From Deep Learning and Transformers to AI Risks and Safety

P.Baldi

Department of Computer Science
AI in Science Institute
Center for Machine Learning and Intelligent Systems
University of California, Irvine

Foundations of Attention Mechanisms and Transformers

P.Baldi

Department of Computer Science
AI in Science Institute
Center for Machine Learning and Intelligent Systems
University of California, Irvine

RoadMap

1. Introduction to Attention and the Standard Model
2. A Taxonomy of Attention Mechanisms (Quarks)
3. Transformers and Attention
4. Applications of Attention
5. Mathematical Theory of Attention

RoadMap

1. Introduction to Attention and the Standard Model
2. A Taxonomy of Attention Mechanisms (Quarks)
3. Transformers and Attention
4. Applications of Attention
5. Mathematical Theory of Attention

What is Attention?

``Everyone knows what attention is... It is the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought...'' William James, Principles of Psychology (1890).

“the ability to focus selectively on a selected stimulus, sustaining that focus and shifting it at will‘’

``the concentration of awareness on some phenomenon to the exclusion of other stimuli''.

Attention

``Everyone knows what attention is... It is the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought...'' William James, Principles of Psychology (1890).

“the ability to focus selectively on a selected stimulus, sustaining that focus and shifting it at will“

``the concentration of awareness on some phenomenon to the exclusion of other stimuli''.

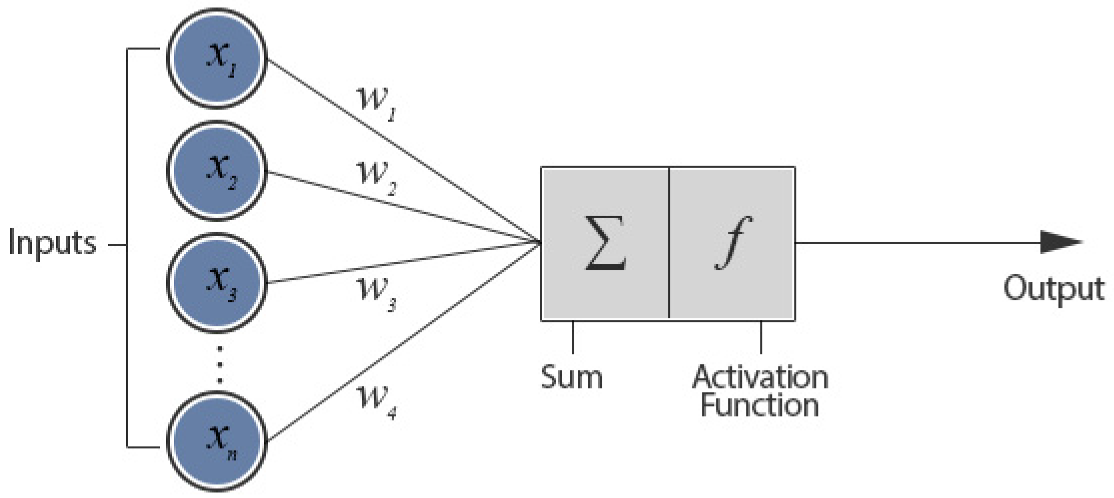
Neurobiology of Attention

- The word “attention” is an inadequate, singular term for a multitude of inter-related processes. We use a host of adjectives to describe attention—for example, we say that attention can be divided, oriented, sustained, or focused, and many of these descriptions likely reflect underlying, dissociable neural processes. Complicating matters, attentional resources can be allocated to either external stimuli, or to internal stimuli such as thoughts and memories. Furthermore, we often confuse the regulation of attention (a covert behavior) with the regulation of movement (an overt behavior) when discussing an “attentional disorder”.

[Arnsten and Castellanos. Neurobiology of attention regulation and its disorders, *Pediatric Psychopharmacology*, 2010].

→ Focus on the most basic building blocks of what attention may be in artificial neural networks (the Standard Model).

The Standard Model



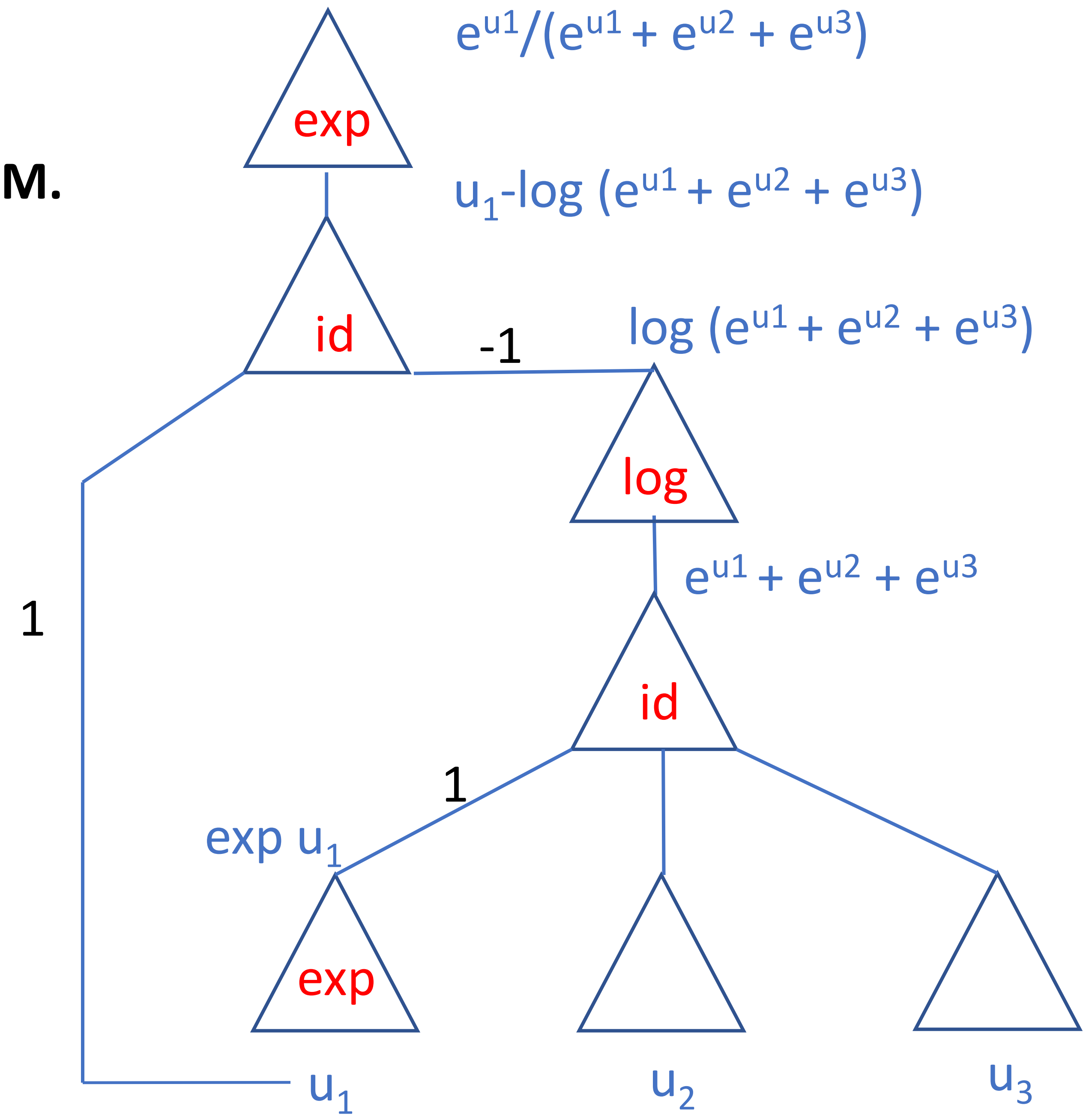
- SM universal approximation properties
- SM extensions (softmax, polynomial activations, product of outputs, ….)

O=f($\sum w_i x_i$)

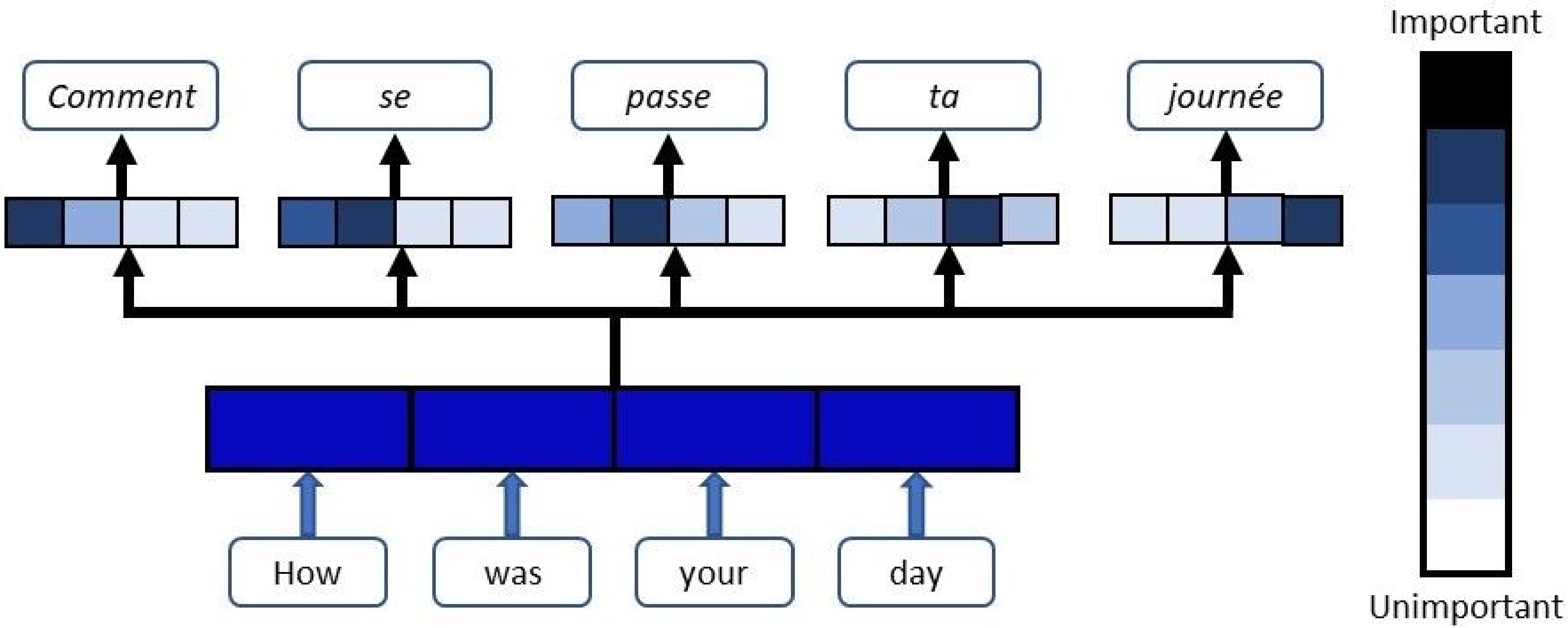
Basic elementary operations:

1) Activation S= Dot product x.w
2) Output O=f(S) (f linear or non-linear activation function)

SoftMax is an extension of the SM.



Attention in DL and NLP applications



Sequence to sequence models

Attention Mechanisms in DL and NLP

Various formulations:

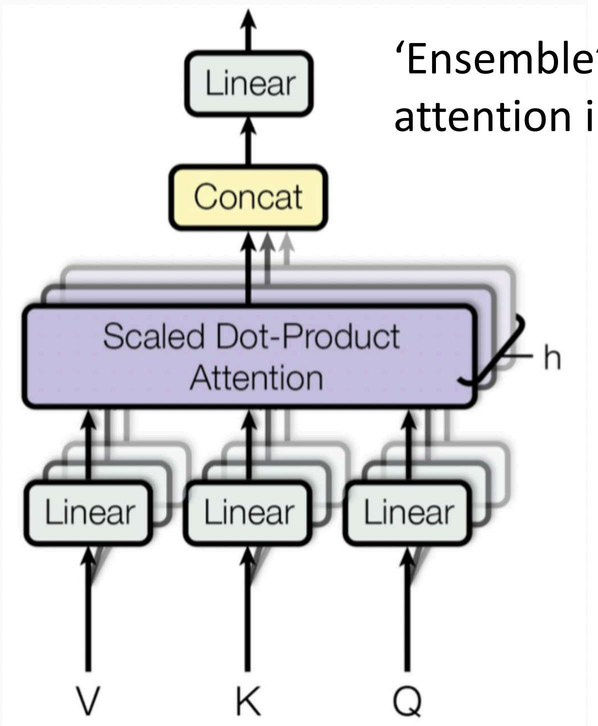
- Content-base attention Graves et al., 2014
- Dot-Product attention Luong et al., 2015
- Additive attention Bahdanau et al., 2015
- Vaswani et al. 2017
- ………..
- **Transformer Architectures**
- Standard modules in DL packages (TensorFlow, PyTorch)
- Google's BERT, OpenAI's GPT , XLNet ….

Transformer Model & (self)-attention

The Transformer Model is **entirely** built on the self-attention mechanisms, **without** using sequence-aligned recurrent architectures.

Every input element has three learnable vectors: Query (Q), Key (K), and Value (V)

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{QK}^{\top}}{\sqrt{n}}\right)\mathbf{V}$$



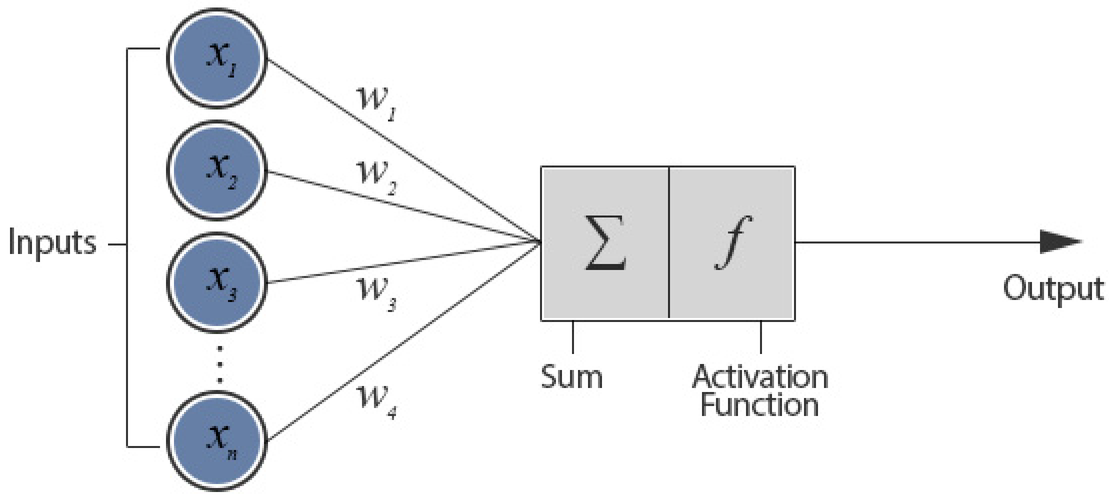
'Ensemble' multiple attention in parallel

Rather than only computing the attention once, the multi-head mechanism runs through the scaled dot-product attention multiple times in parallel.

RoadMap

1. Introduction to Attention and the Standard Model
2. A Taxonomy of Attention Mechanisms (Quarks)
3. Transformers and Attention
4. Applications of Attention
5. Mathematical Theory of Attention
6. Large Language Models

The Standard Model



O=f($\sum w_i x_i$)

Basic elementary operations:

1) Activation S= Dot product x.w

2) Output O=f(S) (f linear or non-linear activation function)

3 variable types:

S, O, w

Classification of Attention Mechanisms (or Extensions of the SM)

- In the SM, there are 3 types of variables: S (activation), O (output), and w (synaptic weights).
- Attention signals can be classified according to their attending Origin, their attended Target, and the underlying Mechanism.
- With two mechanisms, addition and multiplication, this corresponds to 18 possibilities:

	S	O	W
S	+, x	+, x	+, x
O	+, x	+, x	+, x
W	+, x	+, x	+, x

- Multiplicity issues.
- Origin: only of type O → 6 possibilities.

Classification of Attention Mechanisms

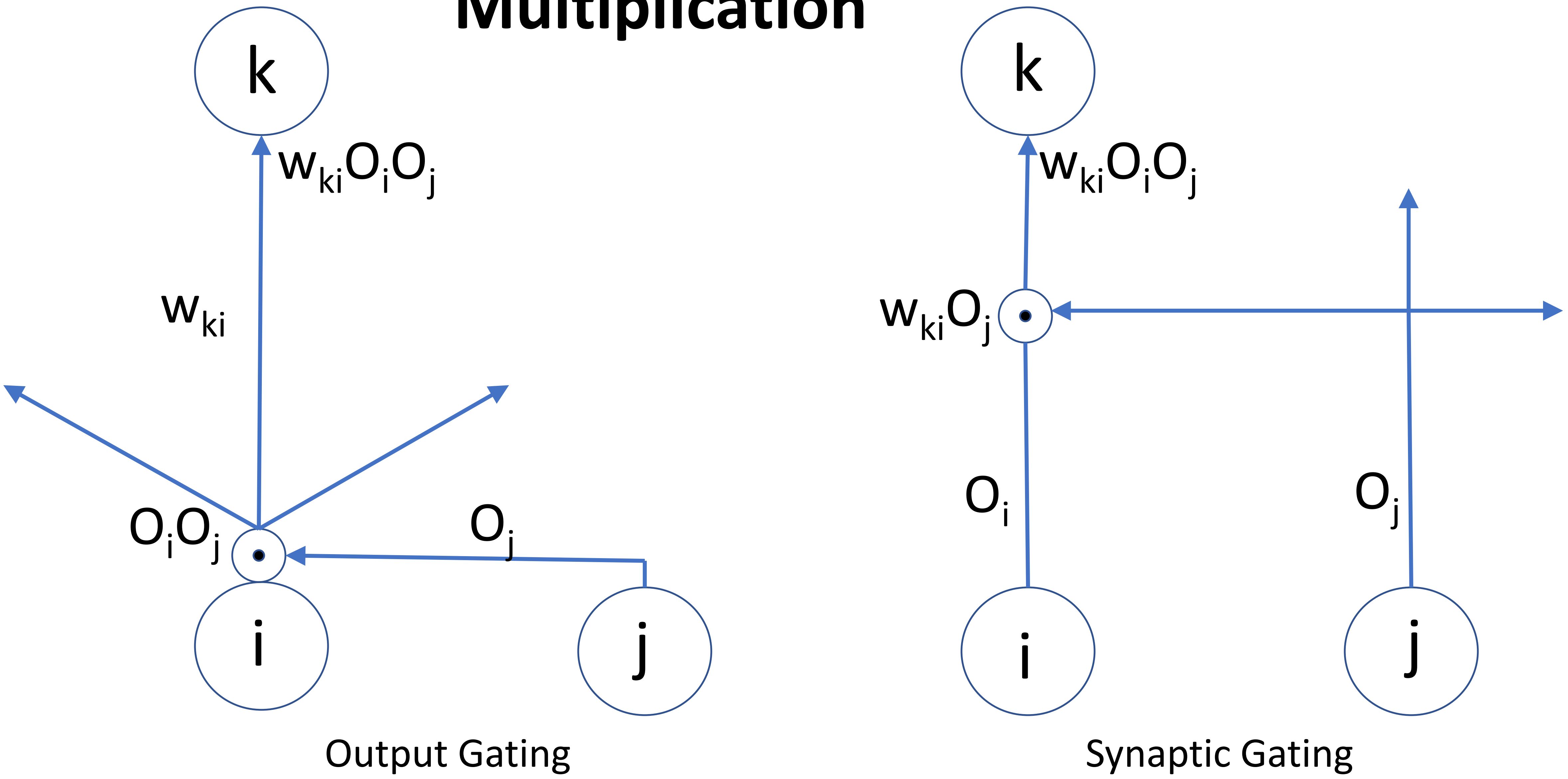
- **Origin is of type O**
- **Six possibilities:**

Target

Mechanism

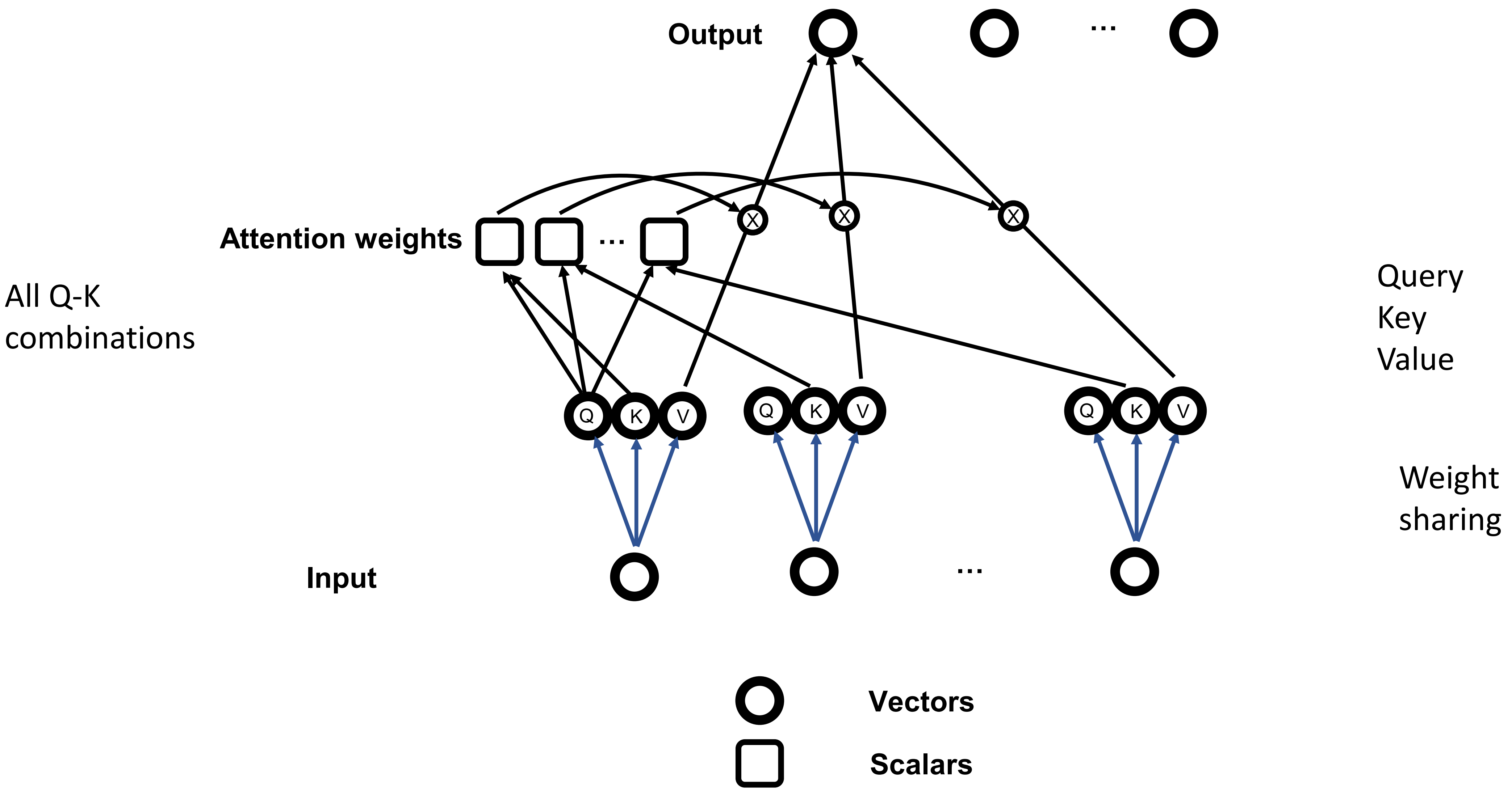
	Activation (S)	Output (O)	Weight (w)
Addition	Activation Attention (SM)		
Multiplication		Output Gating	Synaptic Gating

Multiplication



RoadMap

1. Introduction to Attention and the Standard Model
2. A Taxonomy of Attention Mechanisms (Quarks)
3. Transformers and Attention
4. Applications of Attention
5. Mathematical Theory of Attention


Output
...
Attention weights
...
X
X
X
All Q-K
combinations
Query
Key
Value
Q
K
V
Q
K
V
Q
K
V
Weight
sharing
Input
...
Vectors
Scalars

Database Vocabulary

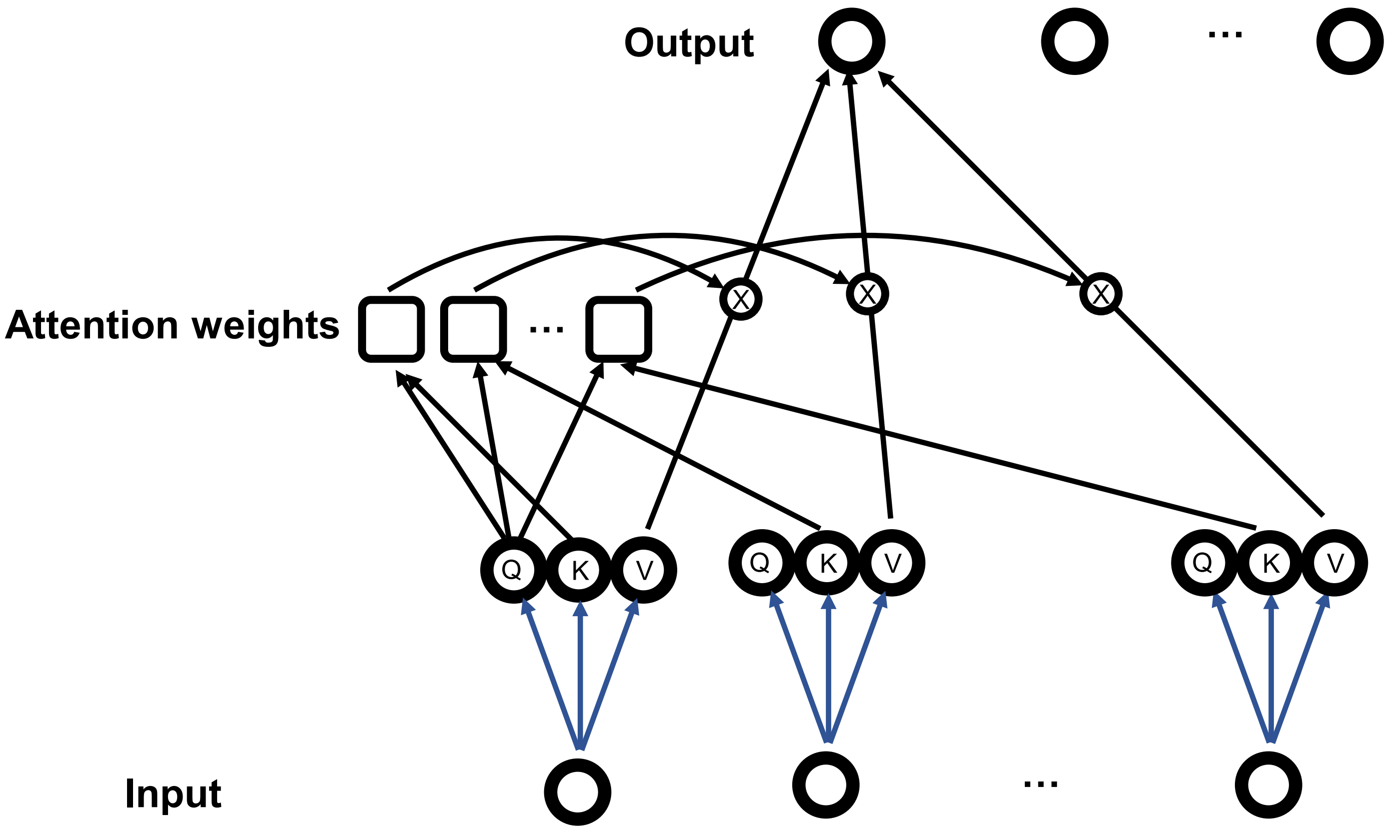
Key

Student ID	Driver License #	Address	First Name	Last Name
	123456			
	123789			
	123770			
	123775			

Values= Rows Contents

Query: 123770?

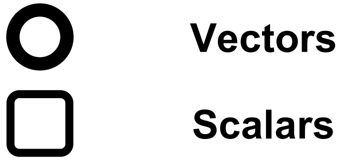
All Q-K
combinations



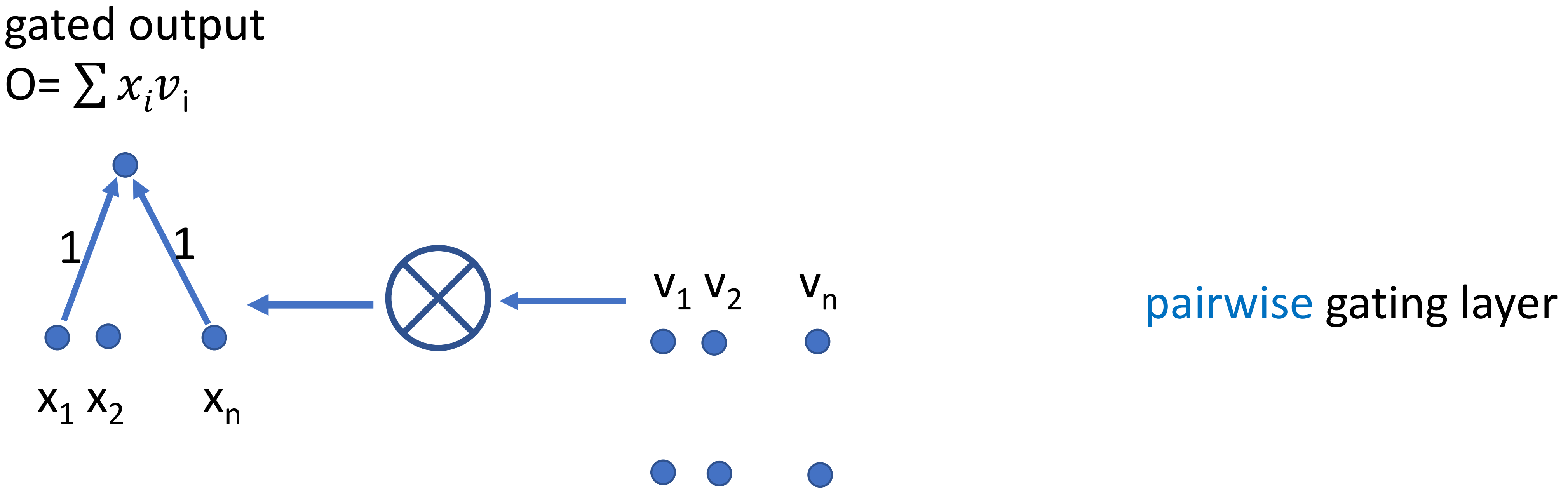
Query
Key
Value

Weight
sharing

Fundamental Property: equivariant to
permutation of the inputs (!).



Attention Enables Computing the Dot Product of the Activities of Two Layers of the Same Size (output or synaptic gating)

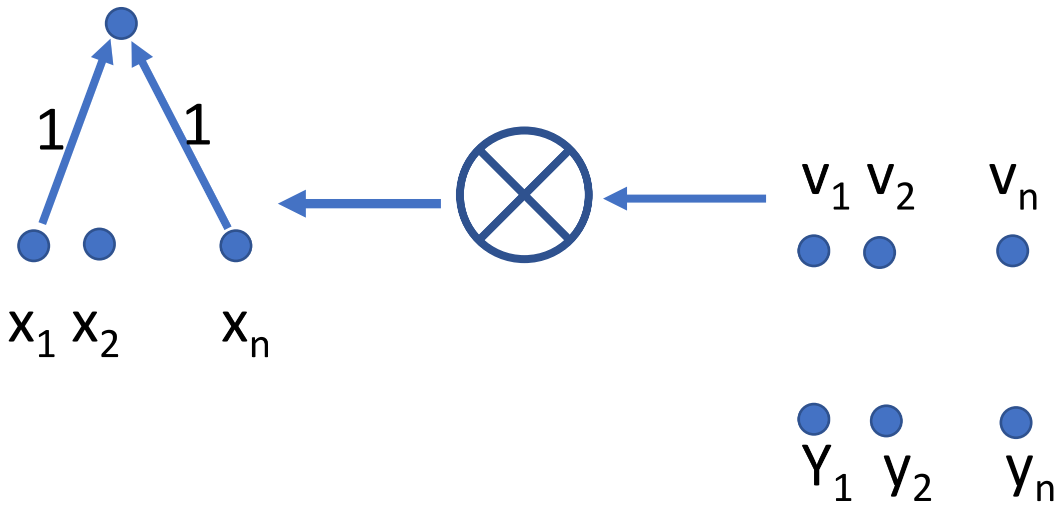


[Can be used to derive alternative proof of universal approximation properties for SM + attention]

Softmax Attention=Dot Product with Softmax (output or synaptic gating)

gated output

O=sum$_i$ v_i x_i



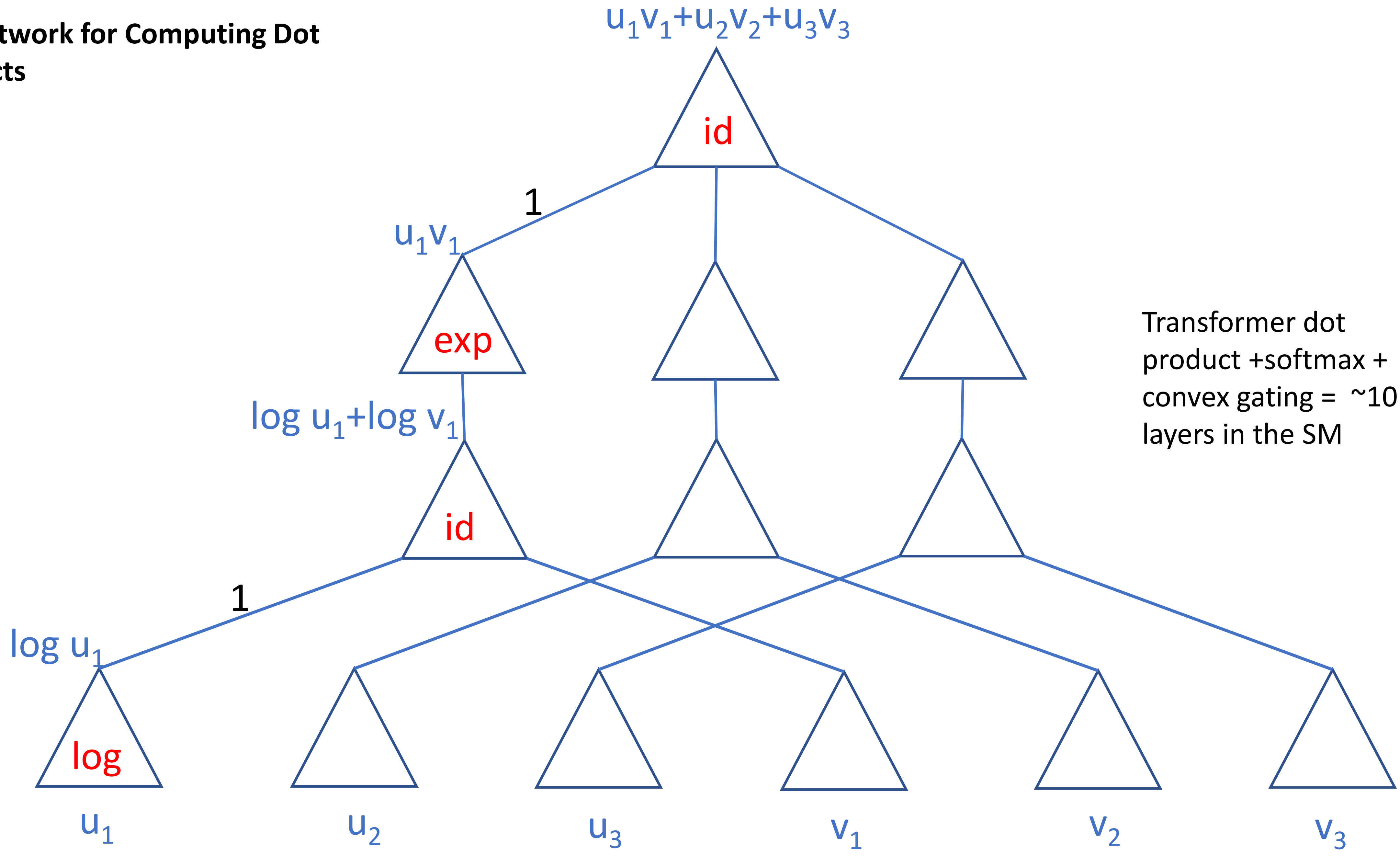
gating layer:

softmax unit

vi=exp y_i / sum$_j$ exp v_j

Attention in NN is based on the ability to compute and fast-store variable-length dot products.

SM Network for Computing Dot Products



Transformer dot product +softmax + convex gating = ~10 layers in the SM

RoadMap

1. Introduction to Attention and the Standard Model
2. A Taxonomy of Attention Mechanisms (Quarks)
3. Transformers and Attention
4. Applications of Attention
5. Mathematical Theory of Attention

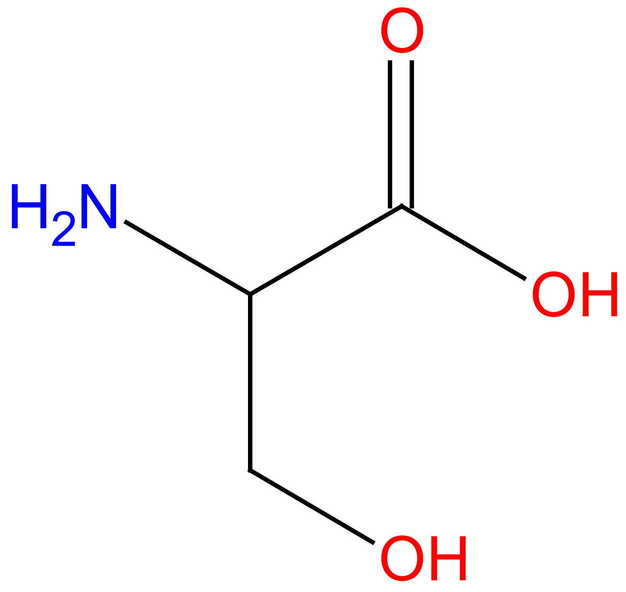
Chemistry Applications

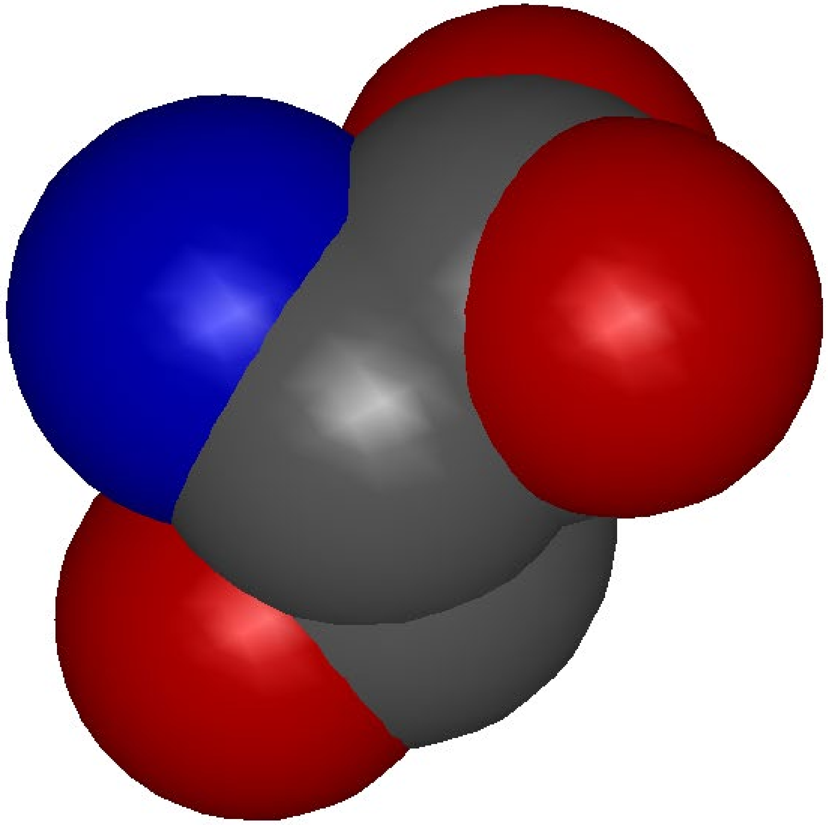
- Prediction of Chemical Reactions

Physics Applications

- Tagging Extreme Jets
- Jet Parton Matching
- Neutrino Classification
- Neutron Stars EoS

Small Molecule Representations





Problem: molecular graphs are undirected

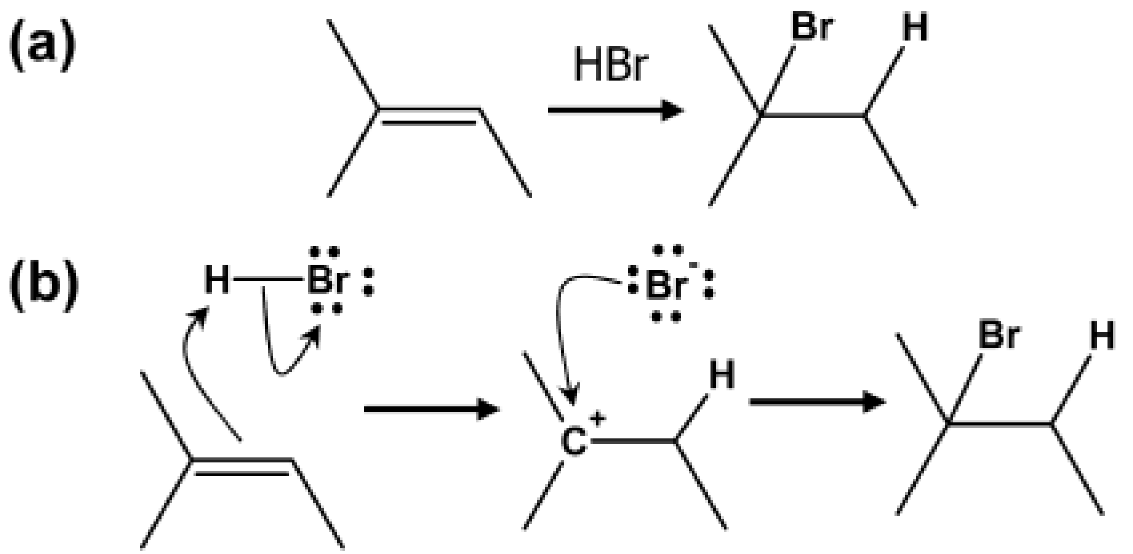
NC(CO)C(=O)O

0010001001010001

Deep Learning Chemical Reactions

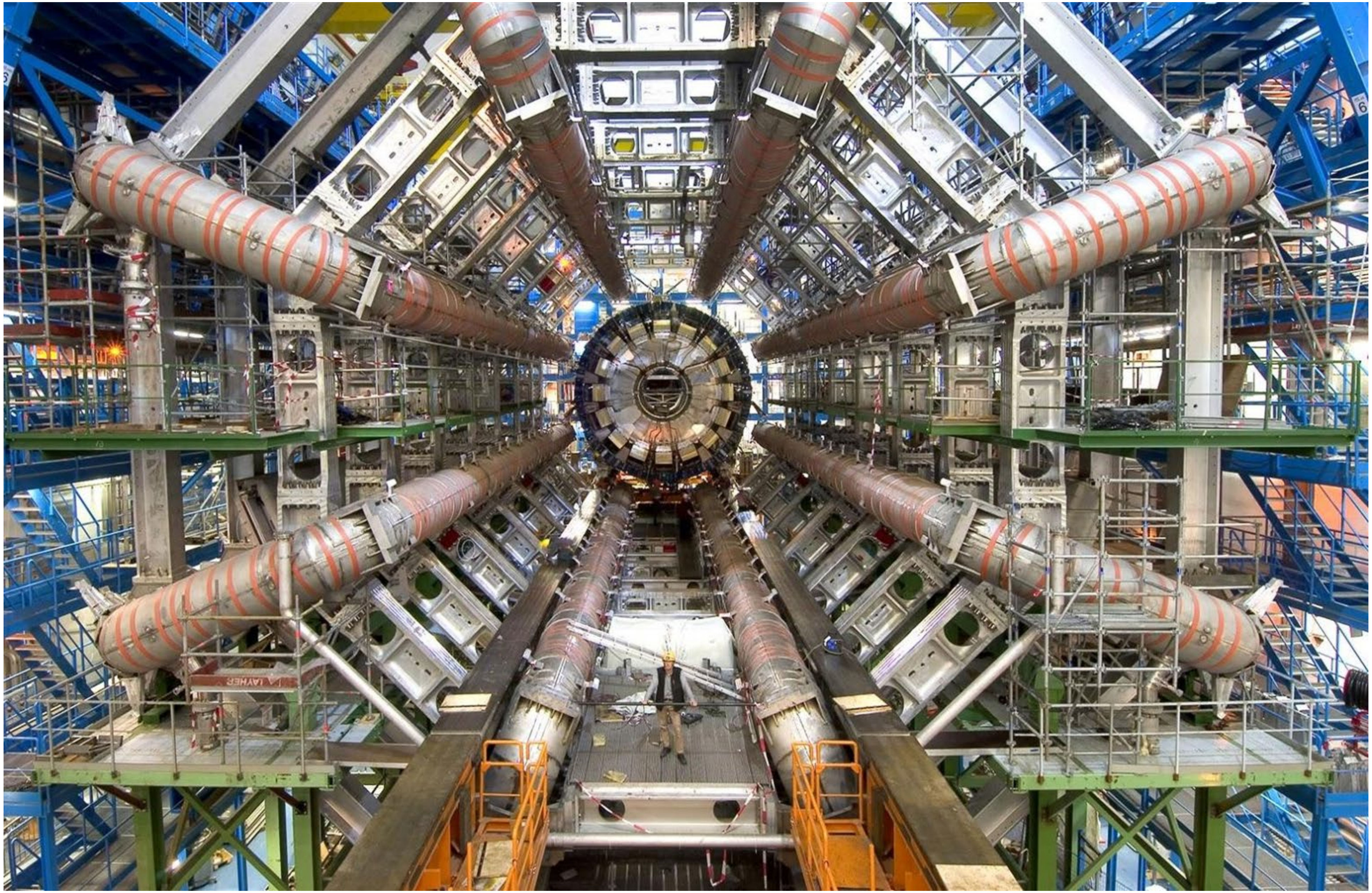
A+B → C+D

RCH=CH2 + HBr → RCH(Br)–CH3



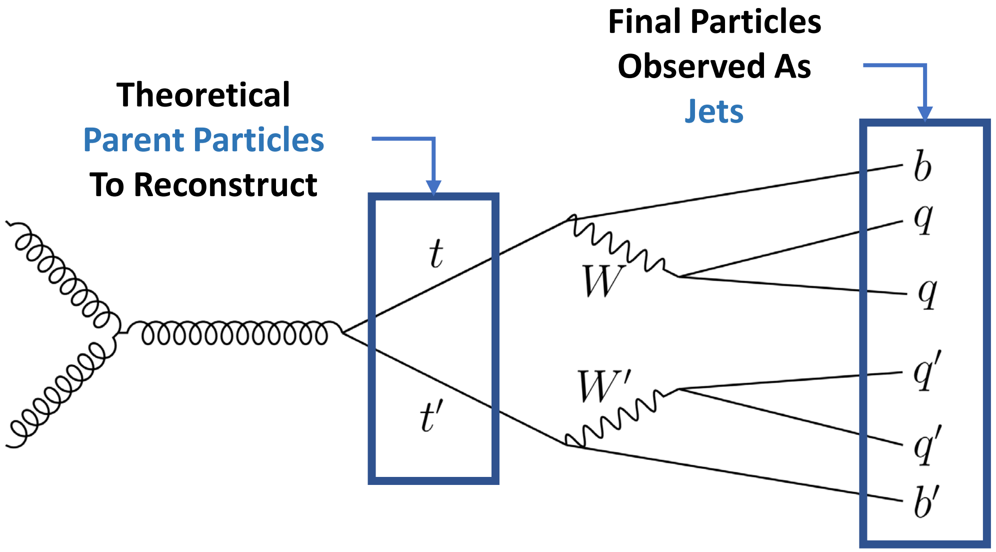
Amin Tavakoli

David Fooshee, Aaron Mood, Eugene Gutman, Amin Tavakoli, Gregor Urban, Frances Liu, Nancy Huynh, David Van Vranken, and Pierre Baldi. Deep Learning for Chemical Reaction Prediction. Molecular Systems Design & Engineering, Royal Society of Chemistry, 3, 442 – 452, (2018).



SPANet Jet-Parton Matching in LHC Top Quark Decays

- Primary (all-hadronic) decay channel produces six particles - two qqb triplets with opposite charge – originating from the top – antitop particle pair which we wish to reconstruct.
- After these particles are produced, they are propagated and measured by the detector as ***jets***.
- Along with the jets from each of the particles, there may be additional jets from other decay products.



$$\{j_1, j_2, j_3, j_4, j_5, j_6, j_7, j_8\}$$

Match Jets to Particle Labels

$$\{b, q', \emptyset, q', b', \emptyset, q, q\}$$

Garbage Jets

This is a difficult matching problem: Observing the jets from the detector, can you determine which jets belong to which particles?
Effective matching requires exploiting the symmetries in this problem!

SPANet Complete Architecture

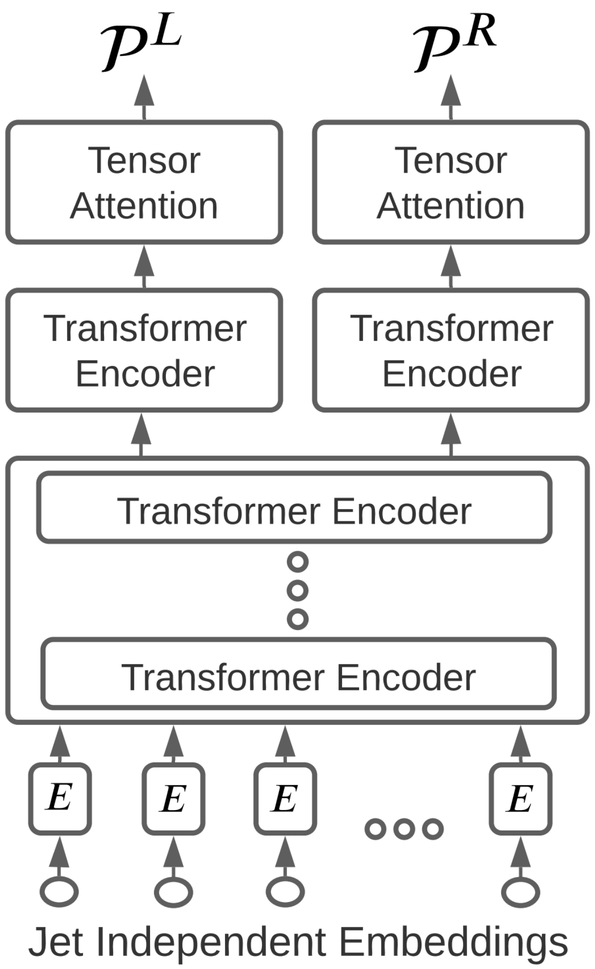
Construct an architecture following the structure of the original Feynman Diagram with attention as its core operation.

Attention Use

Symmetric jet matching

Particle-level encoding

Event-level context-aware encoding



Tensor attention to predict the most likely assignment of jets associated with each particle.

Split the information stream into a finite collection of "particles".

Heavily employ attention in several sections within our network for **context-aware permutation-invariant** learning.

Input is unsorted set of jet 4-momentum vectors.

SPANet Results

- We compare *SPANet* to a classical permutation-based method based on χ^2 probability of assignments.
- *SPANet* uses attention to match all top-quarks while the χ^2 method needs to compute many jet-permutations.
- *SPANet* **reduces the runtime** from $O(N^6)$ to $O(N^3)$ while **increasing efficiency** by $\sim 30\%$ across the board.

	χ^2 Efficiency			SPA-NET Efficiency		
N_{jets}	ϵ^{event}	ϵ_2^{top}	ϵ_1^{top}	ϵ^{event}	ϵ_2^{top}	ϵ_1^{top}
6	61.8%	65.0%	24.2%	80.7%	84.1%	56.7%
7	40.8%	50.4%	24.6%	66.8%	75.7%	56.2%
≥ 8	23.2%	35.5%	20.2%	52.3%	66.2%	52.9%
Inclusive	**37.7%**	**47.0%**	**23.0%**	**63.7%**	**73.5%**	**55.2%**

Runtime on 8 jet events

χ^2 : 369 *ms* per event

Spatter : 4.4 *ms* per event

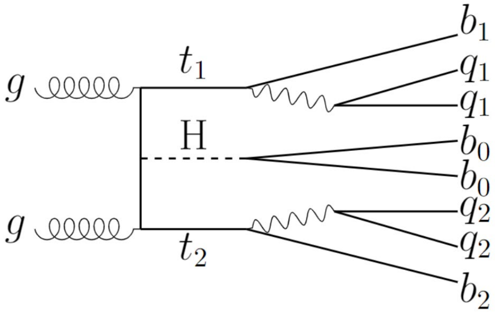
Alexander Shmakov

Michael James Fenton, Alexander Shmakov, Ta-Wei Ho, Shih-Chieh Hsu, Daniel Whiteson, and Pierre Baldi. Permutationless many-jet event reconstruction with symmetry preserving attention networks. *Physical Review* D, in press.

SPANet Results

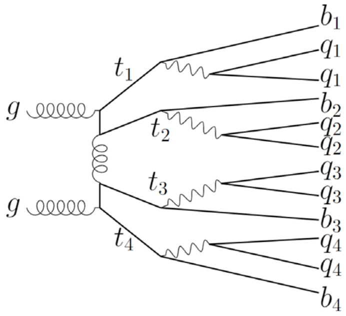
- General formulation allows us to extend this technique to virtually any possible event at the LHC.
- Split particle paths and symmetric attention may be extended to match jets in **incomplete events** – where one or more particles are missing due to detector loss, allowing us to use more training data.
- Extended this technique to two other, more complicated, events at the LHC: ttH and $tttt$.
- $tttt$ Event is so complex and large that the χ^2 method **cannot be tractably computed**!

ttH



	N_{jets}	Event Fraction	SPA-NET Efficiency Event	Higgs	Top	χ^2 Efficiency Event	Higgs	Top
All Events	== 8	0.261	0.370	0.497	0.540	0.056	0.193	0.092
	== 9	0.313	0.343	0.492	0.514	0.053	0.160	0.102
	$\geq$ 10	0.313	0.294	0.472	0.473	0.031	0.150	0.056
	Inclusive	**0.972**	**0.330**	**0.485**	**0.502**	**0.045**	**0.164**	**0.081**
Complete Events	== 8	0.042	0.532	0.657	0.663	0.040	0.220	0.135
	== 9	0.070	0.422	0.601	0.596	0.019	0.152	0.079
	$\geq$ 10	0.115	0.306	0.545	0.523	0.004	0.126	0.073
	Inclusive	**0.228**	**0.383**	**0.583**	**0.572**	**0.016**	**0.153**	**0.087**

tttt



	N_{jets}	Event Fraction	SPA-NET Efficiency Event	Top Quark
All Events	== 12	0.219	0.276	0.484
	== 13	0.304	0.247	0.474
	$\geq$ 14	0.450	0.198	0.450
	Inclusive	**0.974**	**0.231**	**0.464**
Complete Events	== 12	0.005	0.350	0.617
	== 13	0.016	0.249	0.567
	$\geq$ 14	0.044	0.149	0.504
	Inclusive	**0.066**	**0.191**	**0.529**

Alexander Shmakov, Michael James Fenton, Ta-Wei Ho, Shih-Chieh Hsu, Daniel Whiteson, Pierre Baldi. SPANet: Generalized Permutationless Set Assignment for Particle Physics using Symmetry Preserving Attention. *SciPost Physics*, in press.

x5000

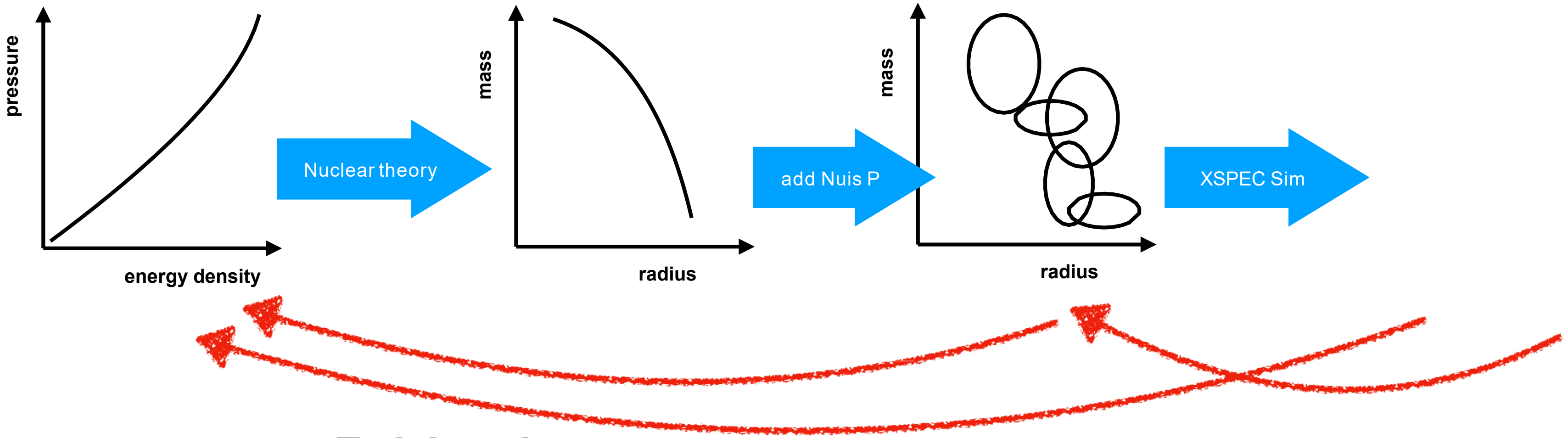
The problem

Equation of State (2 params)

Mass-Radius relation (2 params: M,R)

Star parameters
2 params: (mass, radius)
3 nuis params (dist, temp, dust)

Neutron Star X-ray data
(1024 chan,80% empty)



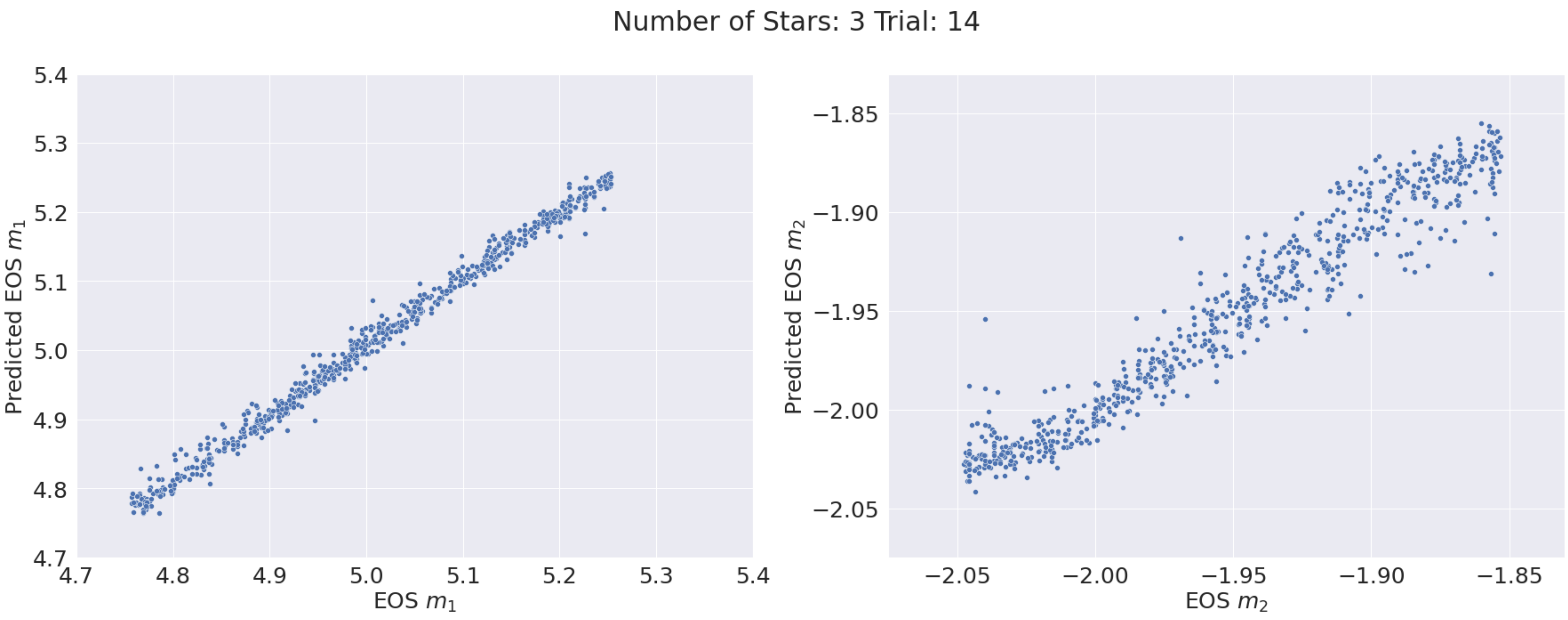
<u>Training data</u>
Fixed EOS, sample of (M,R) pairs
For each M,R pair, add 3 nuisance param
generate sample spectra

<u>Inference</u>
End-to-end: spectra -> EOS
Also might try: spectra-> star
star -> EOS

https://arxiv.org/abs/2002.04699

Jordan Ott

Prediction of EOS coefficients

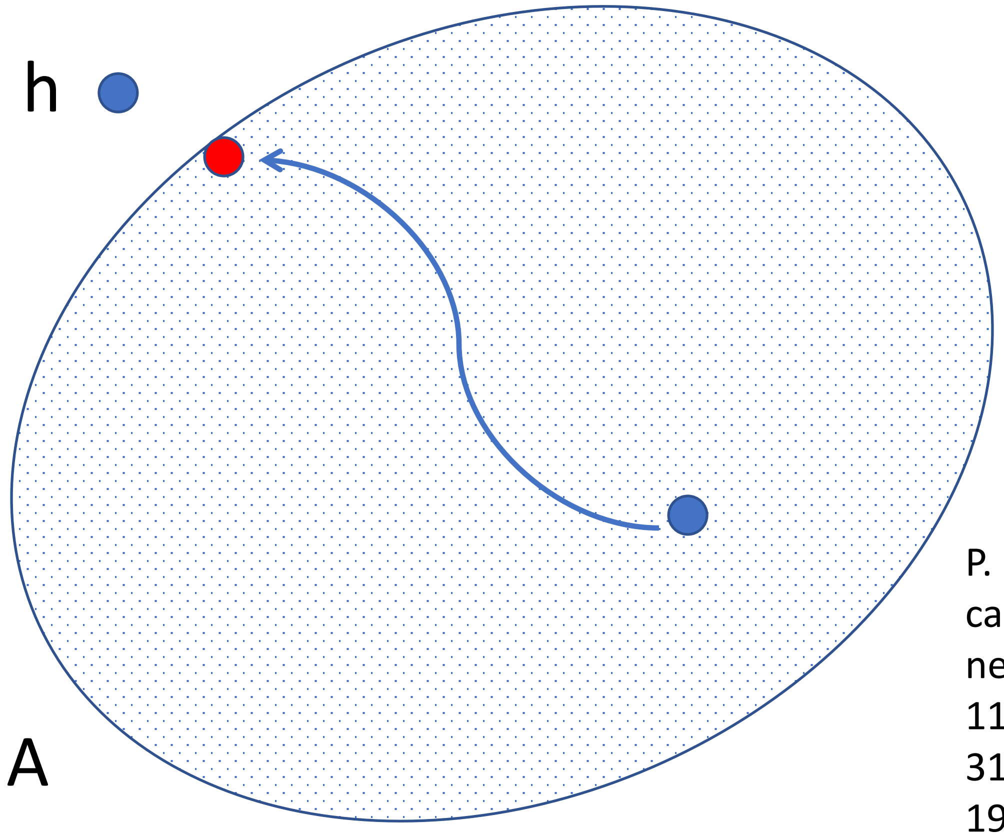


RoadMap

1. Introduction to Attention and the Standard Model
2. A Taxonomy of Attention Mechanisms (Quarks)
3. Transformers and Attention
4. Applications of Attention
5. Mathematical Theory of Attention
6. Large Language Models

Cardinal Capacity

- h = target function (typically known from examples)
- A = class of hypothesis or approximating functions (typically associated with a NN architecture)

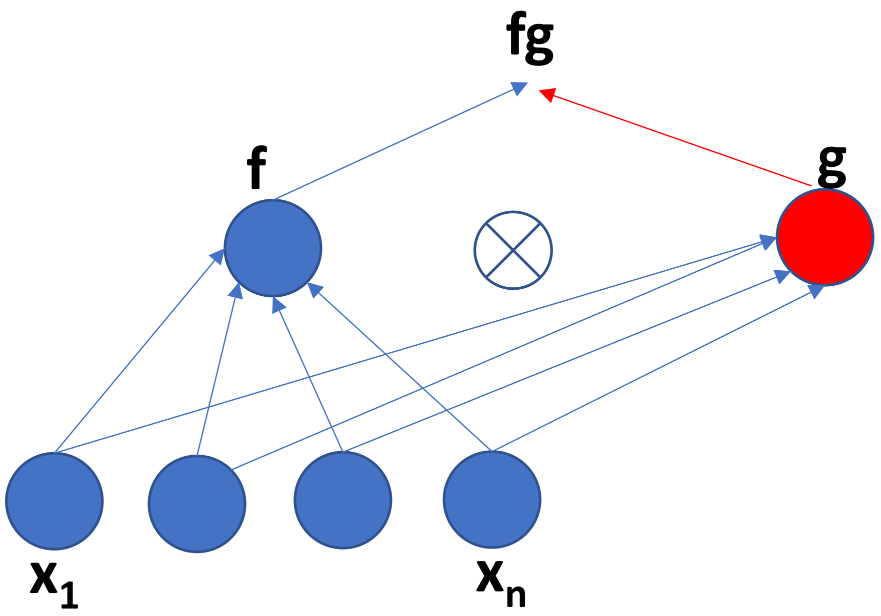


$$C(A) = \log_2 |A|$$

- Average number of bits required to specify a function in A.
- In a neural architecture, number of bits that must be transferred from the data to the synapses during learning

P. Baldi and R. Vershynin. The capacity of feedforward neural networks. Neural Networks, 116, August 2019, Pages 288-311, (2019). Also: Arxiv 1901.00434.

Single Linear (or Polynomial) Threshold Gate Output-Gated by Single Linear (or Polynomial) Threshold Gate

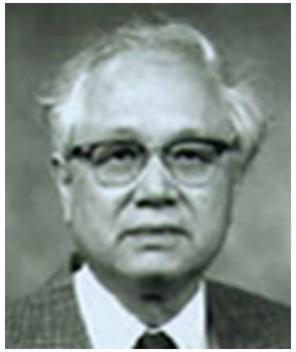


How many Boolean functions can we expressed as the product of two linear threshold functions?

- **Inputs can be 0/1 or -/+ (absorbed by affine transformation)**
- **Outputs are 0/1 fg= f AND g**
- **Outputs are -/+ fg = f NXOR g**

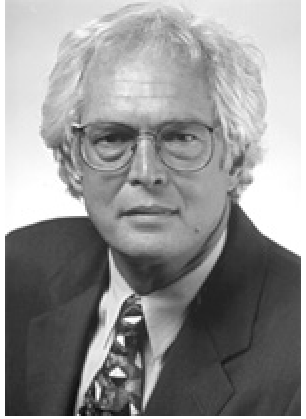
Capacity Of Linear Threshold Gates

$C(n,1) \leq n^2$



$0.5\, n^2 \leq C(n,1)$

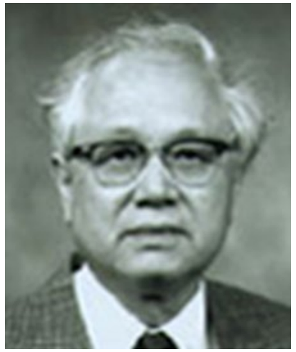
T. Cover 1965



S. Muroga (1965)

Capacity Of Linear Threshold Gates

$C(n,1) \leq n^2$



$0.5\, n^2 \leq C(n,1)$

T. Cover 1965

S. Muroga (1965)

$C(n,1) = n^2\,(1 + o(1))$

Yu. A. Zuev (1989)

Capacity Of Linear Threshold Gates

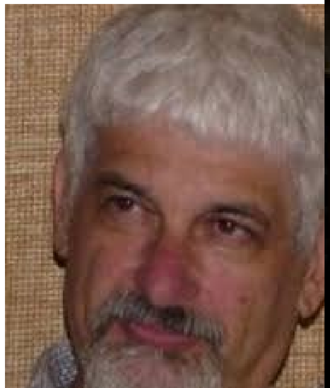
$$C(n,1) = n^2\,(1 + o(1))$$

- $\left(1 - \frac{10}{\log n}\right) n^2 \le C(n,1) \le n^2$

Yu. A. Zuev (1989)

- $C(n,1) = n^2 - n \log_2 n \pm O(n)$

Kahn, Komlos, Szemeredi (1994)



Capacity Of Polynomial Threshold Gates

$$C_d(n,1) \leq \frac{n^{d+1}}{d!}$$

P.B. 1988

Capacity Of Polynomial Threshold Gates

$$C_d(n,1) \leq \frac{n^{d+1}}{d!}$$

P.B. 1988

$$\binom{n}{d+1} \leq C_d(n,1)$$

M. Saks 1993

Capacity Of Polynomial Threshold Gates

$$C_d(n,1) \leq \frac{n^{d+1}}{d!}$$

P.B. 1988

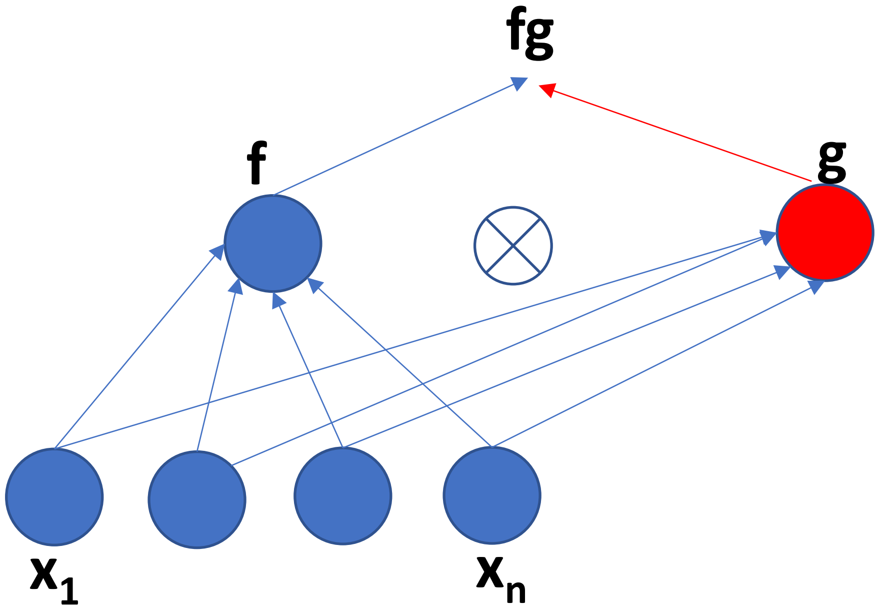
$$\binom{n}{d+1} \leq C_d(n,1)$$

M. Saks 1993

$$C(n,d) = \frac{n^{d+1}}{d!}(1 + o(1))$$

P. Baldi and R. Vershynin. Polynomial threshold functions, hyperplane arrangements, and random tensors. SIAM Journal on Mathematics of Data Science (SIMODS), 1, 4, 699-729, URL: https://epubs.siam.org/toc/sjmdaq/1/3 , DOI: 10.1137/19M1257792, (2019).

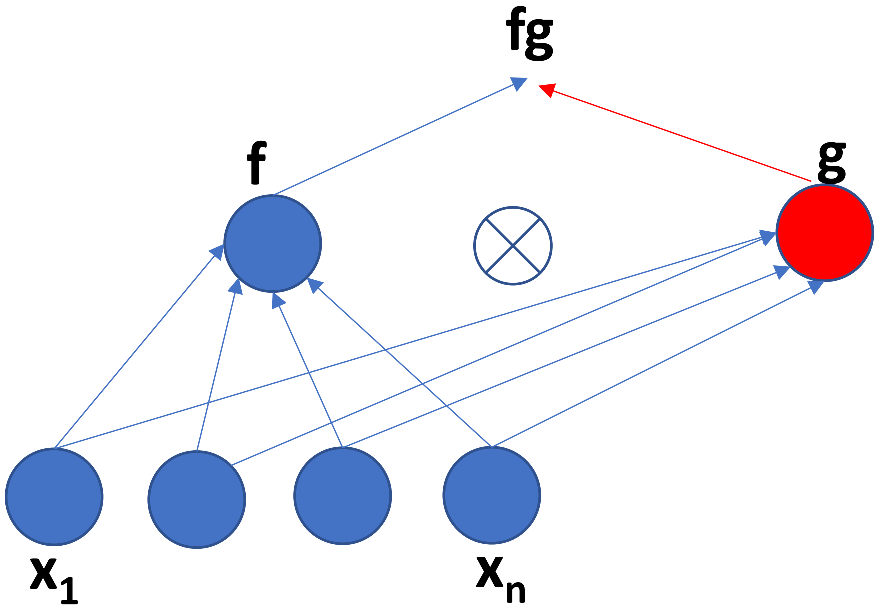
Single Linear (or Polynomial) Threshold Gate Output-Gated by Single Linear (or Polynomial) Threshold Gate



How many Boolean functions can we expressed as the product of two linear threshold functions?

- **Inputs can be 0/1 or -/+ (absorbed by affine transformation)**
- **Outputs are 0/1   fg= f AND g**
- **Outputs are -/+   fg = f NXOR g**

Single Linear (or Polynomial) Threshold Gate Output-Gated by Single Linear (or Polynomial) Threshold Gate



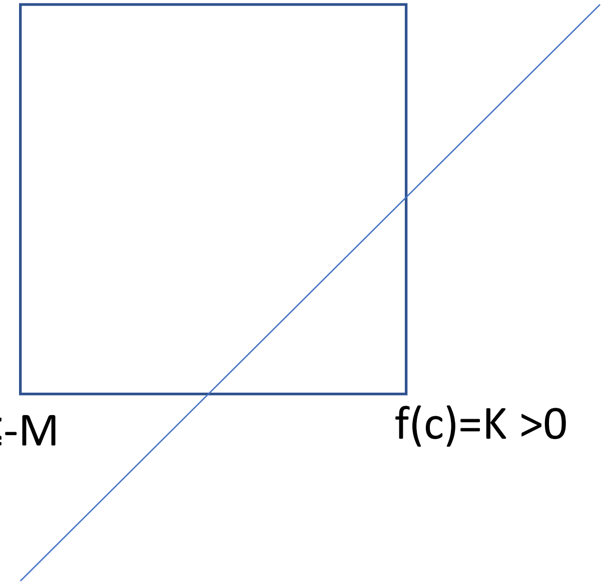
How many Boolean functions can we expressed as the product of two linear threshold functions?

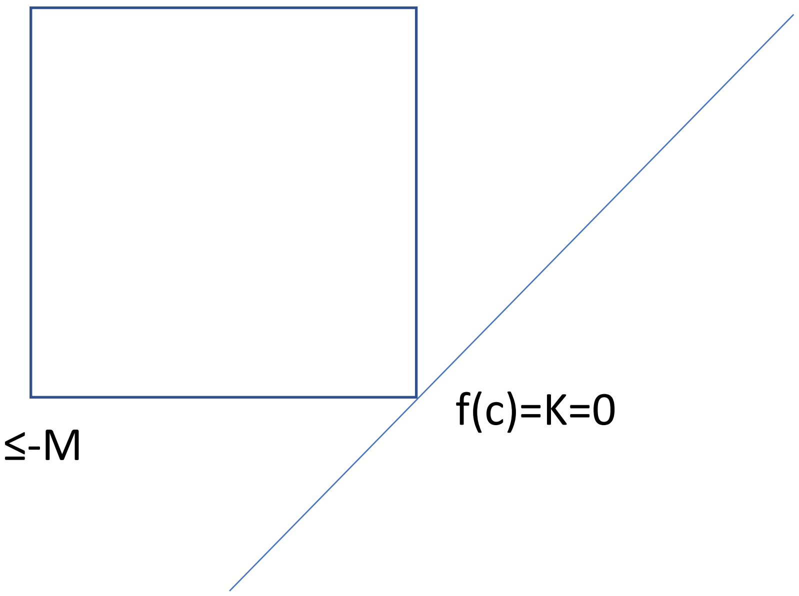
- Inputs can be 0/1 or -/+ (absorbed by affine transformation)
- Outputs are 0/1 fg= f AND g
- Outputs are -/+ fg = f NXOR g

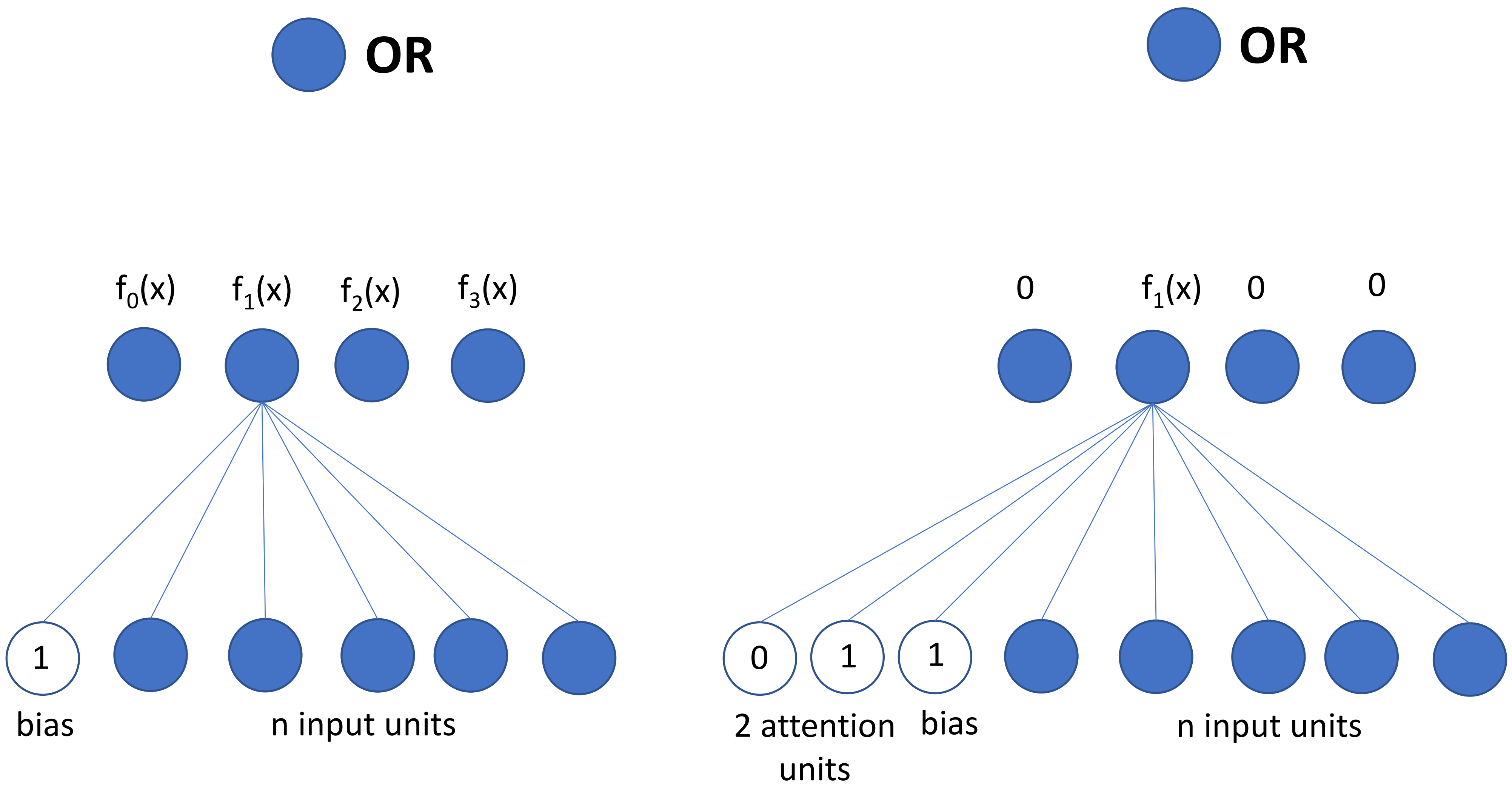
Answer: $2n^2(1+o(1))$ upperbound easy; lower bound?

Simple Lemma

Any corner of the hypercube can be isolated from all the other corners by an affine hyperplane f with large margins (for any 0<M; 0≤ K).

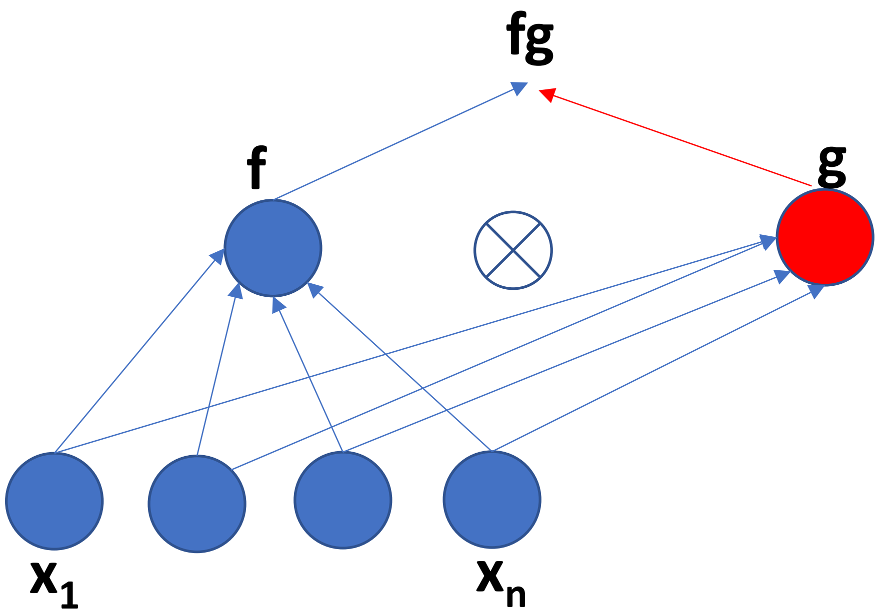






Multiplexing (=Activation Attention)

Single Linear (or Polynomial) Threshold Gate Output-Gated by Single Linear (or Polynomial) Threshold Gate



- Inputs can be 0/1 or -/+ (absorbed by affine transformation)
- Outputs are 0/1 fg= f AND g
- Outputs are -/+ fg = f NXOR g

|f AND g|= |f OR g|

| f XOR g|= | f NXOR g |

Linear Threshold Case:

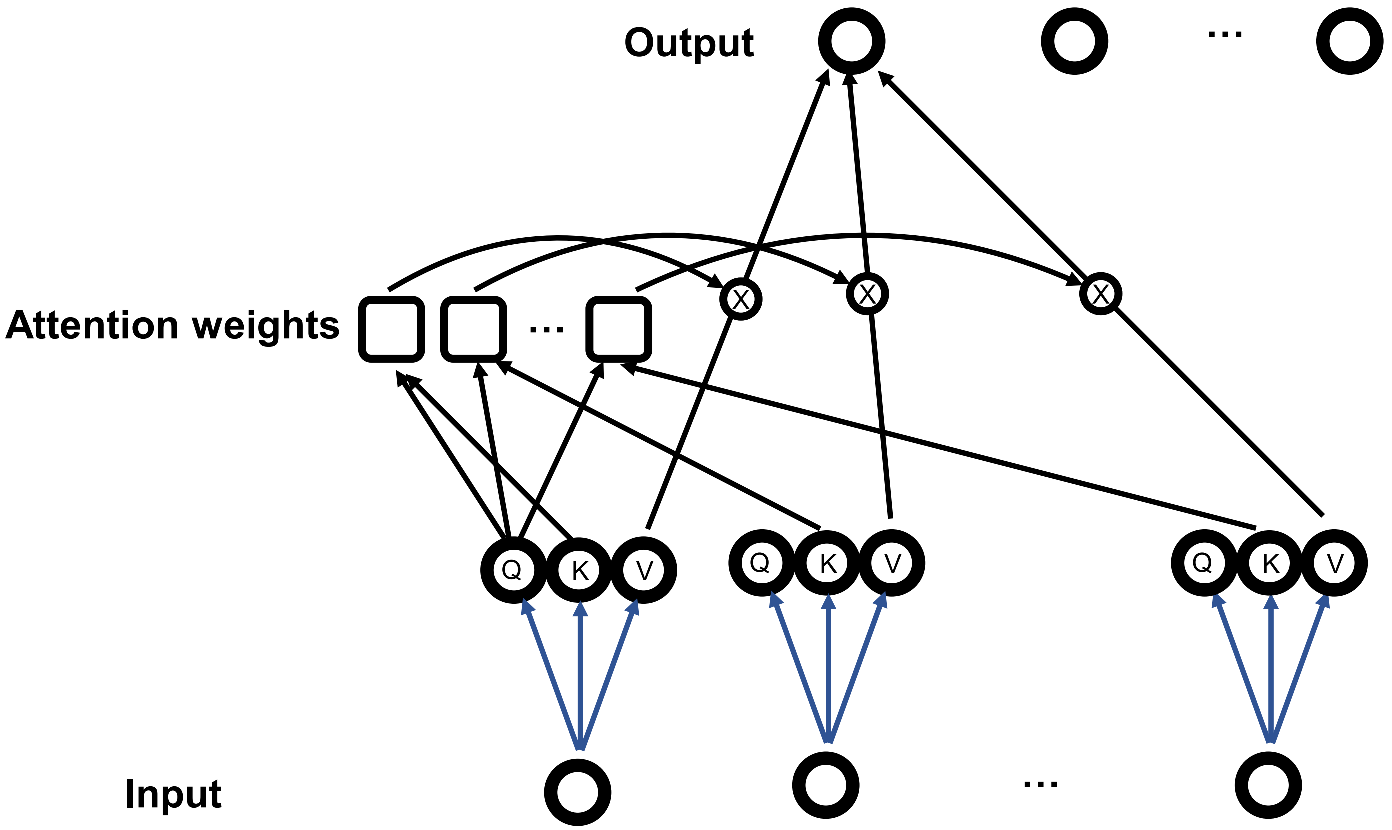
Capacity is equal to $2n^2$ (1+o(1)) (d=1)

Polynomial Threshold Case:

Capacity is equal to $[2n^{d+1}/d!](1+o(1))$ (d>1)

- Gating is a computationally efficient mechanism for tapping into quadratic activation functions in a sparse way
- Much more work is needed to better understand transformers

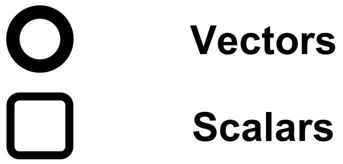
All Q-K combinations



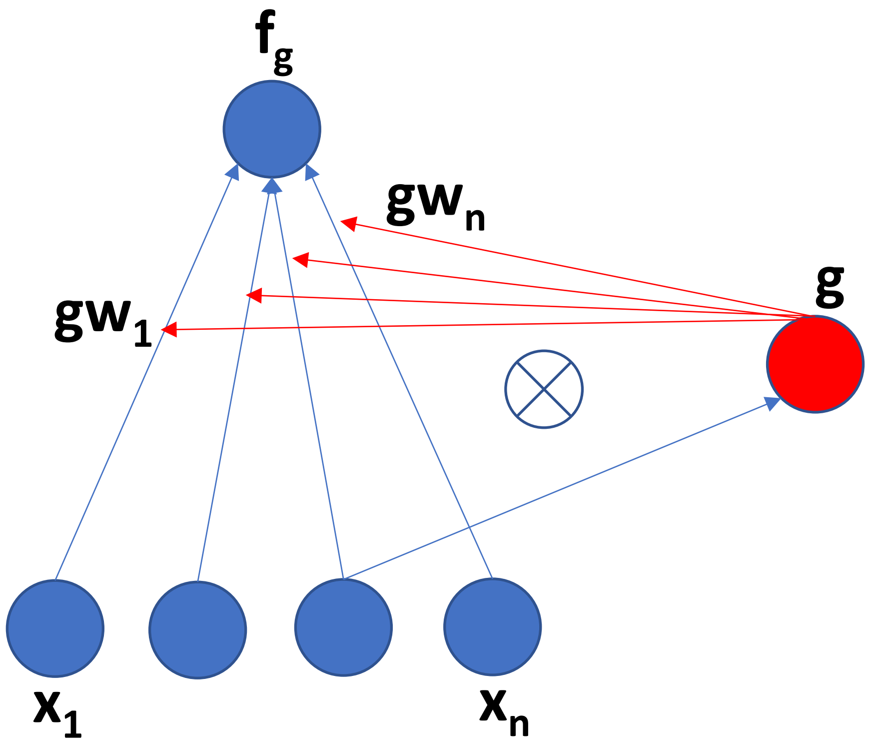
Query
Key
Value

Weight
sharing

Fundamental Property: Invariant to permutation of the inputs (!).



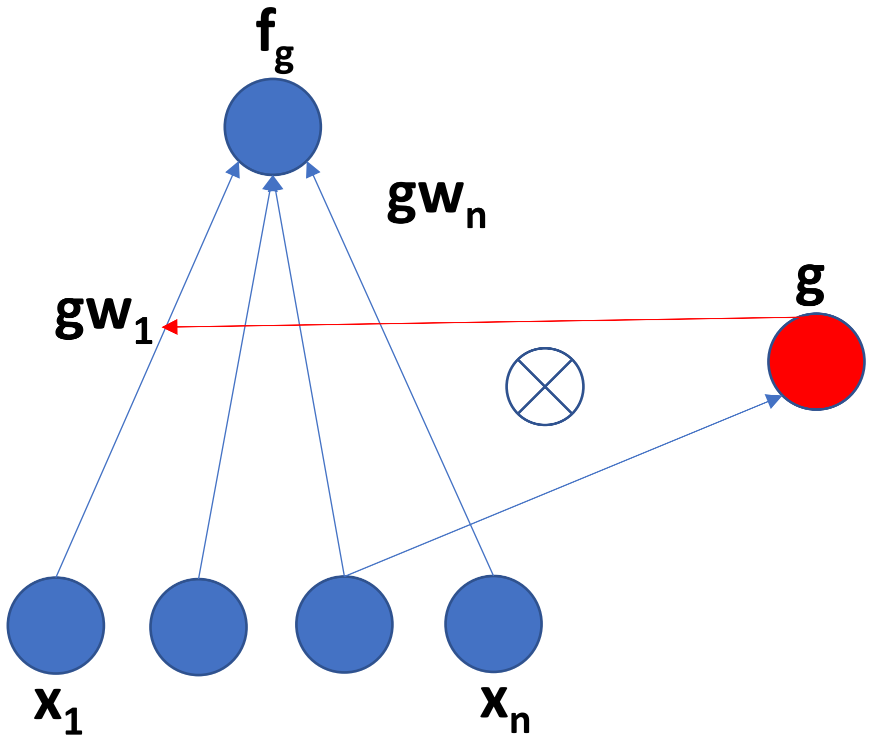
Single Linear (or Polynomial) Threshold Gate Synaptically-Gated by Single Linear (or Polynomial) Threshold Gate (all synapses)



Linear Threshold Case:
Capacity is equal to $2n^2(1+o(1))$ (d=1)

Polynomial Threshold Case:
Capacity is equal to $[2n^{d+1}/d!](1+o(1))$ (d>1)

Single Linear (or Polynomial) Threshold Gate Synaptically-Gated by Single Linear (or Polynomial) Threshold Gate (one synapse)



Linear Threshold Case:

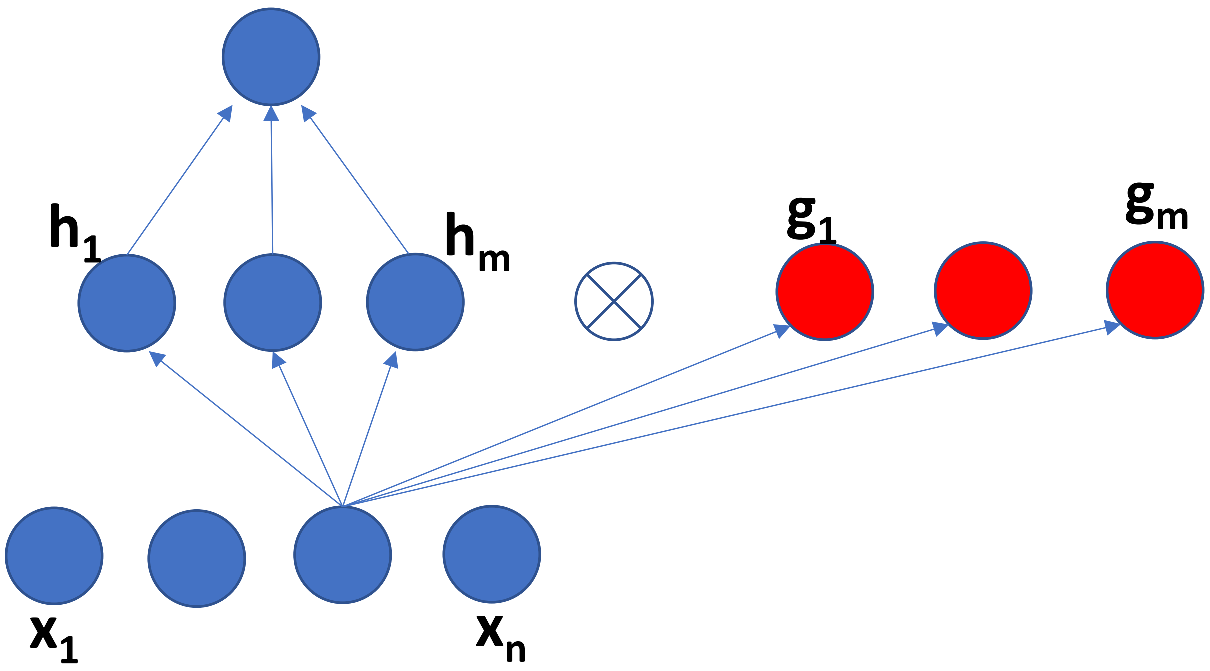
$n^2(1+o(1)) \leq C \leq 2n^2 (1+o(1))$ (d=1)

Conjecture: closer to n^2

Polynomial Threshold Case:

$[n^{d+1}/d!](1=o(1)) \leq C \leq 2 [n^{d+1}/d!](1=o(1))$

Layer of Linear (or Polynomial) Threshold Gates Output-Gated by a Layer of Linear (or Polynomial) Threshold Gates



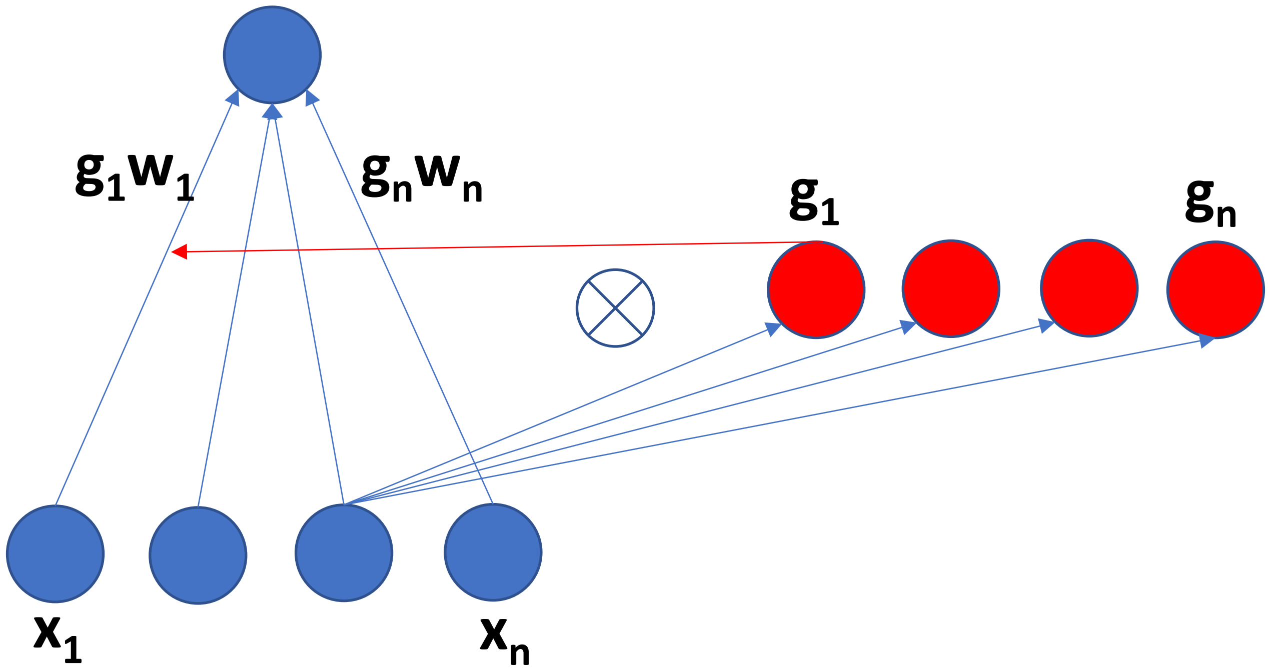
Linear Threshold Case:

$2mn^2\,(1+o(1))$ (d=1)

Polynomial Threshold Case:

$2m[n^{d+1}/d!](1+o(1))$ (d>1)

Linear (or Polynomial) Threshold Gate Synaptically-Gated by a Layer of Linear (or Polynomial) Threshold Gates



Linear Threshold Case:

$n^2(1+o(1)) \leq C \leq n^3\,(1+o(1)) \quad (d=1)$

Polynomial Threshold Case:

$[n^{d+1}/d!](1=o(1)) \leq C \leq 2\,[n^{d+2}/d!](1=o(1)) \quad (d>1)$

RoadMap

1. Introduction to Attention and the Standard Model
2. A Taxonomy of Attention Mechanisms (Quarks)
3. Transformers and Attention
4. Applications of Attention
5. Mathematical Theory of Attention
6. Large Language Models (LLMs)

LLM Technology

- Autoregressive generative models
- Current generation is mostly based on transformers
- Language is tokenized and place encoded
- LLM: Trained in self-supervised mode to predict the next word, i.e. the next token
- Softmax output= distribution over the vocabulary of tokens.
- At production time, sample from the distribution. Greedy sampling does not work well. Usually, top k sampling is used.
- Initially trained on text alone
- Potentially trained on entire humanity's knowledge (far more than any individual human)
- Quality of training data matters
- Running out of data. LLM generated data. Distillation issues.

LLM Technology

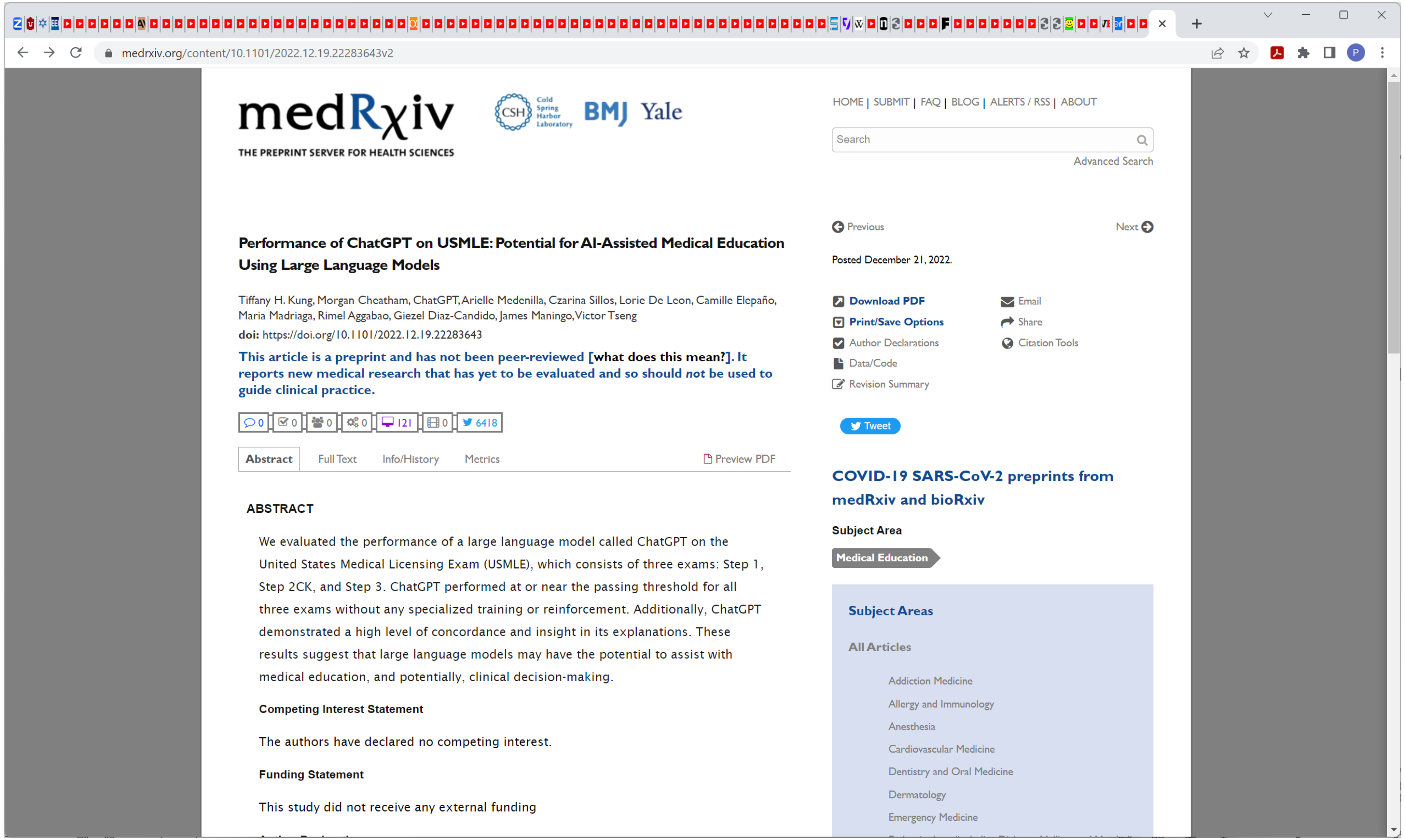
- Training the Base Model
- Aligning the Base Model. Post-training.
- Supervised post-training, RLHF (Reinforcement Learning from Human Feedback)
- Prompt engineering. System prompt.
- Inference with the Base Model
- Reasoning with the Base Mode
- To a first order of approximation: reasoning = rumination.
- Multi-modal versions are now common
- Can be interfaced with other programs, agents, and robots

LLM Landscape

- Many different LLMsmodels: GPT, CLAUDE, GEMINI, GROK, DEEPSEEK, LLAMA, MISTRAL, etc
- Available with different flavors, sizes, reasoning capabilities.
- Available under a subscription model or as “open weights” model (open weight is not the same thing as open source)
- Initially trained on text alone. Currently multimodal version are common.

LLM Capabilities

- Capable of conversing, translating, programming, etc.
- Can make errors and hallucinate
- Many benchmarks—Humanity Last Exam



Other example of application: Pharmacy Automation

ORIGINAL CLINICAL RESEARCH REPORT

Clinical Knowledge and Reasoning Abilities of AI Large Language Models in Anesthesiology: A Comparative Study on the American Board of Anesthesiology Examination

Mirana C. Angel, MSc,*† Joseph B. Rinehart, MD,‡ Maxime P. Cannesson, MD, PhD,§ and Pierre Baldi, PhD*†

BACKGROUND: Over the past decade, artificial intelligence (AI) has expanded significantly with increased adoption across various industries, including medicine. Recently, AI-based large language models such as Generative Pretrained Transformer-3 (GPT-3), Bard, and Generative Pretrained Transformer-3 (GPT-4) have demonstrated remarkable language capabilities. While previous studies have explored their potential in general medical knowledge tasks, here we assess their clinical knowledge and reasoning abilities in a specialized medical context.

METHODS: We studied and compared the performance of all 3 models on both the written and oral portions of the comprehensive and challenging American Board of Anesthesiology (ABA) examination, which evaluates candidates' knowledge and competence in anesthesia practice.

RESULTS: Our results reveal that only GPT-4 successfully passed the written examination, achieving an accuracy of 78% on the basic section and 80% on the advanced section. In comparison, the less recent or smaller GPT-3 and Bard models scored 58% and 47% on the basic examination, and 50% and 46% on the advanced examination, respectively. Consequently, only GPT-4 was evaluated in the oral examination, with examiners concluding that it had a reasonable possibility of passing the structured oral examination. Additionally, we observe that these models exhibit varying degrees of proficiency across distinct topics, which could serve as an indicator of the relative quality of information contained in the corresponding training datasets. This may also act as a predictor for determining which anesthesiology subspecialty is most likely to witness the earliest integration with AI.

CONCLUSIONS: GPT-4 outperformed GPT-3 and Bard on both basic and advanced sections of the written ABA examination, and actual board examiners considered GPT-4 to have a reasonable possibility of passing the real oral examination; these models also exhibit varying degrees of proficiency across distinct topics. (Anesth Analg 2024;139:349–56)

KEY POINTS

- **Question:** How might recent advancements in artificial intelligence (AI) large language models influence the field of anesthesiology?
- **Findings:** Large language models may now be sophisticated enough to pass the anesthesiology written and oral examinations.
- **Meaning:** The rapid development of these models holds the potential to shape the future of both anesthesiology education and practice, but we need to be aware of their limitations.

In recent years, artificial intelligence (AI) primarily in the form of machine learning, in particular deep learning, has experienced a significant expansion driven by progress in computational power and big data availability.[1] In the medical field, AI's potential to increase accuracy and expedite diagnoses has led to its application in numerous areas, including radiology, pathology, and genomics. For example, AI has

From the *Department of Computer Science, University of California Irvine, Irvine, California; †Institute for Genomics and Bioinformatics, University of California Irvine, Irvine, California; ‡Department of Anesthesiology & Perioperative Care, University of California Irvine, Irvine, California; and §Department of Anesthesiology & Perioperative Medicine, University of California Los Angeles, Los Angeles, California.

Accepted for publication November 27, 2023.

DOI: 10.1213/ANE.0000000000006892

Funding: This work was supported by NIH R01EB029751 (to M.P.C, P.B., and J.B.R.).

The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.anesthesia-analgesia.org).

Reprints will not be available from the authors.

Address correspondence to Pierre Baldi, PhD, Department of Computer Science, University of California Irvine, Irvine, CA 92697. Address e-mail to pfbaldi@ics.uci.edu.

AI and Veterinary Medicine: Performance of Large Language Models on the North American Licensing Examination

Mirana Angel
Institute for Geomics and Bioinformatics
University of California Irvine
Irvine, USA
mcangel@uci.edu

Anuj Patel
Department of Computer Science
University of California Irvine
Irvine, USA
patelad2@uci.edu

Haiyi Xing
Department of Computer Science
University of California Irvine
Irvine, USA
haiyix2@uci.edu

Dylan Balsz
Internal Medicine
Anivive Life Sciences
Long Beach, USA
dylan@anivive.com

Cody Arbuckle
Internal Medicine
Anivive Life Sciences
Long Beach, USA
cody@anivive.com

David Bruyette
Internal Medicine
Anivive Life Sciences
Long Beach, USA
david@anivive.com

Pierre Baldi
Department of Computer Science
University of California Irvine
Irvine, USA
pfbaldi@uci.edu

2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS) | 979-8-3503-1890-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/SNAMS60348.2023.10375414

***Abstract*—This study aimed to assess the performance of Large Language Models on the North American Veterinary Licensing Examination (NAVLE) and to analyze the impact of artificial intelligence in the domain of animal healthcare. For this study, a 200-question NAVLE self-assessment sourced from ICVA's website was used to evaluate the performance of three language models: GPT-3, GPT-4, and Bard. Questions involving images were omitted leaving a 164 text-only sample exam. Results were analyzed by comparing generated responses to the answer key, and scores were assigned to evaluate the models' veterinary medical reasoning capabilities. Our results showed that GPT-4 outperformed GPT-3 and Bard, passing the exam with 89 % of the text-only questions correctly. GPT-3 and Bard only achieved an accuracy of 63.4 % and 61 % respectively on the same set of questions. Language models hold promise for enhancing veterinary practices through expanded educational opportunities in the veterinary curriculum, improved diagnostic accuracy, treatment times, and efficiency. However, potential negatives include challenges in changing the current educational paradigm, reduced demand for professionals or paraprofessional concerns surrounding machine-generated decisions. Responsible and ethical integration of language models is crucial in veterinary medicine.**

***Index Terms*—Artificial Intelligence, LLM, ChatGPT, Bard, Veterinary Medicine, Medical Education, Societal Impact**

I. INTRODUCTION

In recent years, the rapid growth of artificial intelligence (AI) has significantly influenced various industries, including healthcare. The development of increasingly powerful AI models, such as large language models (LLMs) has facilitated the automation of diverse tasks and the enhancement of decision-making processes. Consequently, the adoption of AI technology has emerged as a pivotal factor in gaining a competitive edge and boosting efficiency across industries [1]. Here we provide an initial assessment of the applicability of LLMs in veterinary medicine by testing their ability to pass a standard veterinary education test.

The veterinary field encompasses a wide array of professions and specializations, all dedicated to the care and well-being of animals. Veterinarians, who are extensively trained to diagnose and treat various conditions in numerous species ranging from domesticated animals and livestock to wildlife, are a cornerstone of this field. As the veterinary field continues to evolve, new technologies and techniques are revolutionizing the diagnosis and treatment of animal health issues [2].

The advent of diverse AI technologies, such as state-of-the-art text, sound, image, and video data analysis algorithms, have significantly advanced veterinary medicine in areas such as disease diagnosis, treatment planning, and precision medicine [2, 3, 4]. However, current AI models are typically task-specific and lack the capability for independent medical reasoning [5]. This limitation has prompted researchers to explore the potential of large language models, which have demonstrated remarkable cognitive reasoning abilities, in addressing these shortcomings in all fields.

Among large language models, Generative Pre-trained Transformer (GPT) and Bard have emerged as frontrunners, exhibiting outstanding performance in various applications [6, 7, 8]. GPT-3 and GPT-4, as well as Bard, adopt the decoder-only architecture of the transformer model [9]. GPT-3 encompasses 175 billion parameters and showcases remarkable versatility across a range of tasks. In an advancement over GPT-3, GPT-4 boasts an unprecedented one trillion parameters, addressing many of the limitations previously associated with GPT-3. Both GPT iterations were pre-trained on extensive text corpora and subsequently fine-tuned for specialized tasks [6, 7].

Concurrently, Google's Bard initially employed the Lan-

Clinical Knowledge and Reasoning Abilities of AI Large Language Models in Anesthesiology: A Comparative Study on the ABA Exam

Mirana C. Angel MSc[1,2], Joseph B. Rinehart MD[3], Maxime P. Canneson MD PhD[4], Pierre Baldi PhD[1,2,*]

1. Department of Computer Science, University of California Irvine, Irvine, CA 92697, USA
2. Institute for Genomics and Bioinformatics, University of California Irvine, Irvine CA 92697, USA
3. Department of Anesthesiology & Perioperative Care, University of California Irvine, Irvine CA 92697, USA
4. Department of Anesthesiology & Perioperative Medicine, University of California, Los Angeles, Los Angeles, CA 90095

ELSEVIER

Computers in Human Behavior: Artificial Humans

Volume 5, August 2025, 100170

Research paper

Evaluating the Intelligence of large language models: A comparative study using verbal and visual IQ tests

Sherif Abdelkarim [a] [1], David Lu [a] [1], Dora-Luz Flores [c], Susanne Jaeggi [a] [b], Pierre Baldi [a]

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.chbah.2025.100170

Get rights and content

Open access

Highlights

- Evaluated cognitive performance of popular LLMs using verbal and visual IQ tests.
- Found a positive correlation between LLM size and cognitive performance across tasks.
- Significant performance variability across problem types suggests nuanced differences in reasoning.

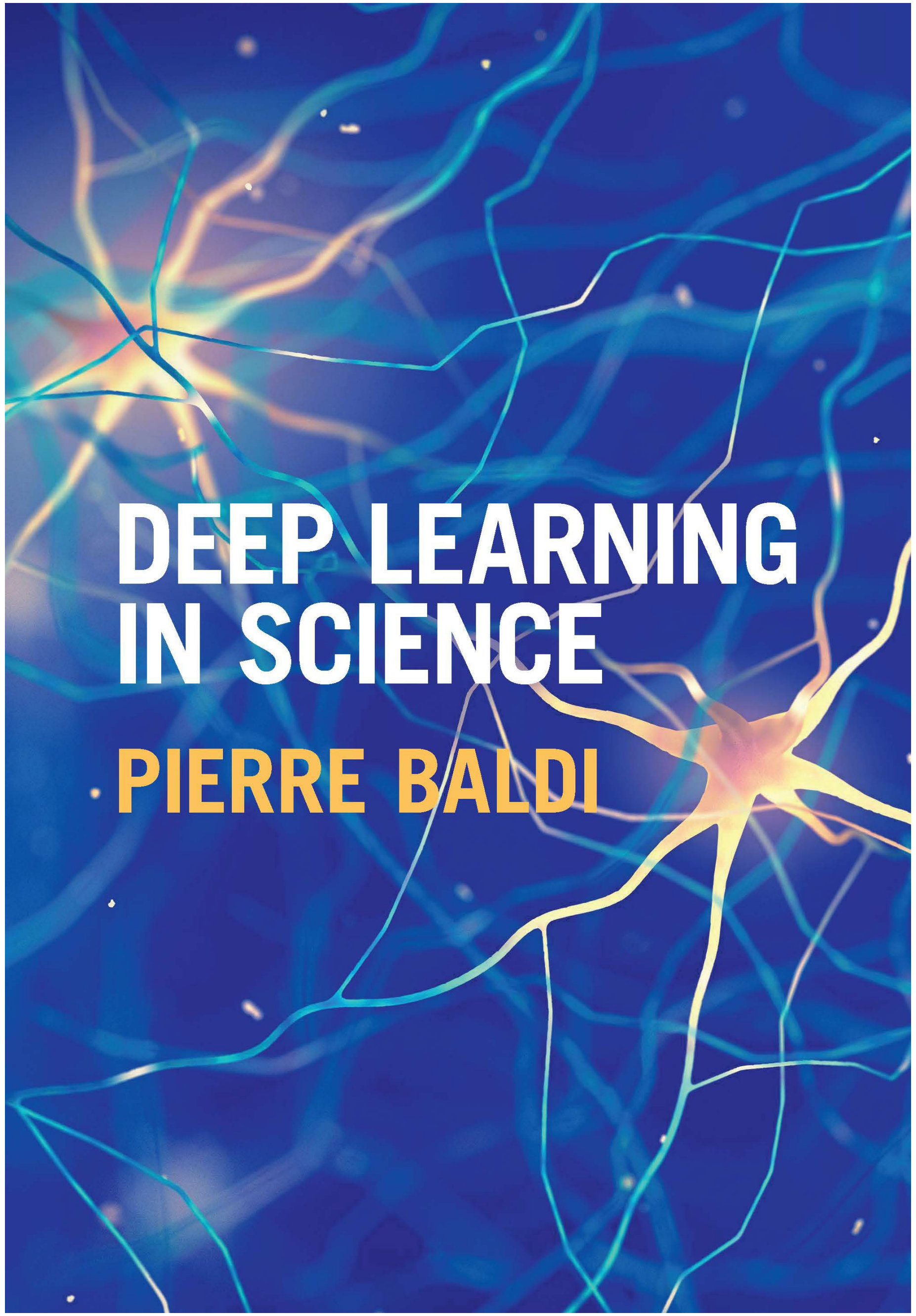
LLM Capabilities

- More or less pass the Turing test
- Recently: >50% on HLE (GROK); gold medal math Olympiad (Gemini; GPT)
- Can LLM achieve AGI? SI?
- Argument against: stochastic parrots, no knowledge of the real world.
- Argument for: keep rapidly beating all benchmarks. The case of Helen Keller…

Conclusion

- Taxonomy of elementary building blocks for attention
- Output gating and synaptic gating extend the SM towards the space of quadratic activations without incurring the full cost
- Output gating and synaptic gating are used in all the existing attention based architectures, including transformers (output gating alone is enough)
- Transformer have permutation invariance properties which are attractive for applications beyond NLP (physics, chemistry)
- Mathematical theory of attention capacity (efficient mechanism to tap into quadratic activations)
- LLMs pass the Turing test

THANK YOU



Cambridge University Press
TOC and sample chapters on
my web site.