say, like, show me what you think a dog is. An outcome only bones. Then you know like, yeah, probably went the wrong direction. And that's how I am.

**Björn:** 4:53

My lab get really excited about generative models. Now it turns out that it's a challenging task being able to learn the visual world out there and being able to reproduce what we see out there Just in a pixel space. We explored several different models and you very easily sort of run in the standard dilemmas that we're having these days you, on the one hand side, want to have, like, great quality and, on the other, you want to capture the world in all of its breadth and diversity. That is there, and quite recently, novel model architectures came about transformers that people might have heard of, or, even more recently, diffusion models, and these promised actually to bridge this, the wide, this gap that we had between the quality and the diversity. And it turns out, yeah, that is to a certain extent true, but at the same time, I saw that new issues actually arrived. These models get more and more complicated, consume more and more compute and needed more and more training data. So we were wondering, like, where will this eventually lead us? Aren't we moving in a direction, and that's what it seemed actually before. We were working on stable diffusion and we're moving in a direction where, in the very near future, only big companies would have the computational resources not just to train these models but even to take the models that have been trained and apply them. And so it turned out right. I mean, there were these visual synthesis models and that goes beyond like to other disciplines as well which were so big that you needed supercomputers to just run these models, not even trained, but to just run these models.

**Björn:** 6:38

And I was concerned about that to a certain extent, because if you really consider this generative AI becoming what I believe, it now has turned out to be a critical technology which becomes really the foundation and that's what we call it to a certain extent also foundation models, a foundation for everything else, not just the research that we're here doing in our fields, but everything else where we create values essentially in our societies. And then you get sort of tied to just very few companies that can actually lay these grounding for this new technology. That comes with a lot of risks, right. I mean, on the one hand side you could say like sure, we need creativity, like the more clever minds the development of future technologies built on, the faster it goes and the more diversity outcome also would be.

**Björn:** 7:31

That was one concern, but there are others.

**Björn:** 7:33

Right, if it's a critical technology for our society and it's eventually held in the hand that the future development is in the hands of just like a few companies that you have out there, that has actually profound implications, let alone like privacy issues and so on, if other businesses have to run their precious data to just those companies.

**Björn:** 7:53

And so we were wondering how can we actually democratize the development of future generative AI, and it turned out. For us, it's key that this technology is not just powerful but at the same time it's also widely accessible, and that means not just open source, which I believe is crucial for foundational models because they are this critical commodity, but third, it also means this technology needs to be really accessible to average users, average users who don't have a supercomputer in their backyard. So that meant for us we targeted consumer hardware $300, $400 of a standard GPU that you could run it on stability fusion these days runs on your mobile phone even, and with that you open it up to really a broader set of researchers and to a broader set of entrepreneurs that can use this technology now, if you don't just need a supercomputer.

**Craig:** 8:49

When you were looking at generative AI, was that after the transformer algorithm appeared or before?

**Björn:** 8:57

So we were looking at that beforehand. But sure, the transformers changed things quite a bit and we were working with technology before transformers variational autoencoders. We had this paper we called it Variational Unit which tried to disentangle appearance from posture so that you can animate human beings, change their gate and their posture as they're walking without changing their appearance at the same time. So that you want to disentangle prior to, like transformers, really keeping getting traction in the vision community. But then we continued and adapted our approaches as new technology really came about and the latest was evidently then diffusion models which led to what people know now as a diffusion and diffusion models for listeners.

**Craig:** 9:45

Can you give layman's description of what a diffusion model is?

**Björn:** 9:49

Sure, it's actually not too complicated. So one key foundation I should mention there, and that is beforehand. I mentioned discriminative AI and discriminative models. That meant you show the computer a lot of images and say, like cat, cat, cat, dog, dog, dog, and the model actually figures that out. What we're these days doing in generative AI is mostly what we call self-supervised learning. That means we show the computer just images. Now how would the computer start to make sense if you just presented images and don't say what's there or what structure is there For that? Utilize self-supervision.

**Björn:** 10:23

So in a nutshell, in diffusion models you give the computer a training image. You add noise to this training image just a tiny bit, so that you and I would not even notice the difference. But you repeat this process of adding noise hundreds or even thousands of times so that the end result looks like you pluck the cable from your TV set Pure noise, nothing left. You do that so that you can turn that process around, present the computer a noisy image and say like, hey, this tiny bit of noise you can make up for right and sure the computer in that case in your network an auto encoder, as we call it can do that. Then you start with your pure noise that you have and eventually you end up not exactly with the image that we started with, with a clear image, but with something which is conceptually very similar to understand, to represent, I should say, the distribution of all the images there out there.

**Craig:** 11:16

Then you connected to a language model so that there's a conversational interface, a natural language interface or

**Björn:** 11:25

Correct. So diffusion models itself don't need that right. They are, as we call them, unconditional models, or they can be unconditional models that just learn the distribution of all the images, like set up a space of how images are situated with respect to another, and this is important just to stress this point there. Think of your training images as being tiny islands in the Pacific with lots of water around them. And what the diffusion model is? It learns a space so that islands which are somewhat similar get closer together. So you essentially start building bridges across the water linking those islands together, so that you can interpolate, that, you can hallucinate, so to speak, what's between two islands where so far has just been water. You can essentially recreate additional islands that haven't been there, additional landscape that hasn't been there, which kind of mixes or blends together the islands that you had in your training data there.

**Björn:** 12:17

But a user typically wants to have control and not just synthesize arbitrary images, right, and that's where textual site information or other site information comes into the game. So you take this diffusion process and say, like, as you're denoising, I tell you it should be an image of a cat, for instance, right and a brownish cat or whatever else, and then the model learns to incorporate the site information as well. Now, in what we were setting up there, we used the transformer architecture to incorporate the textual site information there. But it's not just that. You're limited to textual information. You can also utilize images and say like hey, create an image, but it should look somewhat similar to this other image that I provide you, and then stylize that image. For instance, right, you said like I take a photo from you but render it in the style of a certain artist, and then the image is changed and recreated, so to speak. This photography of yours is recreated with this other concepts in mind.

**Craig:** 13:16

I mean, there is style transfer, but that's not generative.

**Björn:** 13:20

We can do style transfer with generative models and actually most of the style transfer these days is actually done with generative models or, as the public calls it, with generative AI, because, again, like we are generating new data there and for that we have learned in a generative manner the distribution of all the images of a particular style.

**Craig:** 13:40

What are the challenges of making this available on, as you said, consumer hardware?

**Björn:** 13:47

The biggest challenge is that you want to learn a model which fits on consumer hardware, has limited memory. In that case, for those that are more versatile there, we aimed at something that has 10 gigabytes of memory these days works on mobile phones with 2 gigabytes or less but at the same time you want to have a model that kind of captures all the visual world that is out there, and for that you need to present it quite a few images, billions of images in our case, hundreds of terabytes. So how do you get hundreds of terabytes fit into 10 gigabytes, or even take essentially the entire internet out there and make it fit in your mobile phone so that you can take it with you? This requires a lot of abstraction. I believe that is the core of intelligence right Taking lots of data and abstracting it so that you capture the essence of the reality as it's depicted in your training data.

**Craig:** 14:47

You talked about the act of nature of perception, but also differentiating, like what does the model focus on in the scene? Can you talk about that a little bit?

**Björn:** 14:57

Yeah, so there are different takes on how perception actually works about, and psychophysics has been dealing with this for a century and more. What I find quite appealing is this understanding of the brain really constructing a model of the reality as it's presented to our eyes, and this is an old sort of take on what we have there Hammond von Helmholtz more than 100 years ago, gestaltis 100 years ago. We're already considering that to a certain point, that our brains are actually that that perception is an act. It's not a reaction to something. It's an achievement actually for us to create this impression of what we're seeing there and it's not just something that totally passively comes over us. I know that this is hard to conceive of because we are constantly seeing and apparently it happens so effortless and we see this with Asher rendings every now and then that apparently there's more to that and that there is actually a cognitive process going on and on a day to day basis. We're not wondering about them seeing this, apart from you being a vision researcher like myself and then you get hung up about some of those details and you think about like a super complicated presenting being presented to your eyes. You have this process like where do you actually allot your attention? What do you focus on to actually make sense? And we we can. Most people probably don't even know this. We're not like a computer camera. We're a tiny, foveal area of tiny area where we can only read letters where we can see sharp and the rest we don't even see. That's why we constantly are doing saccades. We are focusing our gaze at certain parts, and the same happens with our attentive process as well. It's only that this happens so quickly, in a very intriguing manner, that we paste all of these tiny crop outs of reality together that we get this impression that we see Everything sharp and all with the same detail at the same time. Now the question is, what of that should we carry over to the computer? Because it turns out it is a complicated problem, like.

**Björn:** 17:05

One of the most complex problems that I find very appealing to also research on is, for instance, the binding problem, as you have millions of pixels in the image and it's not that it's just like individual grayscale or colors that you actually perceive, and they make up trees and buildings and so on. Right, I mean, a building is more than just individual colors that you see there, it's structure. It's a very complicated structure that has actually emerged. That requires to bind together information from faraway spots in your scene that's presented, to then assemble the structure that we actually see. In my talk I was showing this example of a flock of birds. They are flying in a triangular shape in the sky. How do you see this triangular shape? None of the individual birds has something triangular on them. It only happens when you bind the individual birds, which are individuals. I can hear they're separate and the next day they found they're flying in opposite directions or build a square shape in the sky or whatever else that is. So this emerges out of this ensemble of birds that are coming together and your brain is only able to actually see the triangular shape by binding together these faraway spots, and that requires attentive processes, pre-attentive processes linking that together. So the question for us and a long story short as vision researchers was always how can we actually make this feasible computationally feasible also, because you cannot try out all combinations. Apparently, the brain also is not doing that. You have rapid feed-forward process that are able to mitigate this impression of there being a triangular structure and, with attention, we now have concepts, for instance, that go in that particular direction but which also don't scale to arbitrary many pixels.

**Björn:** 18:56

The kicker here was when I want to represent scenes.

**Björn:** 18:59

There's a lot of local detail, texture, colors and so on that you want to separate from representing this triangular shape.

**Björn:** 19:09

For instance, of the birds that I have there, the feathers that each individual bird has.

**Björn:** 19:15

You don't care where each individual bit and piece of the feathers were pointed right, you just care about the overall color that this bird had, like how feathery or whatever else. It was a bit of the texture, but the rest I wouldn't even notice if a single pixel is off. Now, if I were taking something like a diffusion model that we talked about beforehand, which is great in capturing these long-range interactions, the long-range context of a scene, if I were to take this model and apply it to represent the local texture, it would just get hung up on all of those local details. It turns out that other architectures convolutional neural networks that we had before and a great perceptual compressors to abstract away all of those details. So it turned out that actually a combination of these two architectures is the best way to go about One that compresses and captures the local detail standard convolutional architecture and then taking a more modern, like long-range contextual representations represented by a diffusion model and combining these two together and then you have stable diffusion in essence.

**Craig:** 20:22

Because when a camera sees a scene, even if at some level it's delineating objects, it doesn't everything is equal.

**Björn:** 20:31

I guess that's the fallacy. I don't consider our brain as this entire seeing apparatus, together with the eyes, as being a camera. Right, as I said, our perception is a much more active process. A camera just takes these individual pixels and processes that to a certain part, but a camera is not really seeing something. It's at most perceiving individual pixels, but it's not seeing objects in there, it does not care about them and it does not get that together. The brain does so and it is, as I would argue, an active process that's happening there. It's not a perception just coming over you without your brain having done anything to that

**Craig:** 21:13

Right. So what you're talking about is the computer system, whether it's stable diffusion or convolutional neural nets. They're creating a representation of the scene and then in that representation space, it's much easier to determine the focus piece.

**Björn:** 21:32

Yeah, so when you want to be able to synthesize images, for instance, you need to eventually go through a bottleneck, have a very compact representation. Other than that, you would never be able to get these billions of images, sort of summarize and be able to synthesize something new from that. This hazard that this compressed space needs to have certain regularities, ways in which you can present it, a bunch of training images, and it can then generalize I guess that's the key word here Generalize to novel data that it hasn't seen in the training phase, which is somewhat off from the training data, but in a way that we would probably not even notice as human beings. And that's what you want to teach your computer, because you will always present it just a finite amount of training data and you want it to generalize to new data grids. So the island, this water between the islands we want to create new islands in between.

**Craig:** 22:27

Yeah, and that mechanism is in stable diffusion. Is it the attention mechanism that's used in the transformer?

**Björn:** 22:38

So first off, I would say it's the diffusion process itself, the attention process. So diffusion models is a bit tricky. Now we're getting into details. It's a bit tricky calling diffusion models itself a model, because it's just a diffusion process there, as a training paradigm, so to speak. The question is really how you incarnate that, and for us it was this encoder-decoder architecture. Something that takes a representation has to go through a bottleneck, so you get the compression and then deflate it again to get in the same dimensionality as your input there. That itself goes a certain way.

**Björn:** 23:13

But now you start talking about blends of architectures. It meant for us that we not just have this but in addition have what's called self-attention, for instance, to put this encoder-decoder architecture, so to speak, on steroids. Then you want to have site information like text coming in. That meant for us incorporating cross-attention, a concept by means of which you can take textual tokens, as we say, a textual representation and, as we're doing, denoising, then control the denoising process happening in this encoder-decoder architecture. So you really see that it's not a one-size-fits-all approach. These days we would say, hey, I have a single architecture and to end, throw the data in and then hope for the best. But it's actually that we try to utilize the best of different worlds there and bring that together. So make up for the weaknesses that one architecture has by the strength of another architecture.

**Craig:** 24:09

You've chosen to make this open source.

**Björn:** 24:13

As I said before, and I consider this to be crucial technology for our society. One thing that is foundational, and one of the limitations that we had beforehand was predominantly that these models did not run on consumer hardware, and with that, only a few companies were limited to that. The goal was to lay something like Linux or other open platforms that as many people as possible can utilize, and open source was evidently another crucial part in coming up or realising this vision.

**Craig:** 24:44

I've talked to a lot of people about open source and that's what we started talking about this dilemma between the concentration of compute in the private sector and how that limits other researchers or smaller companies from developing similar products. And so Meta has chosen to open source its MAMMUM series. You guys are you, I don't know who have open source stable diffusion. There are still more powerful models in private, proprietary models, in different corporate, under different corporate ownership. One theory that I've had I've talked to people about this open source proprietary debate ultimately open source financial resources or to pay for the compute, and so the private sector is always going to be ahead and always going to have models that are productised that are superior to open source. But then there are other people that argue open source ultimately wins. So I'd like to hear what you have to say about that.

**Björn:** 26:09

We'll see who wins and I hope that society overall is winning. And we're developing technology. I mean it's fun to be working on that and so on, but I guess our field has stepped into an arena where what we're creating starts to matter, starts to have implications. Now I believe society has most benefit from that. It turns out that our European Commission also starts to think in the same manner. You've probably heard open source models there play a special role in the regulation, are exempt from certain regulations for that very reason.

**Björn:** 26:44

We see private companies now also going in the direction of open source. We see that open sourcing things seem to be also beneficial to hear their, attract VC money to create and create attention of people, which these days is fairly important factor, especially when you are a startup and want to acquire money. And we have with a French company, seen that open sourcing there can help quite a bit with Germany, german companies as well. Rather than doing this all under the rug, totally close source for many of the lab, I see open source for that reason being also attractive. Now there come other things. These models have great costs to them, computational costs, for instance, if several entities could share in the developmental cost and with that also not just like the economic costs, but also costs like this CO2 footprint and so on, like why have several labs do the same thing again and again and then pollute the environment with that? If we could share in that, create larger sort of utilization of the value that we have there.

**Björn:** 27:56

All of that you see more with open source models and that's why I see that it's, I would argue, more beneficial for societies. It is also a way to create more revenue for the companies to a certain extent, and it's just a matter of time essentially and your mode is not an infinite mode and that's just a question whether it's days, weeks, month or whatever else it is, but eventually others will come to that point as well Is that worth investing billions of dollars for a fairly short amount of time? And that we need to ask ourselves potentially, even with partial governmental support to startups and so on, it's probably advisable to then give also something back to society. And then there are other things like risk mitigation and so on. If these models become critical for society, you want probably lots of researchers testing them, red teaming there, for instance, which becomes all harder if it's just a close source model.

**Craig:** 28:53

Yeah, yeah. Do you think that governments will try and tip the scales toward open source? I mean again, the open AI and the barclays, I think woke everybody up to the danger of not having some oversight. I mean, they're all good people.

**Björn:** 29:13

Just imagine this is the crucial technology that in the future we will build our further platforms on, and then you're dependent on this one company and whatever they might even have all good intentions but then something weird happens inside the company. And what you then do? If you're a startup or a company or even a government, and you build everything that you have based on that Like in earlier days you used operating systems from some well-known companies and you buy it and it works pretty much itself. Yeah, you're there. You unfortunately need patches, but you're good to go for quite a while. But these days you would need the servers, and if tomorrow they're shut down, then essentially all your business ain't going to work anymore. So I guess that woke up quite a few people being after last week's developments in the European Union and that seems that at least there is this clear talking about open source models in that case. But we'll still need to see how this eventually turns out, because that will have tremendous implications.

**Craig:** 30:16

Yeah. So going back to the computer, I mean governments can play a role in making compute available to a wider audience Like.

**Björn:** 30:26

If we actually consider these models to be critical for what's going to happen in our societies there, then compute is the critical commodity to make this happen. Much like electricity, water and so on in the past, it's in the future going to be compute of the kind that we're needing it there and with that being foundational, governments probably should also have a play in that case. I'm very much for private and public partnerships in actually supporting that. There's a lot of small companies that would just not be able to bring up the money, let alone like the people that could actually host such infrastructure, and they would probably be happy to just buy in there and then sort of have the running cost being for the system then being taken over. So I see their win-win situation for everybody.

**Craig:** 31:18

Yeah, every country is going to approach it differently. I know there's a lot of talk in the US about creating a cloud credit system or something to give researchers or small companies access to GPUs, and I imagine that's happening in Europe. It reminds me a little bit of how different countries have approached the internet, and there are smaller countries that have free Wi-Fi as a way of generating economic activity. Do you think that kind of thing will happen, that some countries will form compute resources, whether it's on a credit system or even building national server farms to supply compute to the private sector not only private sector, but to the research community?

**Björn:** 32:13

Progress in our field of generative AI has happened super fast over the last year and I guess politics with that have just started to learn what it means to have this technology around, this technology being so rapidly developing. I've been at Wisinghewer there and been asked by politicians to set the stage, explain everything that's going on, and from that I can't just tell that they started to understand the necessity in that case of this technology and the commodities that we need to further develop this technology. And my understanding is that all Western countries and politicians there until now have finally sort of got their head around like and this technology is there, it's probably also going to stay and there are certain sort of responsibilities that probably also central governments have in putting that together. But I see in the Western world at least striving towards that direction Now in the European Union, for instance, there's discussions about that, but to what extent this will be realized on a cross European level we'll still need to see.

**Björn:** 33:26

I mean, for me it's like the Airbus is a great airplane company or they have built great airplanes but it took quite some years for them to come off the ground right, and I doubt that we have this time in generative AI here in the US I'm sure I'm there. This, this the story is a little bit different. At the same time, there is also economic constraints in there, and for that very reason I was sort of opting for private public partnerships, because companies are making revenue from this technology or these resources being there right away, constantly, sort of adding new hardware to that. So maintaining I mean not just like electricity, but also investing in new GPUs there, but also help with just getting the startup of that that, these server farms that we need for that being funded, and the same, I would hope, also in European countries. I hope that politicians are more and more sort of seeing the responsibility that they have with that, but we also need to give them that they've just learned what generative AI is within a few months.

**Craig:** 34:30

It's it. Things have moved so quickly. I've worked for a couple of years on this National Security Commission on AI, and that's the mission is to educate, in that case, the National Security Community, but more broadly, the government. There was a lot of optimism and hope in the recommendations that were being written, but the follow through has been painfully slow. So the other flip side of that is to make models that require less compute. How do you go about doing that?

**Björn:** 35:06

So there are different takes.

**Björn:** 35:07

One was what we've been doing with stability fusion create models which, right from the bed, consume less hardware resources that we have there, and I guess that's an advisable direction, especially when you set up a new foundational model.

**Björn:** 35:21

The other side is taking models which are already existing and making them more lightweight.

**Björn:** 35:28

Standard approaches for that are distillation type of approaches.

**Björn:** 35:31

That, however, is a costly process and we've seen a lot of startups who are now waving with hey, we have this super fast model and so on, but this itself, in a lot of situations, can take as much resources as training a model from scratch almost. So you train a large model and then you train a small model and essentially, like, have tremendous sort of extra costs, what I find super repeating. There is then the third stage, which is take models which are already existing, potentially even smaller models, and empower them, make them better. We are having, for instance, ongoing research there where we have published like first bits and pieces of where we take a smallish model which can also be an older model there and increase its resolution. Yeah, turn a model which is just giving a low res output images and turn that into a megapixel size image there, without requiring much extra compute during inference and without any costly training stage there, by embracing more novel architectures, like flow matching approaches, for instance, which turn out to be complimentary to what we've been doing in diffusion models.

**Craig:** 36:39

Yeah, can you explain?

**Björn:** 36:40

diffusion models I mentioned before and how the training goes. Now there's a different understanding, like what's actually happened with a diffusion model, and that is that the diffusion model actually describes a trajectory through image space. I was talking about these islands in the Pacific. Now there's been recent developments, flow matching, for instance, optimal transport approaches which straighten these trajectories. With that you get less diversity but you get much faster inference and with that you see that the two are complementary to another. One gives you diversity but takes quite a bit. The other is fast but less diversity. So we've been taking a standard diffusion model out of the box without any distillation, any of these costly processes, and then take the output of this model, or not the output In the latin space that we had in stable diffusion, and add a flow matching approach which takes the lowest and matches that In an optimal transport fashion, so as directly as possible things trade trajectories over to the higher s output, latin, which then our order encoder that I mentioned before and can turn into pics. I'm. That turns out to be a super effective approach without requiring costly distillation.

**Björn:** 37:52

So One of my key takes there is that intelligence, at the end of the day, arrives when you have to solve complicated problems with finite resources. That's what our brains are actually doing. We use the same brains as our ancestors use. They just had sticks and stones and we're dealing with complicated machines there, and this finite hardware there required us to come up with intelligence, intelligent ways of actually dealing with problems, and I guess, as a research lab without infinite compute, you also need to go for these intelligent solutions there, rather than taking what you have and just put that on steroids by investing more into gps. I guess that I see from a research perspective as something very appealing and that I see also from the application site for users as something that is appealing, and I guess stable diffusion to a certain point Shows that there is something to this theory that I just outlined.

**Craig:** 38:52

Oh, I wanted to ask about world models and, and a lot of what you've been talking about sounds like world models where you're getting input directly from images, for example, presumably video, that seems to be coming very quickly as opposed to through the filter of language. Greater grounding in reality than than the at least the original LLMs amazing we can talk about the original LM's, so only a few years now but that that we're purely synthesizing the knowledge that's contained in language. Is it related? What your research to world models?

**Björn:** 39:36

I find constructivism very appealing, constructivism in the sense of pg, our brains, creating a mental model of the outside world, and being kind of this active scientist that actively explore the world out there, doing tiny little experiments and then making predictions, essentially in the in the sense of predictive coding, and then we probe and see does that work? Does my prediction about the next frame In a video sequence work out or not? Than I need to adjust things, and so on. And then I find very appealing. I guess we need to be humble there. I guess that is also one of the lessons that I've learned from two decades, or whatever, of vision research.

**Björn:** 40:11

There these problems are super challenging that we're dealing with, making sense of the world by just observing a bunch of pixels, is a challenging problem. It is every now and then that you're granted this opportunity to move on a tiny mountaintop and see what's before you, but you see the next rich of mountains. You cannot see behind that. And with that comes that, yeah, if you look back, what gans, for instance, jantive, adversarial networks we're doing all that looks like these humble models and almost not doing anything compared to what we're having Now. If you look one year back like what we did in stable diffusion looks like tiny baby steps compared to the nice results that we're having these days, and the same will probably be true in a year and two from now with video and the like.

**Björn:** 41:01

But saying that we will have a full up model of our world, which essentially I was alluding to visual understanding in my talk there as well in this is a goal, but we will only achieve that in part, and I would be careful to sort of go too far there, cuz I am. That's one lesson that we've learned from development in your networks in a winters and so on. We came up with these promises there hey, I have one year and you give me a few billion more and I will eventually have the brain, and then we figure out structures probably a bit more complicated. It's not just like Linearly adding up there and the same host true for many of the developments that we're having these days. So I'm by no means concerned that I will run out of research challenges in the upcoming years title.

**Craig:** 41:50

Title of your talk was something like the fallacy of scaling, or the big companies that are in this are gonna continue to scale because the scaling laws are holding, so far they appear to, but but

**Björn:** 42:09

I would say, like in 2007, what we call dinner scaling has ended. So this integration, that rigid five men like six decades ago, had predicted that you can make transistors smaller and smaller and with that they reduce, they need less power and their frequency goes up and so on. And that has come to a certain end. I'm the integration. I guess every one your audience can see that the processors that they're buying these days don't have many more sort of tech, that their frequency than they had like five years ago. Whatever else compare to what we had like 1020 years ago where every Year you were seen increased air. Then we had parallelization with GPUs and so on, and that what deep learning has benefited from quite a bit.

**Björn:** 42:50

But this is by no means as quickly ramping up the GPUs in their performance compared to what the demand site, our models in a I actually would be able to actually take in. And with that come limitations in the scaling. It's like the canary bird and when you go under the ground and a coal mine or whatever else, it's first the smaller Companies, research labs and so on noticing in. These days it's even the big companies there noticing that. If you look at GPT for after first being released linearly better, but their parts where it's even gotten worse. Even with Large donations or contributions, I should say, by a big company, it wasn't that wasn't infinite amount of money and that wasn't infinite compute that they had. We see that even with these enormous amounts of money you cannot just like, throw it at it and it's constantly getting better. There are trade offs involved in that development and with even that big companies with that much resources.

**Björn:** 43:52

Noting it, the writing is on the wall that just skating. We will see skating up. This is important, don't get me wrong on that. But this is not the only way to go about.

**Björn:** 44:01

We will see classical engineering scaling things up happening, but I think the scaling up Is the single most consistent factor if you go back half a century or whatever else, which which correlates with a lot of developments there, but with that it overshadows important technological sort of breakthroughs and paradigm shifts that we had at the same time. It's not that the GPUs that we're having these days, I just electron tubes or relays from hundred years ago on steroids, made smaller and faster. We had developments like the transistor, like the integrated circuit afterwards, and only they were able to actually get this there. I guess the same thing that we're seeing as well. We have not just the mps that we had decades ago, and the same way we will have different architectures as well. So taking something that was there and just getting up, just get to a certain way. We will have these paradigm shifts, hopefully in the very near future as well, because there are challenging problems which I believe we cannot only solve the means of scaling that's it for this episode.

**Craig:** 45:13

That’s it for the episode. I want to thank Björn for his time. If you want to learn more about today's conversation, you can find a transcript of the episode on our website. Eye on AI, E, Y, E, hyphen, O N and dot A I. And remember the singularity may not be near, but a AI is changing our world, so pay attention.