Thoughts on the Present State of AI
Bisan lecture, April 10, 2024

It feels like an utterly bizarre and incongruent thing to give a lecture on AI to a Palestinian audience when Gaza is suffering beyond words, so many are dying. I am going to read a statement that I have posted on my blog first.

- OK, I agreed to talk about AI, so here goes.

Statement:

I had first hand experience of the Israeli/Palestinian conflict for the first time in 2004. I experienced the road blocks, the wall, the growth of settlements, hearing other's nightmare stories and interacting myself with the IDF. As a Professor, I was shocked that Israeli laws on “Areas” prevented Israelis from collaborating with Palestinians in the “West Bank”. I was not allowed to go to Gaza or experience first-hand this open-air prison with over 2 million inmates. It is not surprising to me that Hamas took charge there and now has struck back. If someone grabs your throat, you strike back. A person reaps what they have sowed. But the full force of the Israeli electorate's total depersonalization of every living Palestinian only emerged in the present war. Hitler taught the Jews the weapon of starvation in the Warsaw ghetto and they have copied his policy starving the Gazans. I have hundreds of Jewish friends who may now look at me as anti-semitic, but this second *nakba* compels me to say that in my book Israel has become a pariah state.

- AI is a hot topic in the midst of major successes that everyone is discussing. It is also an area that can be studied anywhere: until recently, all research was publicly available at arXiv.org, software to use GPUs and Python code on PCs or Macs can be purchased. Although huge data was needed for the so-called “large language models”, many experiments can be done with small data sets, hence I believe it is a suitable topic for independent researchers.

A useful article attached to my slides

The Surprising Rise of "Tiny AI"

How Small Generative Models Like H2O-Danube-1.8B Are Democratizing AI

Frederik Bussler

Published in Predict, Feb 12, 2024

Dubbed "tiny AI," compact yet powerful models like H2O-Danube-1.8B, Vicuna, Koala, Alpaca, and TinyLlama are bringing advanced generative abilities to the masses. Requiring modest computational resources, these small LLMs are reshaping the AI landscape by making AI more inclusive, innovative, and impactful.

1. The sizes of representations: representing thoughts with very high dimensional vectors as well as low

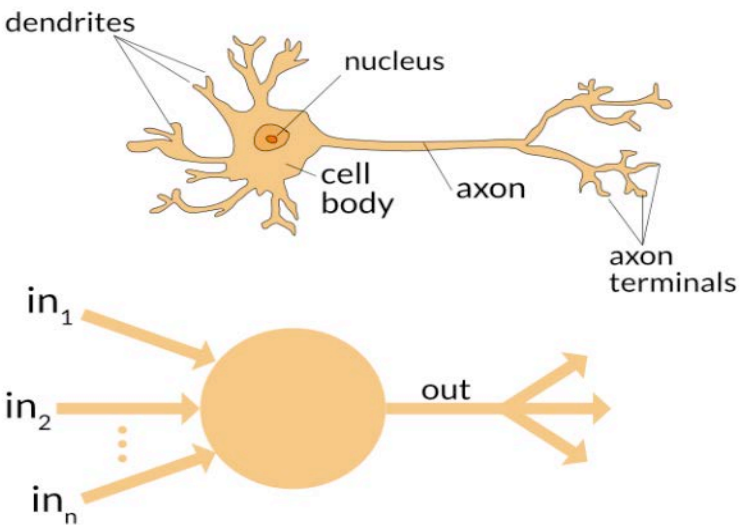
2. *Transformers*, and grammars

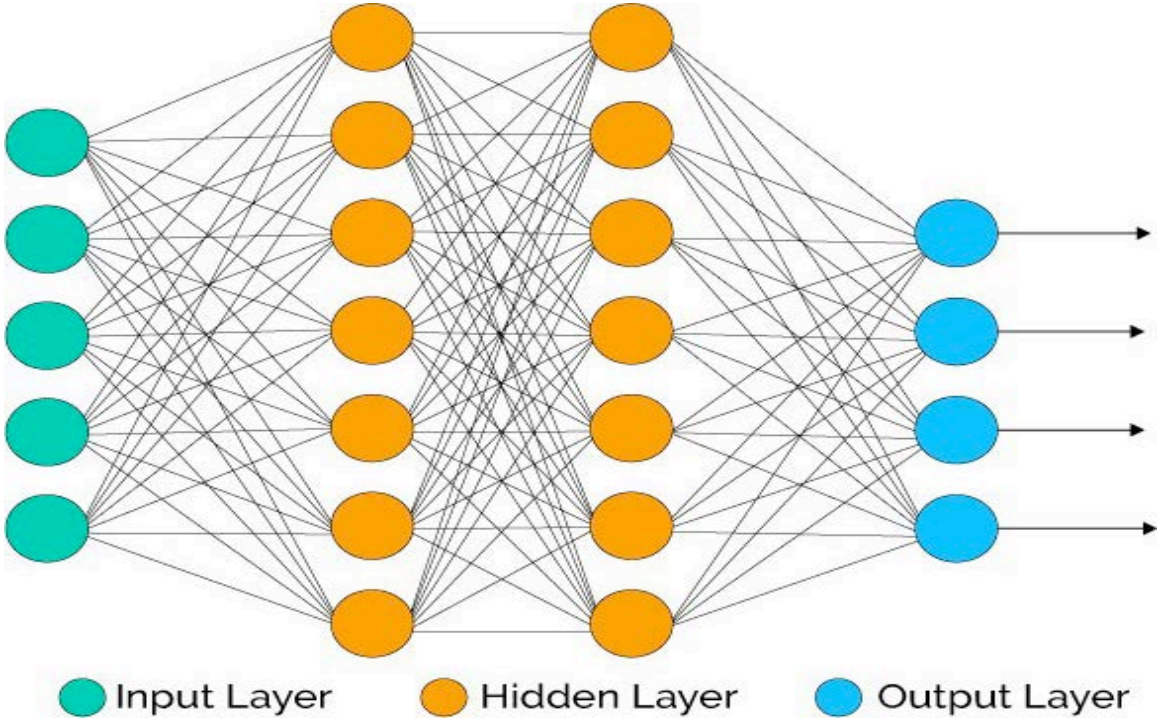
3. Senses are the next general AI step, modularity and feedback.

4. Remarkable *scaling* of both brains and AI code (if time allows)

§1. Sizes of Representations

The workhorse of deep learning is a neural net, modeled on actual neurons. Each unit in each layer has a real number activation which is a linear function of its inputs, truncated to a range, e.g. [-1,+1]. But how many units are needed for effective thinking?

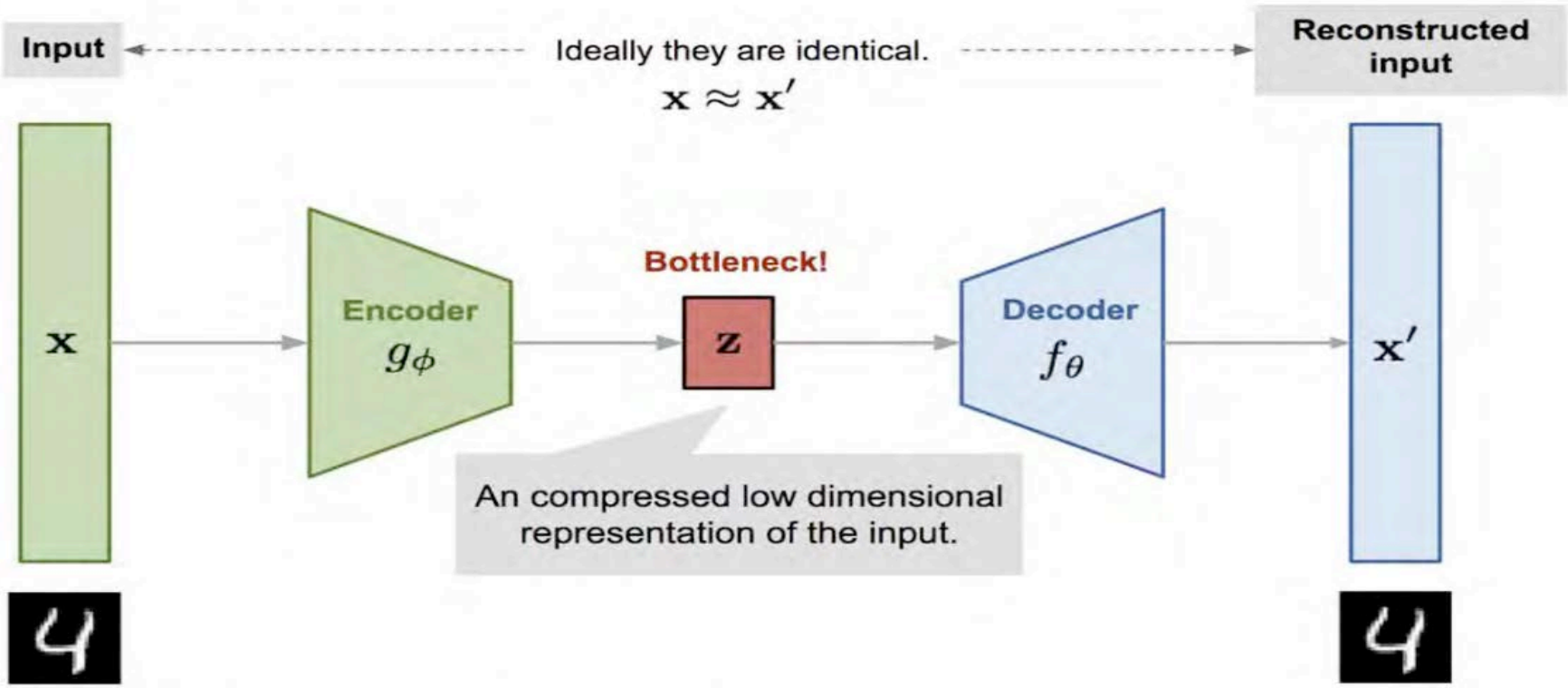




- Surprisingly (to me), the size of representations for thoughts has turned out to be very important.
- A breakthrough was Mikolov's 2013 program *word2vec,* in which high dimensional distance modeled *co-occurence statistics*, (frequency of pairs of words appearing near each other in texts) and he suggested using $\mathbb{R}^{512}$ representations of words. These could support things like [king] – [male] + [female] ~ [queen]
- There are even higher dimensional representations in the large language models. This mirrors finding that higher level visual neurons are not driven by specific stimuli, e.g. shape data is encoded distributively: outside primary sensory-motor areas where it is hard to find single neurons whose activity detects something meaningful.

- If the neural net has a thousand units in its layers, the linear activation functions are matrices with a million “weights” that must be learned by training.
- The large language models have billions, even trillions of such weights.
- To train or run them, CPUs are too slow. But the graphical processing unit (GPU) can be used.
- However, there are famous exceptions as Charley Gross discovered.

Even in monkeys, some neurons respond only to faces and do have clear relations to things like gender/age/skin color. “Auto-encoders” with drastic bottlenecks find these high level categories for face recognition. These computer responses are amazingly similar to those of Macaque infero-temporal neurons.



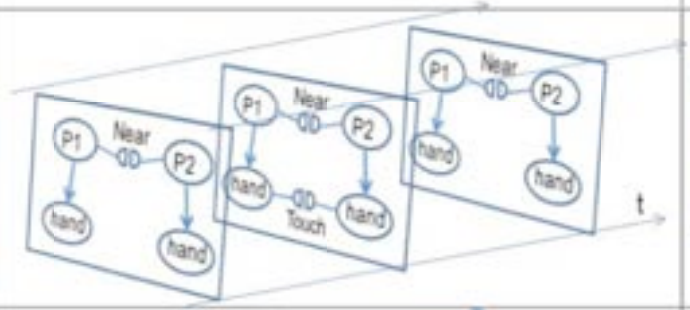
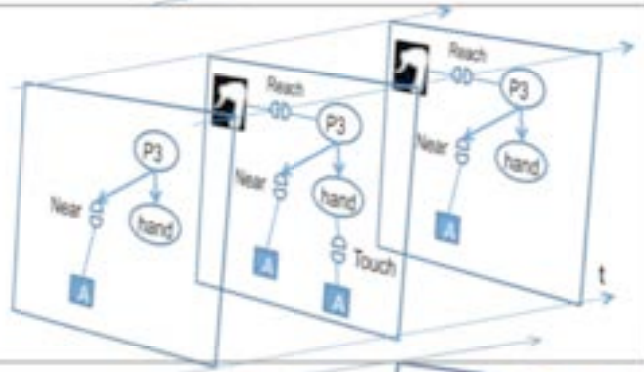
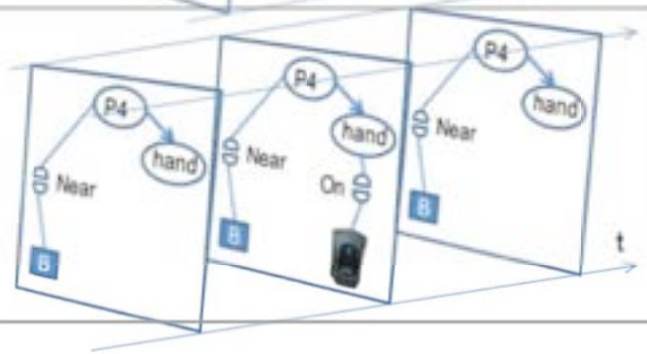
Higgins, Tsao et al, 2021, *Nature Communications,* “Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons”

§2 Transformers and Grammars

- Our thoughts clearly use small sets of discrete categories. Apparently the brain uses *both* high dimensional and low dimensional representations.
- *Transformers* were introduced in Google's 2017 paper "*Attention is all you need*". These project high dimensional representations onto much lower dimensions forcing low dimensional data to carry contextual information.
- Three matrices are used: one for the word being analyzed, one for a second word carrying context and a third for how to merge this information into the representation of the first.

- Transformers apparently encode grammars: Hewitt and Manning (2019) showed how much of traditional parse trees were hidden in BERT's $\mathbb{R}^{1024}$ representations by making suitable linear projections to $\mathbb{R}^{32}$.
- But grammars are essential in images, actions, thoughts in general. All thinking apparently involves grammatical tokens, re-usable groupings of smaller tokens filling “slots”, creating grammatical trees, assembled on the fly instantly.
- e.g. sentence = subject+verb+object OR
  face = eyes+nose+mouth

- Here is an example from Song Chun Zhu's team with evolving parse trees for actions:

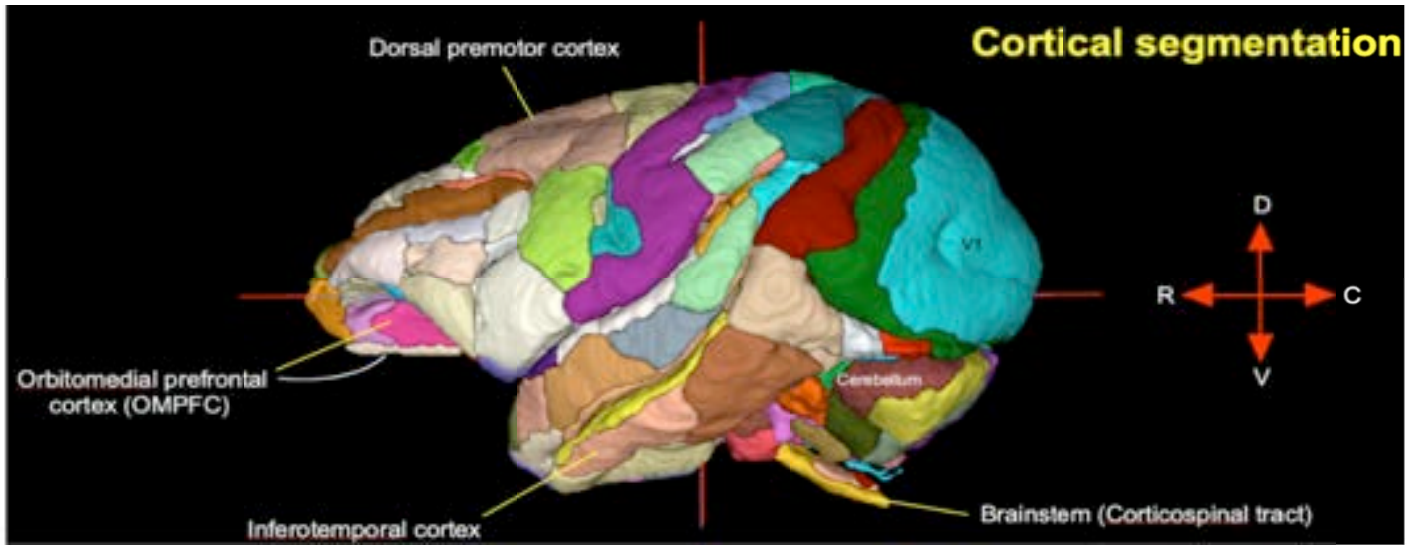
Atomic actions	Grounded relations	Symbols		Video examples
		Foreground	Background	
ShakeHands(P1,P2)	Near(P1,P2) And Touch(P1.hand, P2.hand)			
UseDispenser(P3)	Reach(P3) And Near(P3,A) And Touch(P3.hand,A)			
PickUpPhone(P4)	Near(P4,B) And On(P4.hand,B)			

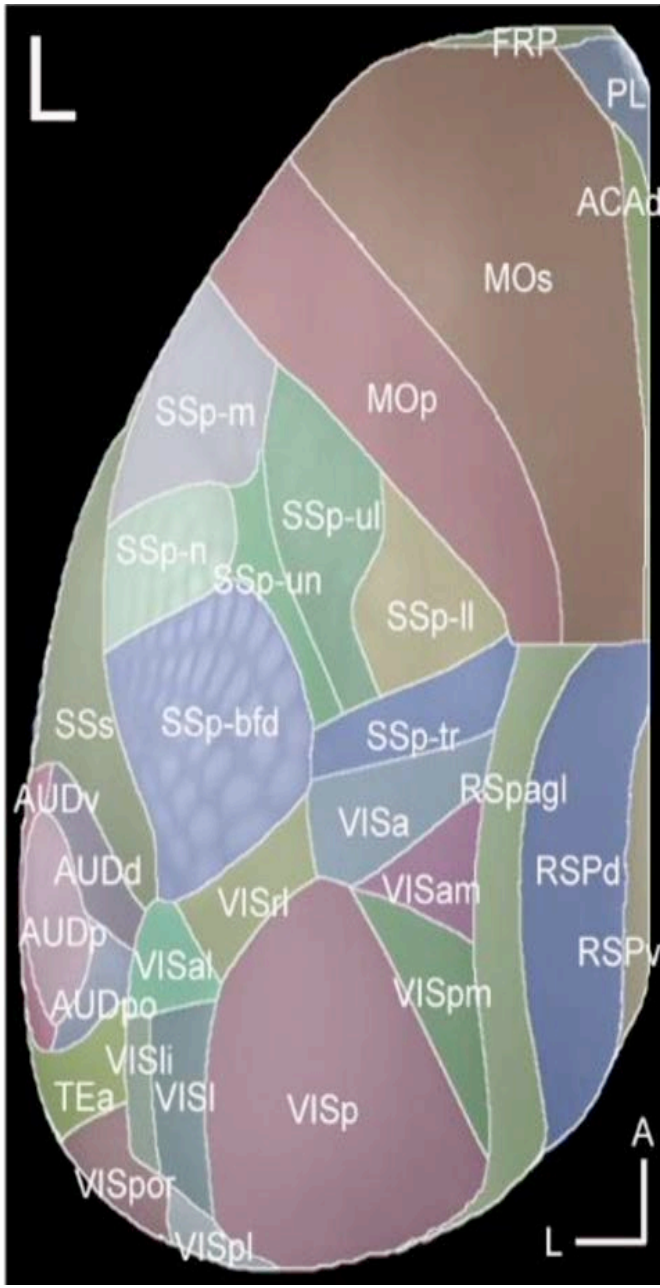
§3. Senses are the next step

- Babies start by mastering hand-eye coordination in their cradles, learning both seeing objects and moving their hands, thus a bit of the 3D structure of the world -- all together.
- The AI can also train on fluid motion, seeing and grasping discrete objects, combining motion + perception in either the real world or a virtual one.
- This involves coordinating multiple modules, seeing and moving.

Brains have multiple Modules

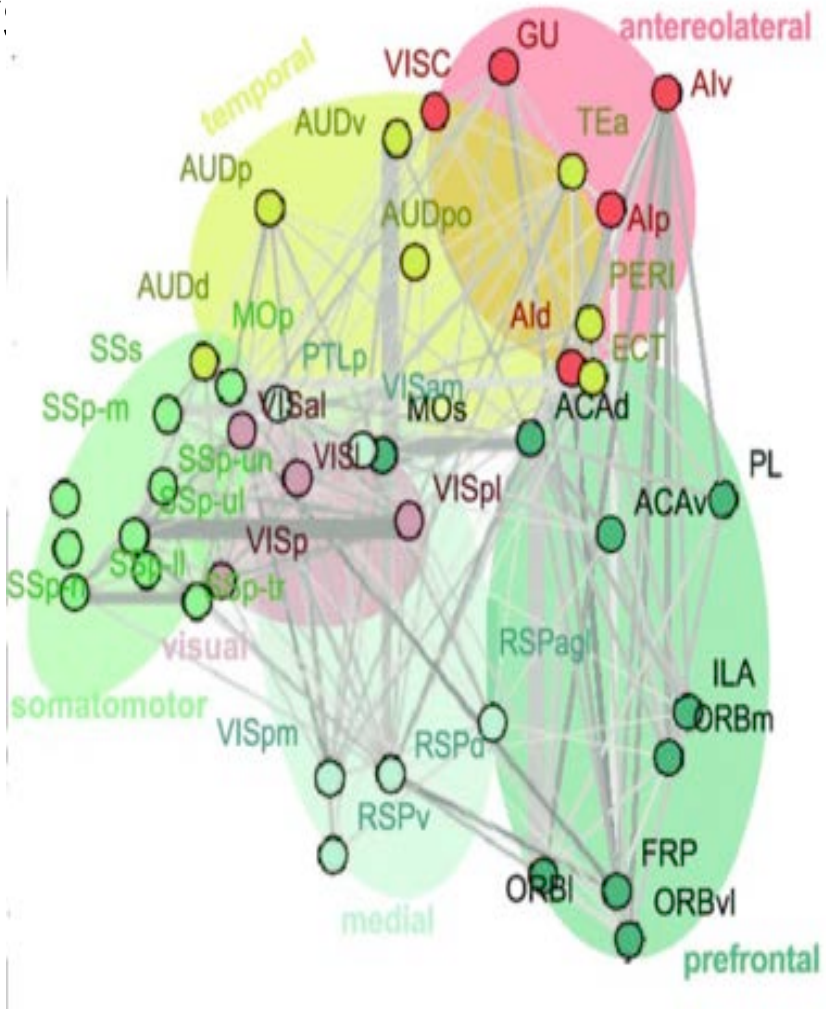
- Mammalian brains are divided into areas, primary and secondary areas for each sense, others for planning and motor control and yet others for memory.
- Right, a flattened enlarged mouse cortex; below, a colorful macaque brain



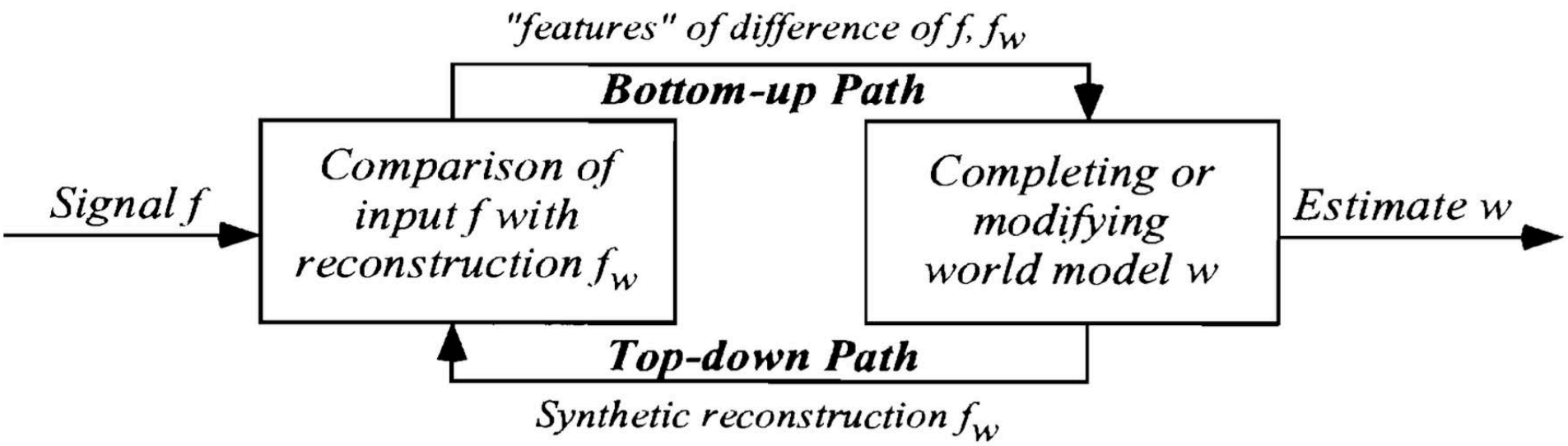


The Connectome

- The white matter of brains is made of the long distance axons, connecting areas. Working out the full connectome of mammals is a hot topic. Below is a recent connectome for the mouse. AI's can be structured in the very similar way
- In brains, every pair of areas with a feed forward connection, also has a feedback connection. *Feedback is ubiquitous and clearly will also be essential for the AGI.*
- The pathways between areas have *much lower bandwidth than local circuits* and drive distinct sets of synapses.



- A theory developed by Grenander's group instantiates Bayesian statistics to explain the way two areas interact. The higher area learns *p(w).* The *feedback* connections learn *p(f | w).*



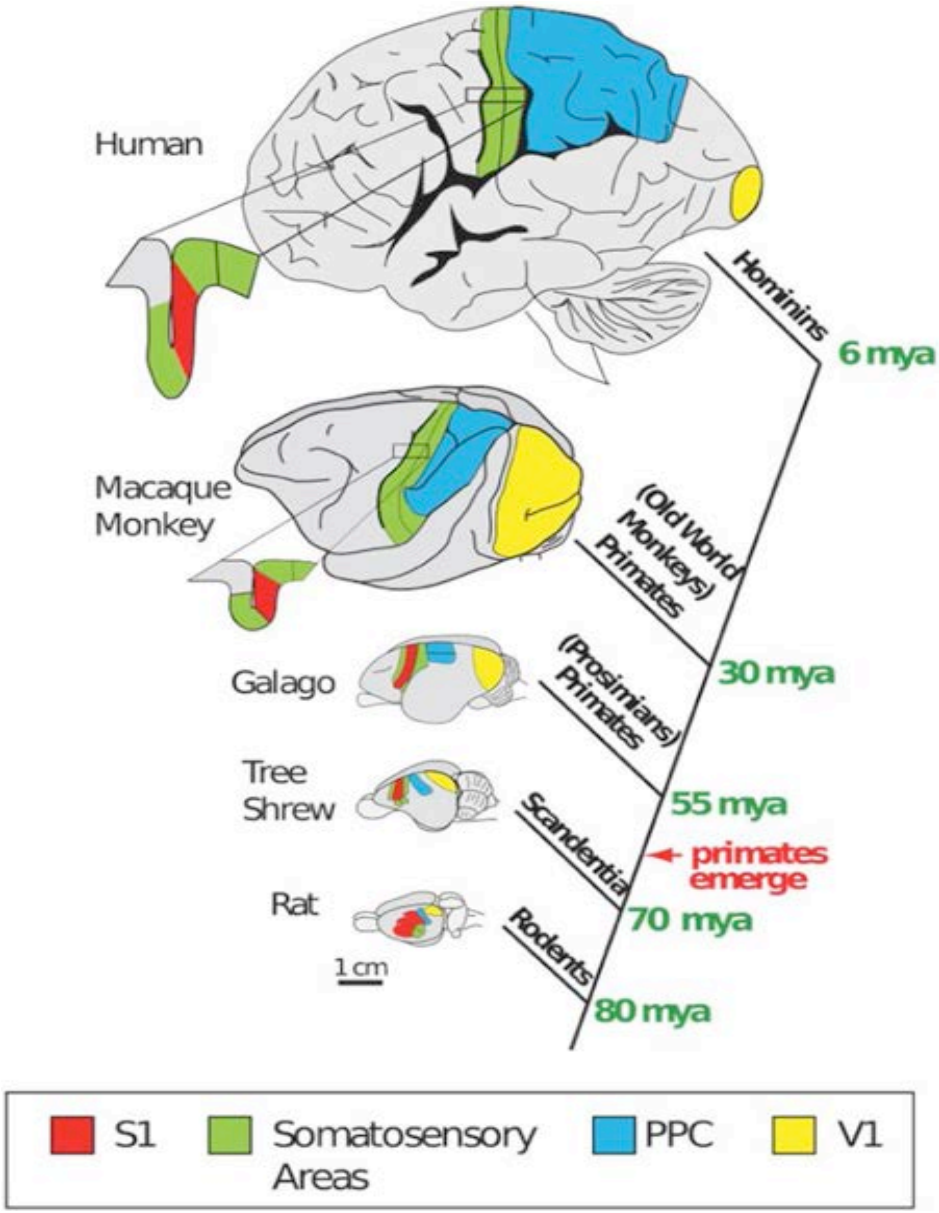
- All this can readily be done in computer code. Mammalian cortical architecture is a logical design to replicate.

Some by products

- With either the real or a virtual world, the key is to introduce a knowledge of the 3D world, occlusion, solid objects, moving around objects.
- Once the AI has senses, it can begin to know what the phrase “real world” means, it can build a memory, hence a concept of “now” in its ongoing interaction.
- Whether it ever has self-awareness is another question (nods to Eckhard Tolle, *The Power of Now*).
- Fully self driving cars may finally emerge, (Musk's oversimplifications notwithstanding) because it will possess common sense knowledge of interactions between hard objects.

§4. Scaling in Brain and Code

- A mouse cortex has roughly 13 million neurons and about 100 billion synapses while human cortex contains some 16 billion neurons, 100 trillion synapses.
  - All *mammalian* brains are almost completely identical – cell types, cell connections, layered structure, modular decomposition, neuro-transmitters, etc.



This is a **thousand fold** scaling!!!!! Moreover, there is no indication of genetic control of detailed connectivity, much seems random and strongly modified by learning.

- This strongly suggests that a few simple rules may be sufficient to generate intelligence.
- Like brains, deep learning code scales incredibly
  - To demonstrate the ideas to non-CS people, I use a 3 layered net with 7 units, 17 weights, that learns when a pair (*x,y*) is a point inside the unit circle or not.
  - LeCun's 1989 zip code reading net had 1256 units and 9760 independent weights.
  - Google's 2017 transformer model did translation and had 65 million parameters
  - GPT(3) in 2020 has 175 billion parameters; GPT(>3) multi-trillion parameters?

- Thousand-fold plus scaling is **never found** in the physical world, the social world, or the business world. When any object, town or corporation expands by an order of magnitude, new organizational structures are ALWAYS needed. *Things that do scale must be very simply built.*
- Both brain and deep learning AI codes are apparently trained from scratch with seemingly simple architectures. Neural nets with gradient descent training appear to be by far the best nonlinear regression algorithm ever devised. It seems to avoid much of the bias/variance conflict and to expand without substantial change.

Thank you for listening.

If there is no World War III, I am certain that fully functional intelligent robots will be built in the next couple of decades. I hope that some of you will be part of this amazing event.

My life has been made possible by international collaboration and I know from personal experience that brilliant people, crying out for research opportunities, *exist in every country*.

I pray for peace, that openness and understanding will spread to all nations.

The Surprising Rise of “Tiny AI”

How Small Generative Models Like H2O-Danube-1.8B Are Democratizing AI

Frederik Bussler

Published in Predict, Feb 12, 2024

The advent of large language models (LLMs) like GPT-4 ushered in a new era of advanced AI capabilities. However, the immense computational requirements of such models have traditionally made them inaccessible to most developers and organizations. This is now changing with the emergence of small generative AI models.

Dubbed “tiny AI,” compact yet powerful models like H2O-Danube-1.8B, Vicuna, Koala, Alpaca, and TinyLlama are bringing advanced generative abilities to the masses. Requiring modest computational resources, these small LLMs are reshaping the AI landscape by making AI more inclusive, innovative, and impactful.

The Democratization of AI

The most significant contribution of small AI models is their role in democratizing AI. By drastically cutting training and deployment costs, tiny AI places advanced capabilities into the hands of a broader audience. This includes individual developers, academics, startups, non-profits, and small-medium businesses.

For instance, models like H2O-Danube-1.8B can run efficiently on basic hardware like a single GPU. Unlike large models necessitating hundreds of GPUs or TPUs costing millions of dollars, tiny AI models have operational costs within the reach of most organizations and developers.

This democratization is a pivotal moment, allowing more people to tap into and mold the power of AI based on their unique needs. It paves the way for customized applications from personalized medicine systems to tailored virtual assistants. Ultimately, the innovations enabled by democratized tiny AI could profoundly impact business, education, healthcare, and society as a whole.

The Power Packed Capabilities of Tiny AI

But are small models compromising too heavily on capabilities in pursuit of efficiency? Not at all! Tiny AI models can perform extraordinarily well across diverse AI tasks:

- **Text Generation:** Models like Koala, Alpaca, and H2O-Danube-1.8B can generate coherent, human-like text
- **Language Translation:** Small models can translate text between languages with high accuracy. For instance, models like Vicuna have achieved great success in English-Spanish translation
- **Question Answering:** Tiny AI models can provide informative responses to a broad spectrum of questions across domains like science, history, and current affairs
- **Task Automation:** Small models can follow instructions to automatically execute tasks like scheduling meetings, drafting documents, and filling forms

Indeed, while large models capture more knowledge and context, tiny AI models offer remarkable generative abilities relative to their size. And constant enhancements in model design and training are unlocking even more powerful small models over time.

Real-World Impact

As tiny models become more powerful and affordable, they can catalyze change across diverse domains:

- Healthcare: Small models can enable personalized medicine by rapidly analyzing patient data for tailored diagnoses and treatment plans. Startups are also using these models for applications like faster disease diagnosis.
- Customer Service: Compact generative models can cut costs and enhance experience by automating customer service processes. They are being applied to create 24/7 chatbots and drafting better responses.
- Creative Applications: Democratized generative AI powers new applications fueling human imagination. Models aid authors in writing novels, help illustrators digitally render images, and assist musicians in songwriting.
- Education: Tiny AI tools effectively adapt to student learning patterns. Education groups employ these assistive models for applications ranging from customized pedagogical agents to automated grading tools.

Indeed, tiny AI is reshaping domains ranging from graphic design to supply chain optimization. And these applications are just the beginning. Tiny models have immense headroom for adaptation to new domains and tasks by users and will enable innovations unimaginable today.

The Bigger Picture

The proliferation of tiny AI marks a pivotal juncture with massive implications for technology and society. However, it also warrants thoughtful discourse on associated opportunities, challenges, and risk mitigation.

On the positive side, small generative models can bring free-of-cost AI assistance to millions globally with transformative humanitarian potential. Further, by expanding AI's reach, they set the stage for unprecedented tech-enabled collaboration, collectively taking scientific progress to new heights.

The Way Forward

Today, we stand at the brink of an AI revolution powered by tiny AI models like H2O-Danube-1.8B. Much like personal computers and mobile phones before them, small models will fundamentally transform how individuals and organizations leverage technology.

In an AI-integrated future catalyzed by tiny models, we will have hyper-personalized healthcare, fluid human-computer collaboration, and boundless outlets for human creativity aided by AI. As these models continue evolving rapidly, supported by advances like sparse AI and liquid computing architectures, their efficiency and capabilities will scale dramatically.

This new era of Tiny AI warrants excitement and cautious optimism. With collective responsibility across researchers, developers, policymakers and end-users prioritizing safety, accessibility and

innovation simultaneously, the proliferation of small models can usher in an AI revolution that uplifts societies holistically.

The Rise of ‘Tiny AI’ and Why It Matters

“Large” language models aren’t everything

See all from Frederik Bussler

Claude 3, ChatGPT, and The Death of LLMs

The End of the Language Era?

AI Showdown: Microsoft Copilot vs. Google Gemini vs. ChatGPT 3.5 vs. Mistral vs. Claude 3

A Comprehensive Guide to the Best AI Assistants in 2024 in AI Technology