

# Attention Is All You Need ("Transformers")

Michael Schatz  
October 21, 2024  
Applied Comparative Genomics



# Class Schedule

M	Oct 14	Epigenome	Project Proposal Assigned
W	Oct 16	Single cell	
M	Oct 21	Transformers	Assignment 5 Assigned
W	Oct 23	Enformer	
M	Oct 28	DL in Genomics	Preliminary Report Assigned
W	Oct 30	Midterm Review	
M	Nov 4	Midterm!	
W	Nov 6	Disease Genomics	
M	Nov 11	Metagenomics	Final Report Assigned
W	Nov 13	No Class (BIODATA24)	
M	Nov 18	Cancer Genomics	
W	Nov 20	Project Presentation 1	
M	Nov 25	Thanksgiving Break	
W	Nov 27	Thanksgiving Break	
M	Dec 2	Project Presentation 2	
W	Dec 4	Project Presentation 3	
M	Dec 16	Project Report Due	



# Project Proposal

## Due: Monday Oct 21, 2024 by 11:59pm

### Project Proposal

---

Assignment Date: Monday October 14, 2024

Due Date: Monday, October 21 2024 @ 11:59pm

Review the [Project Ideas](#) page

Work solo or form a team for your class project of no more than 3 people.

The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to 2 to 3 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)
- Please add a note if you need me to sponsor you for a MARCC account (high RAM, GPUs, many cores, etc)
- Please add a note if you need me to sponsor you for an AnVIL cloud billing account (dynamic resources)

Submit the proposal as a 1 to 2 page PDF on GradeScope (each team should submit one proposal and tag all people in the team). After submitting your proposal, I will provide feedback. If necessary, we can schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting.

Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at [https://academic.oup.com/bioinformatics/pages/submission\\_online](https://academic.oup.com/bioinformatics/pages/submission_online)

Please use Piazza to coordinate proposal plans!

<https://github.com/schatzlab/appliedgenomics2024/blob/main/project/proposal.md>

Check Piazza for questions!

# Assignment 5

## Due: Monday Oct 28, 2024 by 11:59pm

### Assignment 5

First, run the below cells to ensure you install the required dependencies

```
In [ ]: !pip install kipoiseq==0.5.2
        !pip install logomaker

In [ ]: !pip install enformer-pytorch

In [ ]: from enformer_pytorch import from_pretrained
        from enformer_pytorch.finetune import HeadAdapterWrapper
        import kipoiseq
        from kipoiseq import Interval
        import matplotlib.pyplot as plt
        import numpy as np
        import pandas as pd
        import pyfaidx
        import seaborn as sns
        import torch
        import torch.nn as nn
        import tqdm
        from transformers import GPT2Model, GPT2Tokenizer
        import seaborn as sns
        import logomaker
        import torch.nn.functional as F
        from sklearn.preprocessing import StandardScaler, LabelEncoder
        from sklearn.decomposition import PCA
        import torch.optim as optim
        from torch.utils.data import DataLoader, TensorDataset
        from sklearn.metrics import classification_report

In [ ]: device = "cuda" if torch.cuda.is_available() else "cpu"
```

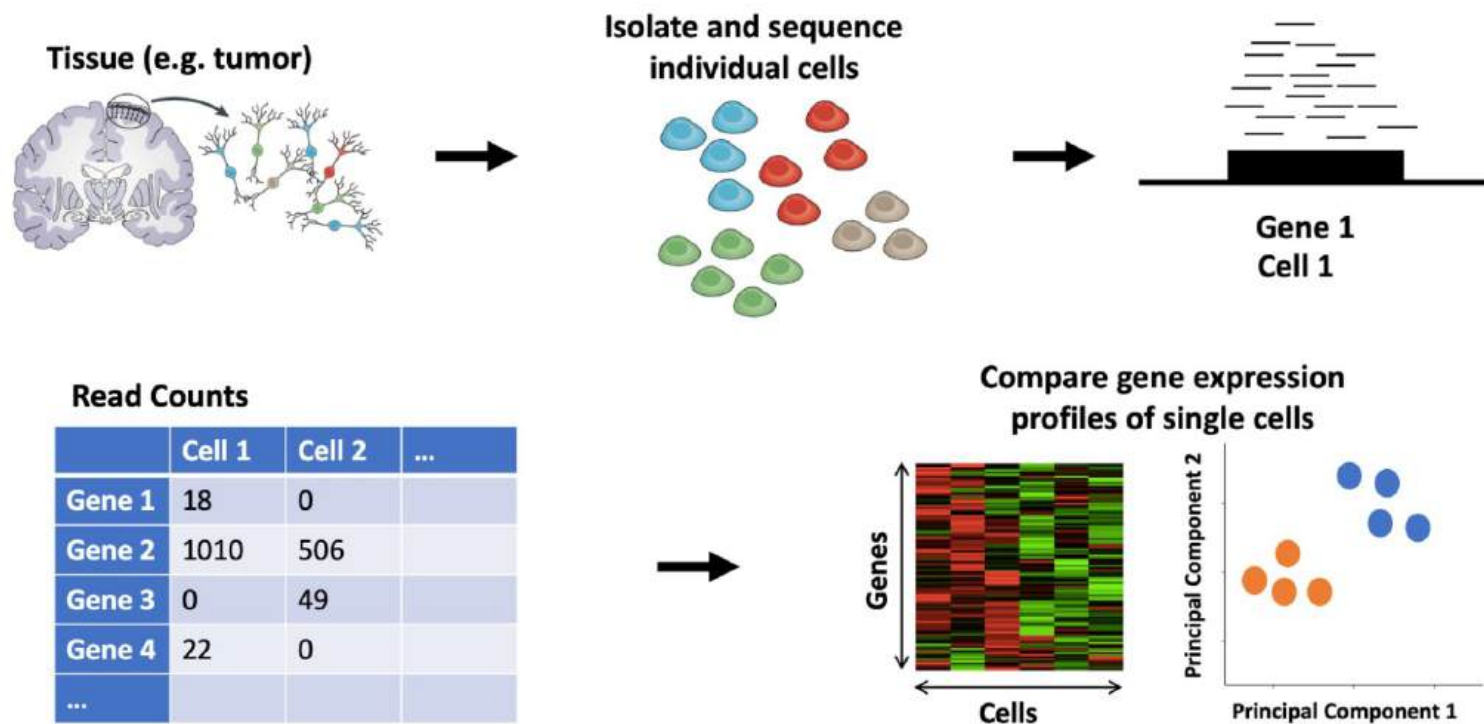
### Question 1: Self-attention

Self-attention allows a model to weigh the importance of different tokens in a sequence relative to each other. This allows it to capture dependencies across the entire input, so the model can learn both local and long-range relationships. In this question, you will take a look at how self-attention works for natural language by using GPT-2 and visualizing the attention weights as a heatmap.

<https://schatz-lab.org/appliedgenomics2024/assignments/assignment5/>

Check Piazza for questions!

# Single-cell RNA-sequencing (scRNA-seq)



# ARTICLE

doi:10.1038/nature11247

## An integrated encyclopedia of DNA elements in the human genome

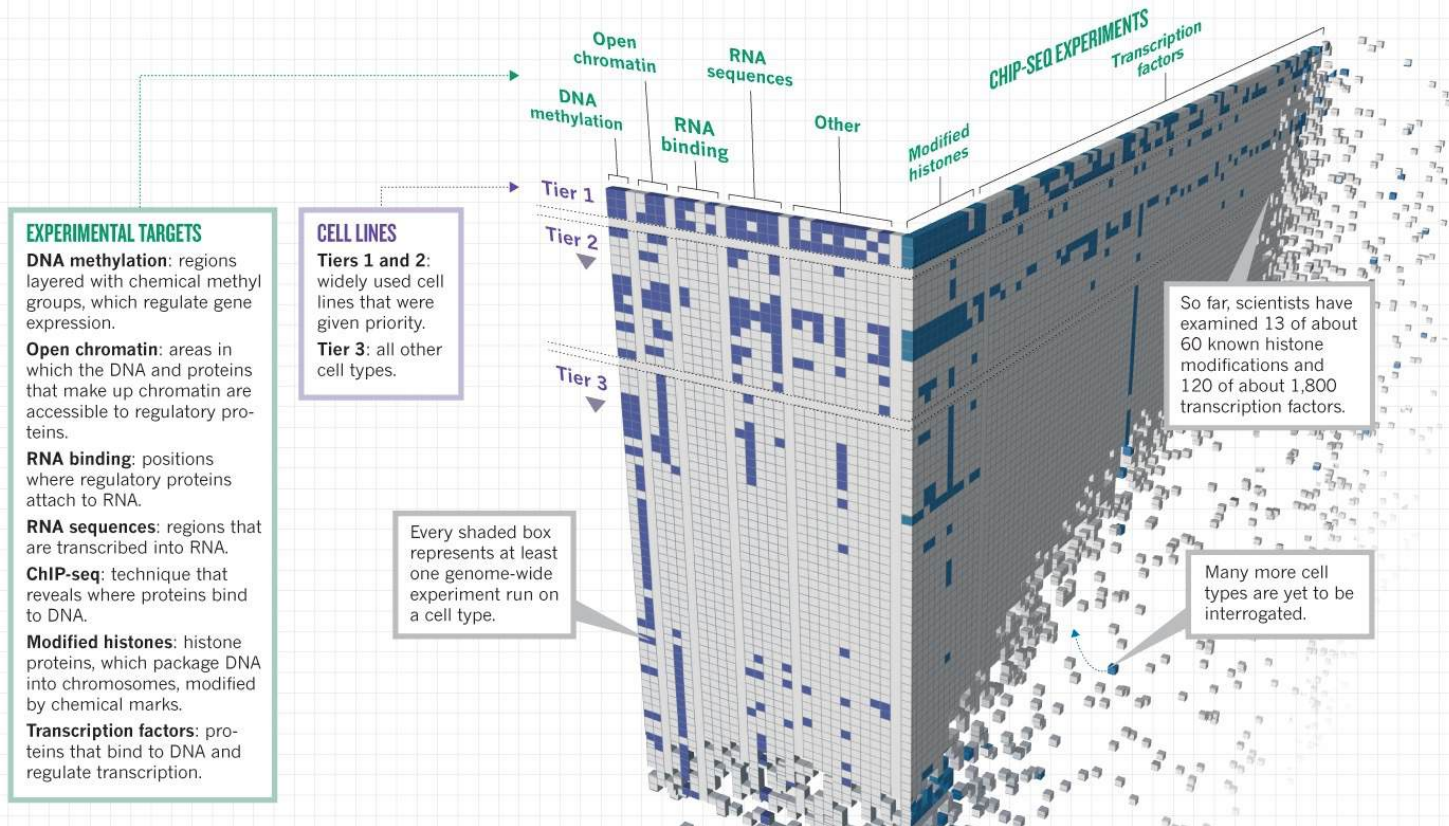
The ENCODE Project Consortium\*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

# ENCODE Data Sets

## MAKING A GENOME MANUAL

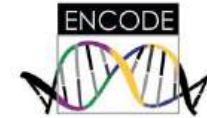
Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.



***1,640 data sets total over 147 different cell types***



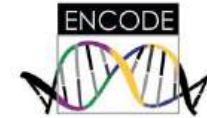
# Major Findings



1. *The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.*
2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*
3. *Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.*
4. *It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.*
5. *Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.*
6. *Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.*

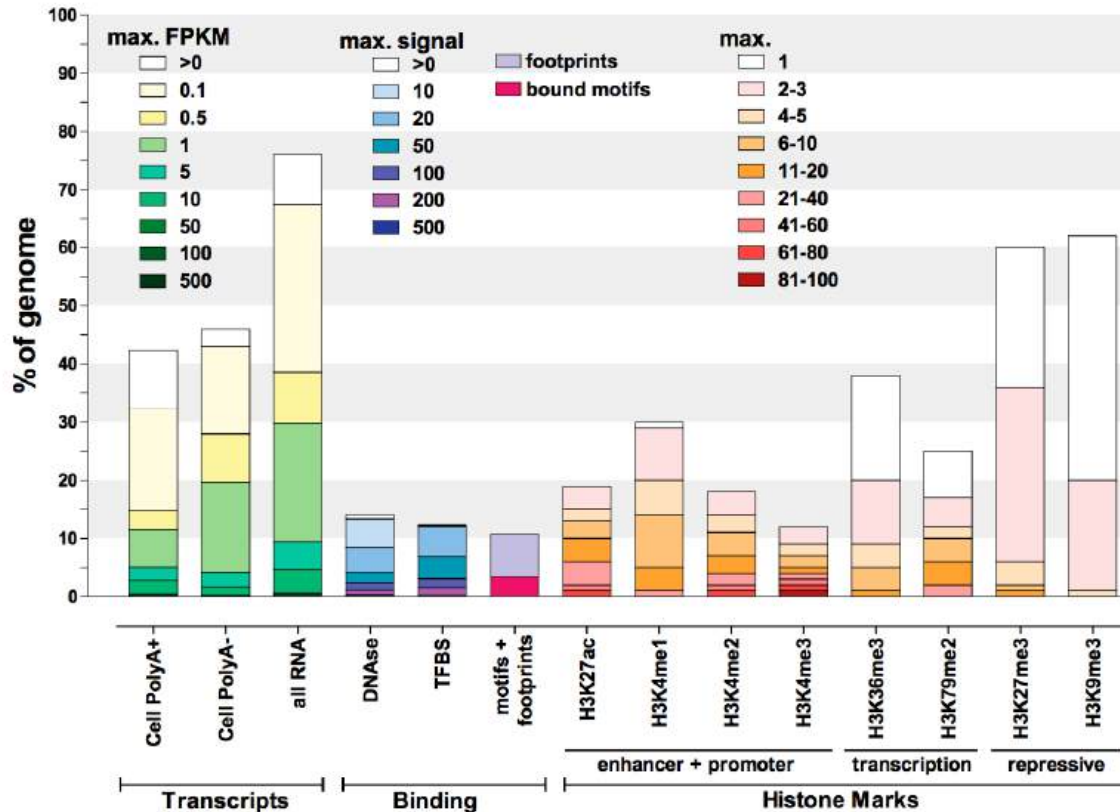


# Major Findings



1. ***The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.***
2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*
3. *Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.*
4. *It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.*
5. *Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.*
6. *Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.*

# Pervasive Transcription and Regulation



*“Accounting for all these elements, a surprisingly large amount of the human genome, 80.4%, is covered by at least one ENCODE-identified element”*

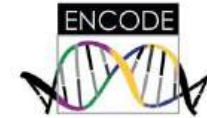
- 62% transcribed
- 56% enriched for histone marks
- 15% open chromatin
- 8% TF binding
- 19% At least one DHS or TF Chip-seq peak
- 4% TF binding site motif
- (Note protein coding genes comprise ~2.94% of the genome)

*“Given that the ENCODE project did not assay all cell types, or all transcription factors, and in particular has sampled few specialized or developmentally restricted cell lineages, **these proportions must be underestimates of the total amount of functional bases.**”*

## Defining functional DNA elements in the human genome

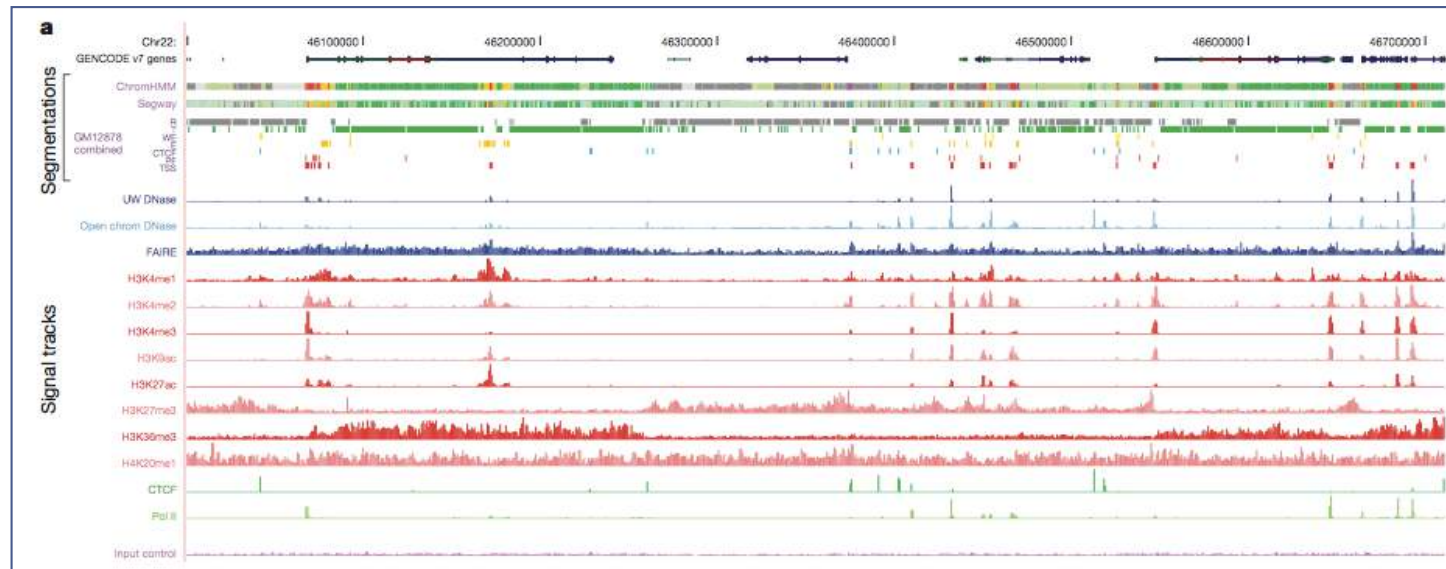
Kellis et al (2014). *PNAS* 6131–6138, doi: 10.1073/pnas.1318948111

# Major Findings



1. *The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.*
2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*
3. ***Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.***
4. *It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.*
5. *Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.*
6. *Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.*

# Signal Integration

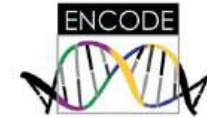


**Table 3 | Summary of the combined state types**

Label	Description	Details*	Colour
CTCF	CTCF-enriched element	Sites of CTCF signal lacking histone modifications, often associated with open chromatin. Many probably have a function in insulator assays, but because of the multifunctional nature of CTCF, we are conservative in our description. Also enriched for the cohesin components RAD21 and SMC3. CTCF is known to recruit the cohesin complex.	Turquoise
E	Predicted enhancer	Regions of open chromatin associated with H3K4me1 signal. Enriched for other enhancer-associated marks, including transcription factors known to act at enhancers. In enhancer assays, many of these (>50%) function as enhancers. A more conservative alternative would be cis-regulatory regions. Enriched for sites for the proteins encoded by <i>EP300</i> , <i>FOS</i> , <i>FOSL1</i> , <i>GATA2</i> , <i>HDAC8</i> , <i>JUNB</i> , <i>JUND</i> , <i>NFE2</i> , <i>SMARCA4</i> , <i>SMARCB1</i> , <i>SIRT6</i> and <i>TAL1</i> genes in K562 cells. Have nuclear and whole-cell RNA signal, particularly poly(A) – fraction.	Orange
PF	Predicted promoter flanking region	Regions that generally surround TSS segments (see below).	Light red
R	Predicted repressed or low-activity region	This is a merged state that includes H3K27me3 polycomb-enriched regions, along with regions that are silent in terms of observed signal for the input assays to the segmentations (low or no signal). They may have other signals (for example, RNA, not in the segmentation input data). Enriched for sites for the proteins encoded by <i>REST</i> and some other factors (for example, proteins encoded by <i>BRF2</i> , <i>CERPB</i> , <i>MAFK</i> , <i>TRIM28</i> , <i>ZNF274</i> and <i>SETDB1</i> genes in K562 cells).	Grey
TSS	Predicted promoter region including TSS	Found close to or overlapping GENCODE TSS sites. High precision/recall for TSSs. Enriched for H3K4me3. Sites of open chromatin. Enriched for transcription factors known to act close to promoters and polymerases Pol II and Pol III. Short RNAs are most enriched in these segments.	Bright red
T	Predicted transcribed region	Overlap gene bodies with H3K36me3 transcriptional elongation signal. Enriched for phosphorylated form of Pol II signal (elongating polymerase) and poly(A) <sup>+</sup> RNA, especially cytoplasmic.	Dark green
WE	Predicted weak enhancer or open chromatin cis-regulatory element	Similar to the E state, but weaker signals and weaker enrichments.	Yellow

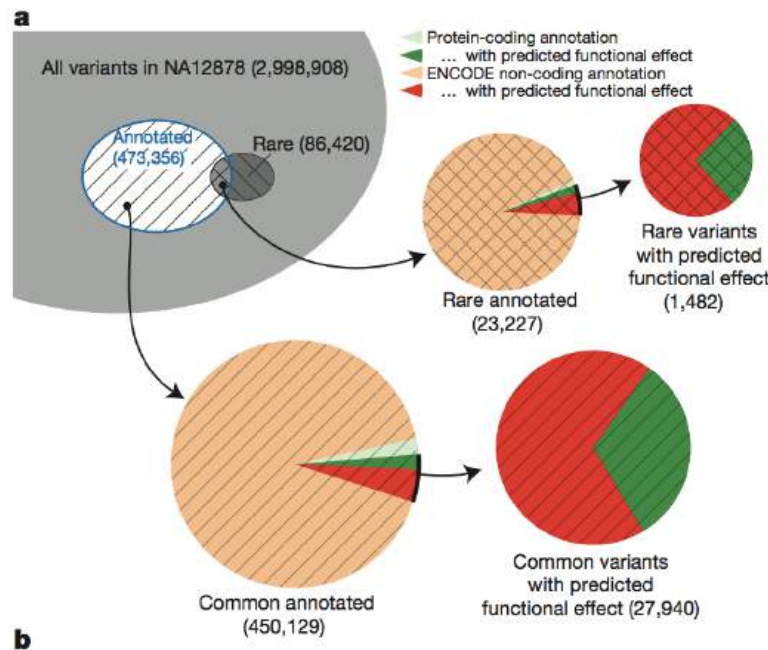
- Use ChromHMM and Segway to Summarize the individual assays into 7 functional/regulatory states

# Major Findings



1. *The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.*
2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*
3. *Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.*
4. *It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.*
5. ***Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.***
6. *Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.*

# Many variants in ENCODE-regions



**Figure 9 | Examining ENCODE elements on a per individual basis in the normal and cancer genome.** a, Breakdown of variants in a single genome (NA12878) by both frequency (common or rare (that is, variants not present in the low-coverage sequencing of 179 individuals in the pilot 1 European panel of the 1000 Genomes project<sup>35</sup>)) and by ENCODE annotation, including protein-coding gene and non-coding elements (GENCODE annotations for protein-coding genes, pseudogenes and other ncRNAs, as well as transcription-factor-binding sites from ChIP-seq data sets, excluding broad annotations such as histone modifications, segmentations and RNA-seq). Annotation status is further subdivided by predicted functional effect, being non-synonymous and missense mutations for protein-coding regions and variants overlapping bound transcription factor motifs for non-coding element annotations. A substantial proportion of variants are annotated as having predicted functional effects in the non-coding category. b, One of several relatively rare occurrences, where

## Breakdown of variants by frequency

- Common or Rare (that is, variants not present in the low-coverage sequencing of 179 individuals in the pilot 1 European panel of the 1000 Genomes project)
- ENCODE annotation, including protein-coding gene and non-coding elements

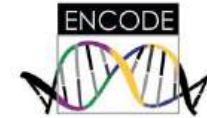
Annotation status is further subdivided by predicted functional effect

- non-synonymous and missense mutations for protein-coding regions and variants overlapping bound transcription factor motifs for non-coding element annotations.

***A substantial proportion of variants are annotated as having predicted functional effects in the non-coding category.***



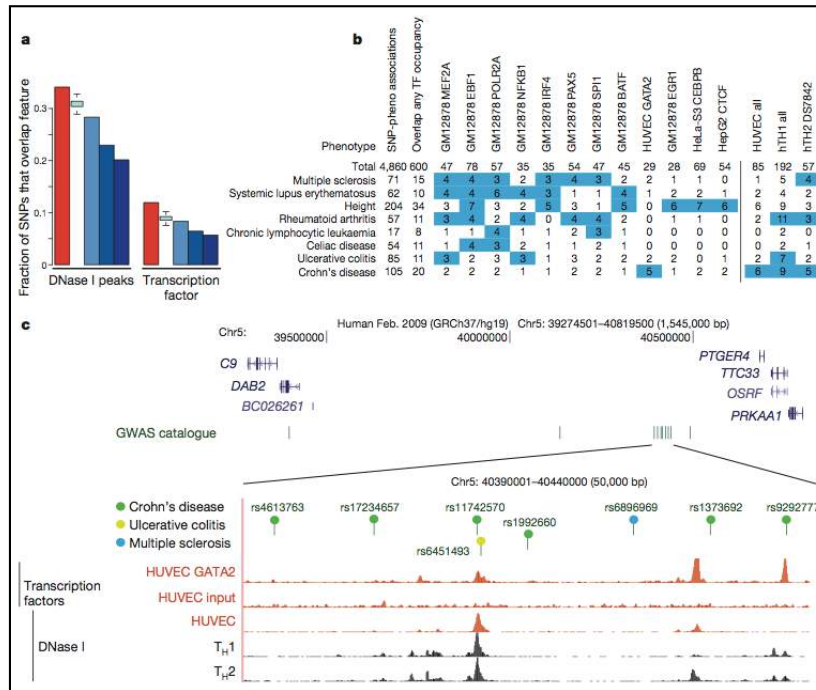
# Major Findings



1. *The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.*
2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*
3. *Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.*
4. *It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.*
5. *Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.*
6. ***Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.***



# ENCODE and Disease

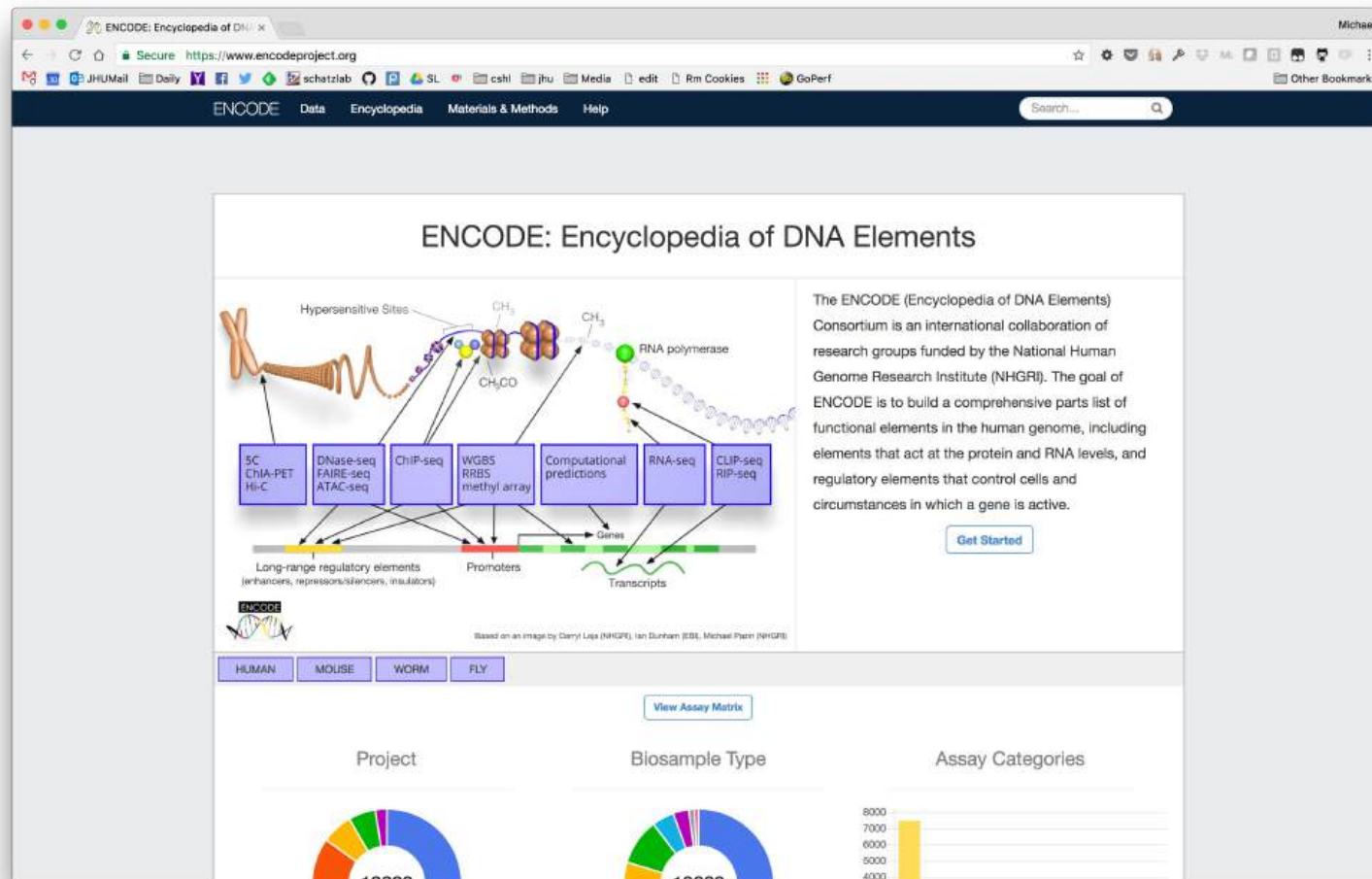


**Figure 10 | Comparison of genome-wide-association-study-identified loci with ENCODE data.** **a**, Overlap of lead SNPs in the NHGRI GWAS SNP catalogue (June 2011) with DHSs (left) or transcription-factor-binding sites (right) as red bars compared with various control SNP sets in blue. The control SNP sets are (from left to right): SNPs on the Illumina 2.5M chip as an example of a widely used GWAS SNP typing panel; SNPs from the 1000 Genomes project; SNPs extracted from 24 personal genomes (see personal genome variants track at <http://main.genome-browser.bx.psu.edu> (ref. 80)), all shown as blue bars. In addition, a further control used 1,000 randomizations from the genotyping SNP panel, matching the SNPs with each NHGRI catalogue SNP for allele frequency and distance to the nearest TSS (light blue bars with bounds at 1.5 times the interquartile range). For both DHSs and transcription-factor-binding regions, a larger proportion of overlaps with GWAS-implicated SNPs is found compared to any of the controls sets. **b**, Aggregate overlap of

phenotypes to selected transcription-factor-binding sites (left matrix) or DHSs in selected cell lines (right matrix), with a count of overlaps between the phenotype and the cell line/factor. Values in blue squares pass an empirical  $P$ -value threshold  $\leq 0.01$  (based on the same analysis of overlaps between randomly chosen, GWAS-matched SNPs and these epigenetic features) and have at least a count of three overlaps. The  $P$  value for the total number of phenotype-transcription factor associations is  $< 0.001$ . **c**, Several SNPs associated with Crohn's disease and other inflammatory diseases that reside in a large gene desert on chromosome 5, along with some epigenetic features indicative of function. The SNP (rs11742570) strongly associated to Crohn's disease overlaps a GATA2 transcription-factor-binding signal determined in HUVECs. This region is also DNase I hypersensitive in HUVECs and T-helper T<sub>H</sub>1 and T<sub>H</sub>2 cells. An interactive version of this figure is available in the online version of the paper.

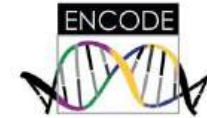
- 88% of GWAS SNPs are intronic or intergenic of unknown function
- We found that 12% of these GWAS-SNPs overlap transcription-factor-occupied regions whereas 34% overlap DHSs
- GWAS SNPs are particularly enriched in the segmentation classes associated with enhancers and TSSs across several cell types

# ENCODE Studies



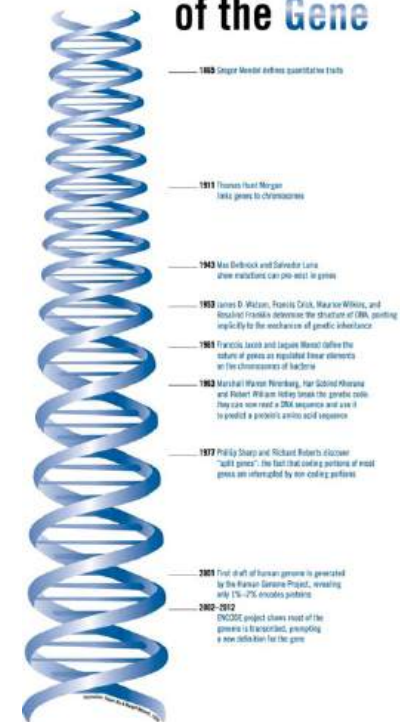
>17000 Citations for main paper; >>25k for all papers

# Summary & Critique



- **Summary**
  - *The unprecedented number of functional elements identified in this study provides a valuable resource to the scientific community as well as significantly enhances our understanding of the human genome.*
- **Critique**
  - Was it correct?
  - What is functional?
  - What is conservation?
  - What was the control?
  - What are the tradeoffs of organizing so much funding (\$288M!) around a single project; will other groups successfully use these data?

## Redefining the Nature of the Gene



# Attention Is All You Need ("Transformers")

Michael Schatz  
October 21, 2024  
Applied Comparative Genomics

# “Transformers”

## Attention Is All You Need

**Ashish Vaswani\***  
Google Brain  
avaswani@google.com

**Noam Shazeer\***  
Google Brain  
noam@google.com

**Niki Parmar\***  
Google Research  
nikip@google.com

**Jakob Uszkoreit\***  
Google Research  
usz@google.com

**Llion Jones\***  
Google Research  
llion@google.com

**Aidan N. Gomez\* †**  
University of Toronto  
aidan@cs.toronto.edu

**Lukasz Kaiser\***  
Google Brain  
lukaszkaizer@google.com

**Illia Polosukhin\* ‡**  
illia.polosukhin@gmail.com

### Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.



**Ashish Vaswani**  
Essential AI



**Llion Jones**  
Sakana AI



**Noam Shazeer**  
Character AI



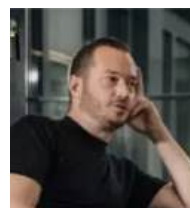
**Aidan Gomez**  
Cohere



**Niki Parmar**  
Essential AI



**Łukasz Kaiser**  
OpenAI

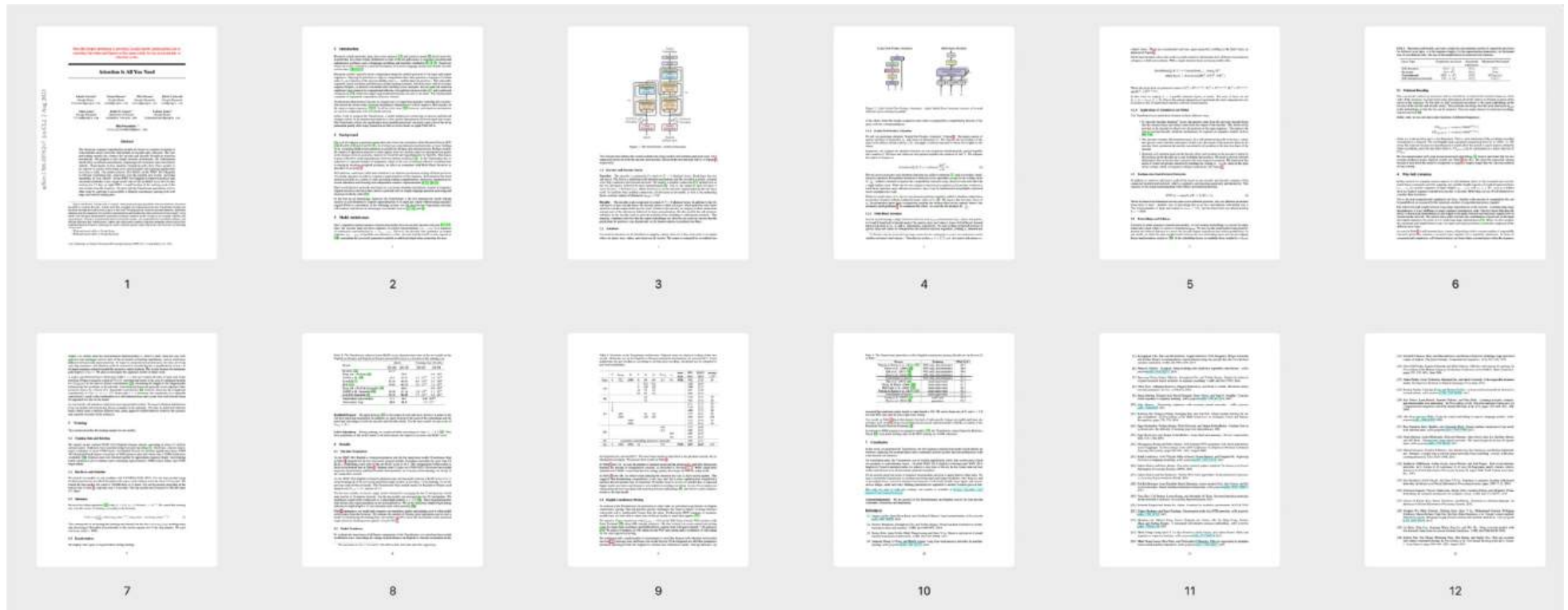


**Jakob Uszkoreit**  
Inception



**Illia Polosukhin**  
NEAR

# The paper



Vaswani et al (2017) arXiv:1706.03762



## Attention is all you need

Authors Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin

Publication date 2017

Journal Advances in neural information processing systems

Volume 30

Description The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder and decoder configuration. The best performing such models also connect the encoder and decoder through an attention mechanism. We propose a novel, simple network architecture based solely on an attention mechanism, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our single model with 165 million parameters, achieves 27.5 BLEU on English-to-German translation, improving over the existing best ensemble result by over 1 BLEU. On English-to-French translation, we outperform the previous single state-of-the-art with model by 0.7 BLEU, achieving a BLEU score of 41.1.

Total Citations Cited by 133826

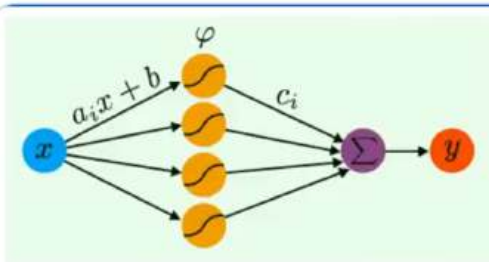




# Outline

1. Recap on ANNs and CNNs
2. The problem
3. Self-Attention
4. Transformers
5. Impact

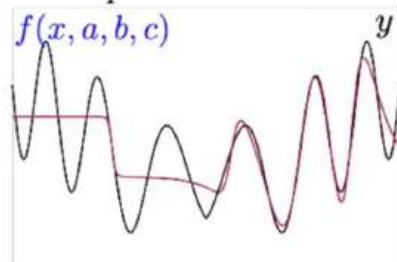
# ANNs are “Universal Approximators”



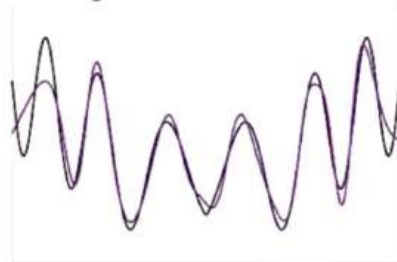
1 hidden layer perceptron:

$$y \approx f(x, a, b, c) \stackrel{\text{def.}}{=} \sum_{i=1}^p c_i \varphi(a_i x + b_i)$$

$p = 6$  neurons



$p = 20$  neurons



## Approximation by Superpositions of a Sigmoidal Function\*

G. Cybenko†

**Abstract.** In this paper we demonstrate that finite linear combinations of compositions of a fixed, univariate function and a set of affine functionals can uniformly approximate any continuous function of  $n$  real variables with support in the unit hypercube; only mild conditions are imposed on the univariate function. Our results settle an open question about representability in the class of single hidden layer neural networks. In particular, we show that arbitrary decision regions can be arbitrarily well approximated by continuous feedforward neural networks with only a single internal, hidden layer and any continuous sigmoidal nonlinearity. The paper discusses approximation properties of other possible types of nonlinearities that might be implemented by artificial neural networks.

**Key words.** Neural networks, Approximation, Completeness.



George Cybenko

## ConvnetJS demo: toy 2d classification with 2-layer neural network

The simulation below shows a toy binary problem with a few data points of class 0 (red) and 1 (green). The network is set up as:

```
layer_defs = [];
layer_defs.push({type:'input', out_sx:1, out_sy:1, out_depth:2});
layer_defs.push({type:'fc', num_neurons:6, activation: 'tanh'});
layer_defs.push({type:'fc', num_neurons:2, activation: 'tanh'});
layer_defs.push({type:'softmax', num_classes:2});

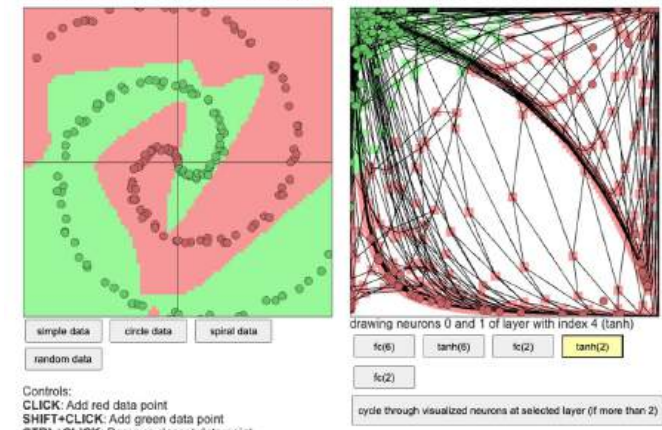
net = new convnetjs.Net();
net.makeLayers(layer_defs);

trainer = new convnetjs.SGDTrainer(net, {learning_rate:0.01, momentum:0.1, batch_size:10, l2_decay:0.001});
```

change network

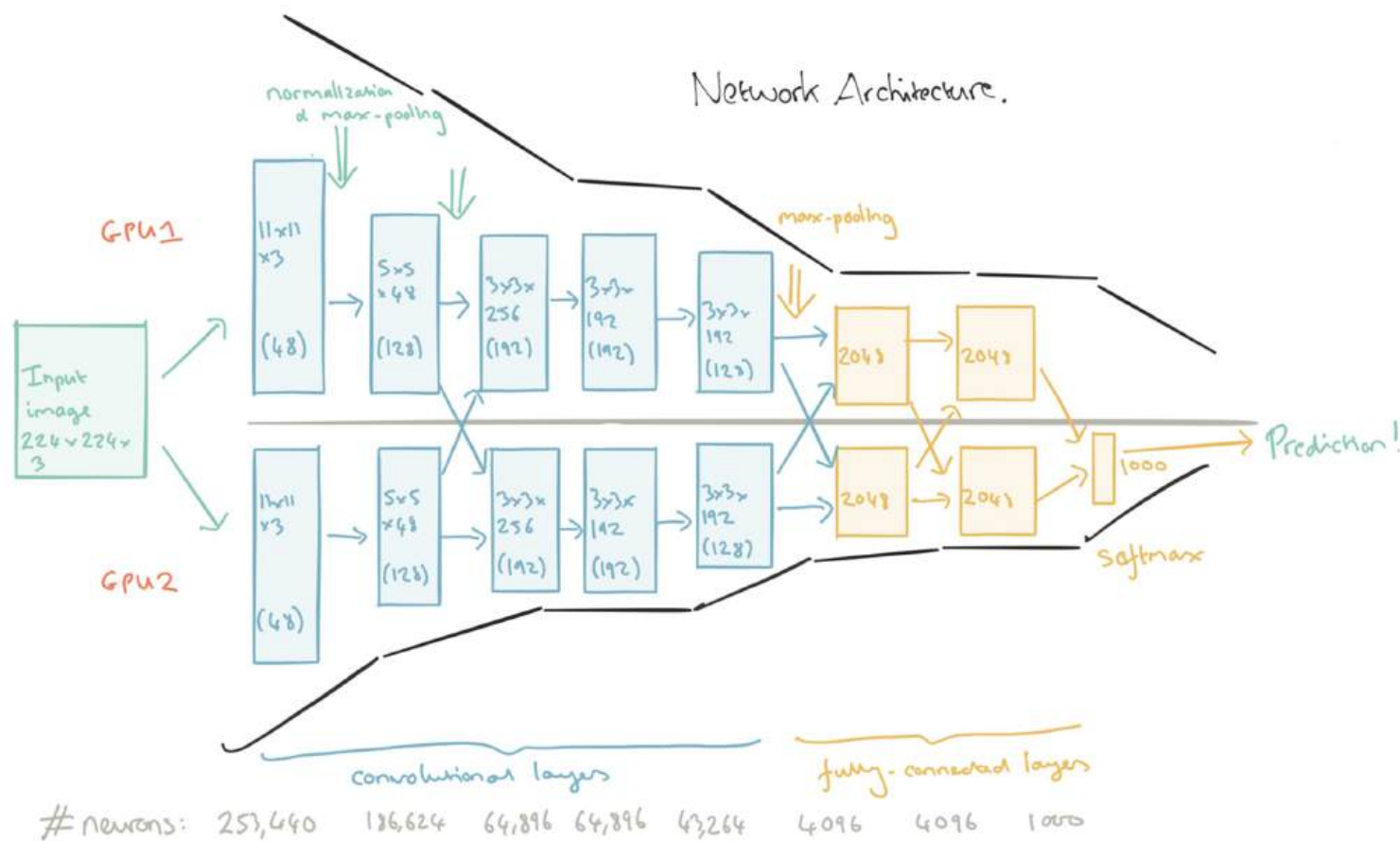
Feel free to change this, the text area above gets eval()'d when you hit the button and the network gets reloaded. Every 10th of a second, all points are fed to the network multiple times through the trainer class to train the network. The resulting predictions of the network are then "painted" under the data points to show you the generalization.

On the right we visualize the transformed representation of all grid points in the original space and the data, for a given layer and only for 2 neurons at a time. The number in the bracket shows the total number of neurons at that level of representation. If the number is more than 2, you will only see the two visualized but you can cycle through all of them with the cycle button.



Controls:  
CLICK: Add red data point  
SHIFT+CLICK: Add green data point  
CTRL+CLICK: Removes closest data point

# AlexNet Architecture (annotated)




DOG

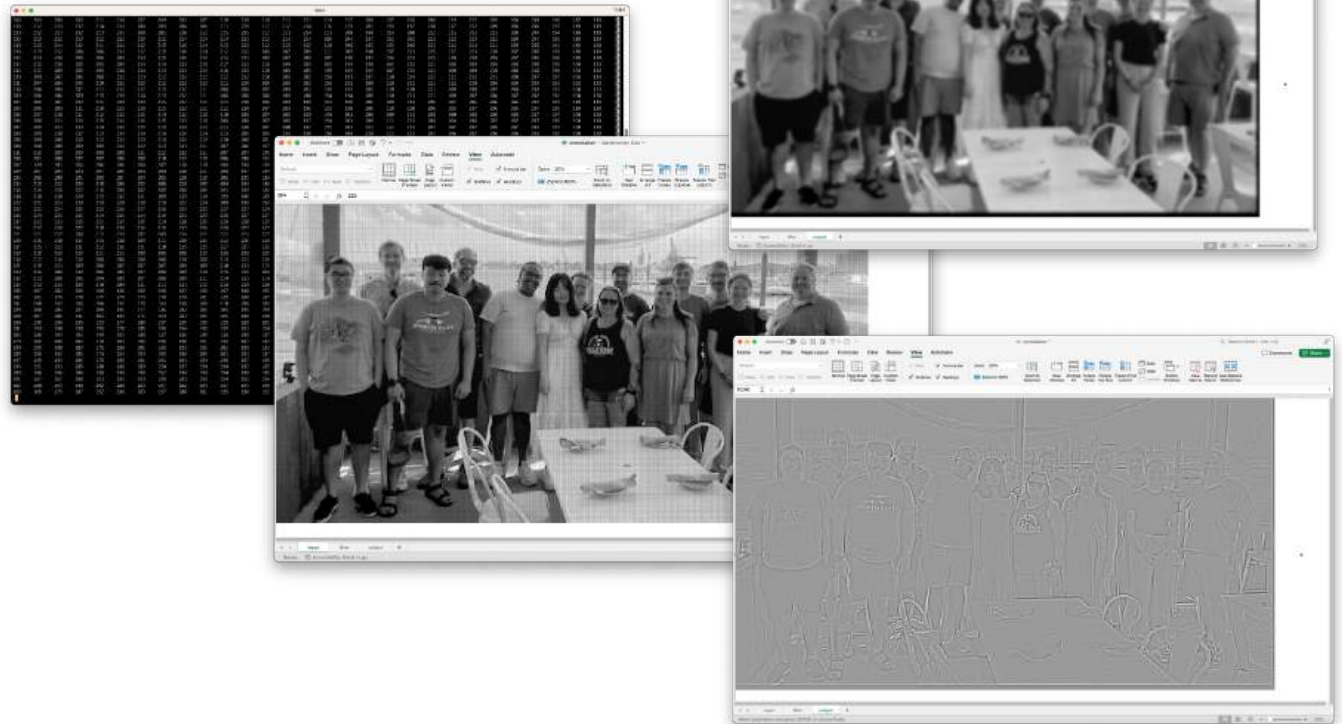
BAGEL

VAN

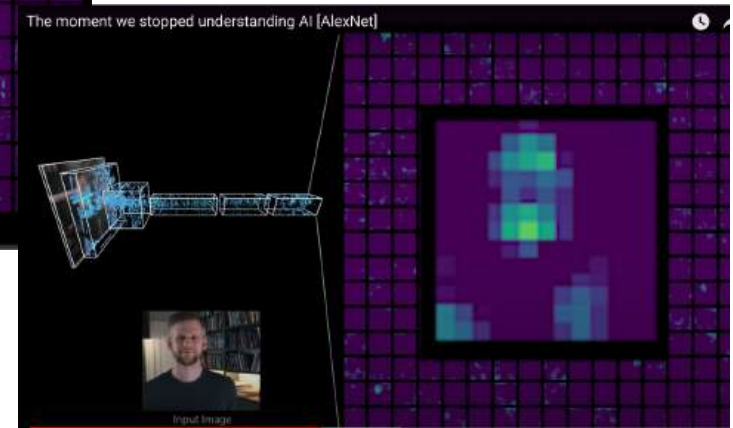
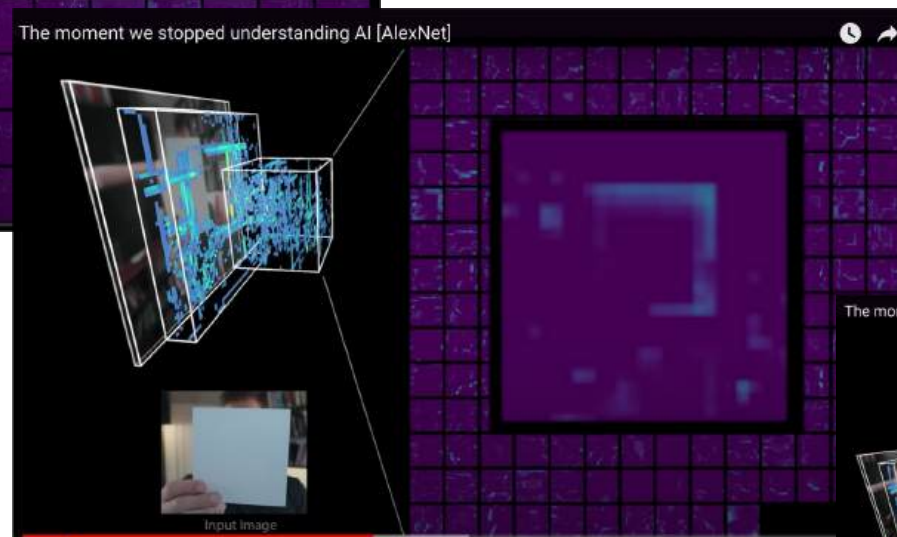
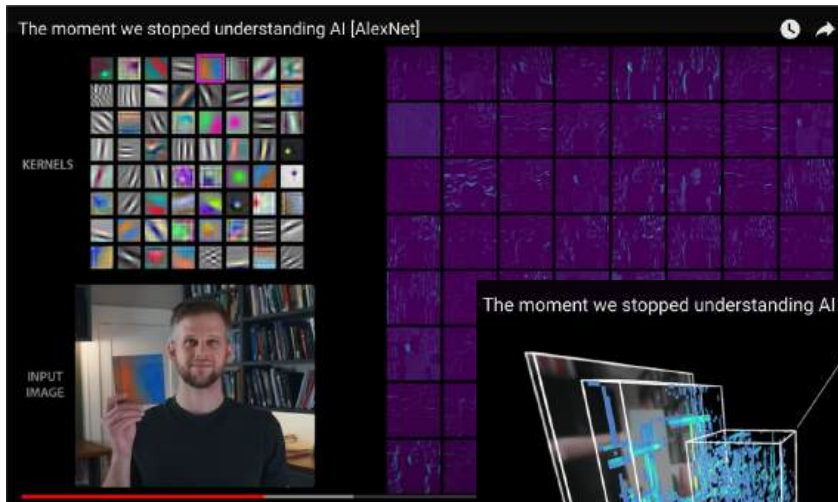
<https://blog.acolyer.org/2016/04/20/imagenet-classification-with-deep-convolutional-neural-networks/>

# Convolutional Kernels

Operation	Kernel $w$	Image result $g(x,y)$
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Ridge or edge detection	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur 3 x 3 (approximation)	$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	
Gaussian blur 5 x 5 (approximation)	$\frac{1}{256} \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}$	
Unsharp masking 5 x 5 Based on Gaussian blur with amount as 1 and threshold as 0 (with no image mask)	$\frac{-1}{256} \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}$	



= input!A1 \* filter!\$A\$1 + input!A2 \* filter!\$A\$2 + input!A3 \* filter!\$A\$3 + input!A4 \* filter!\$A\$4 + input!A5 \* filter!\$A\$5 +  
 input!B1 \* filter!\$B\$1 + input!B2 \* filter!\$B\$2 + input!B3 \* filter!\$B\$3 + input!B4 \* filter!\$B\$4 + input!B5 \* filter!\$B\$5 +  
 input!C1 \* filter!\$C\$1 + input!C2 \* filter!\$C\$2 + input!C3 \* filter!\$C\$3 + input!C4 \* filter!\$C\$4 + input!C5 \* filter!\$C\$5 +  
 input!D1 \* filter!\$D\$1 + input!D2 \* filter!\$D\$2 + input!D3 \* filter!\$D\$3 + input!D4 \* filter!\$D\$4 + input!D5 \* filter!\$D\$5 +  
 input!E1 \* filter!\$E\$1 + input!E2 \* filter!\$E\$2 + input!E3 \* filter!\$E\$3 + input!E4 \* filter!\$E\$4 + input!E5 \* filter!\$E\$5



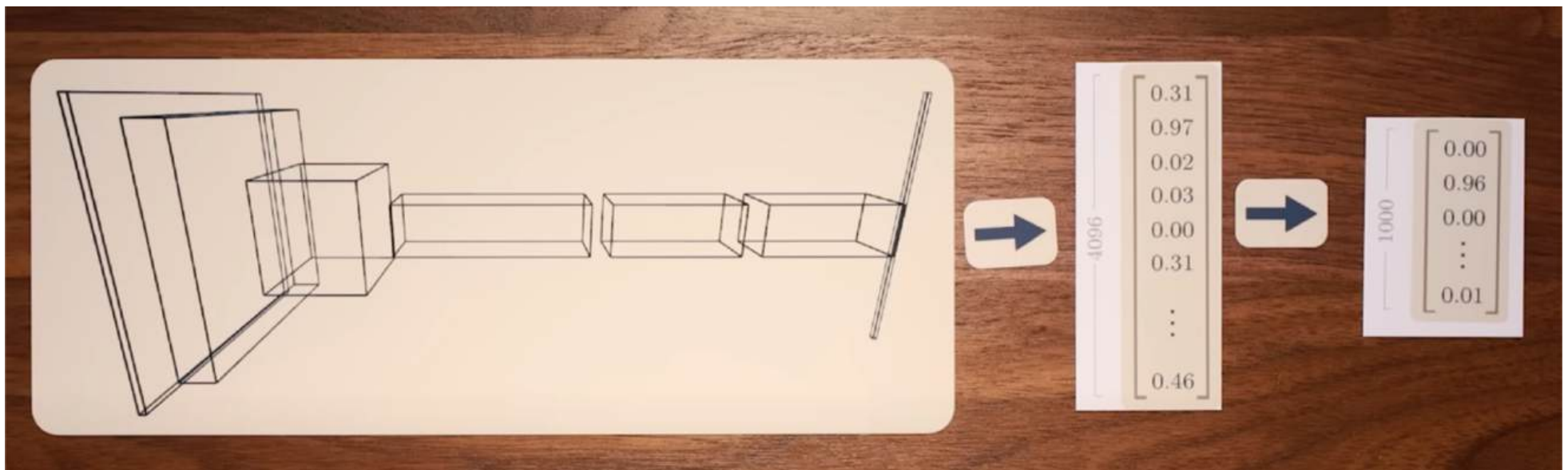
<https://www.youtube.com/watch?v=UZDiGooFs54>



Final two layers are special:

Last layer  $[1 \times 1000]$ : Probability of the image in class  $i$

Second last  $[1 \times 4096]$ : Projection of image into 4096-dim space



<https://www.youtube.com/watch?v=UZDiGooFs54>

## Activation Atlas: tSNE plot of images embeddings



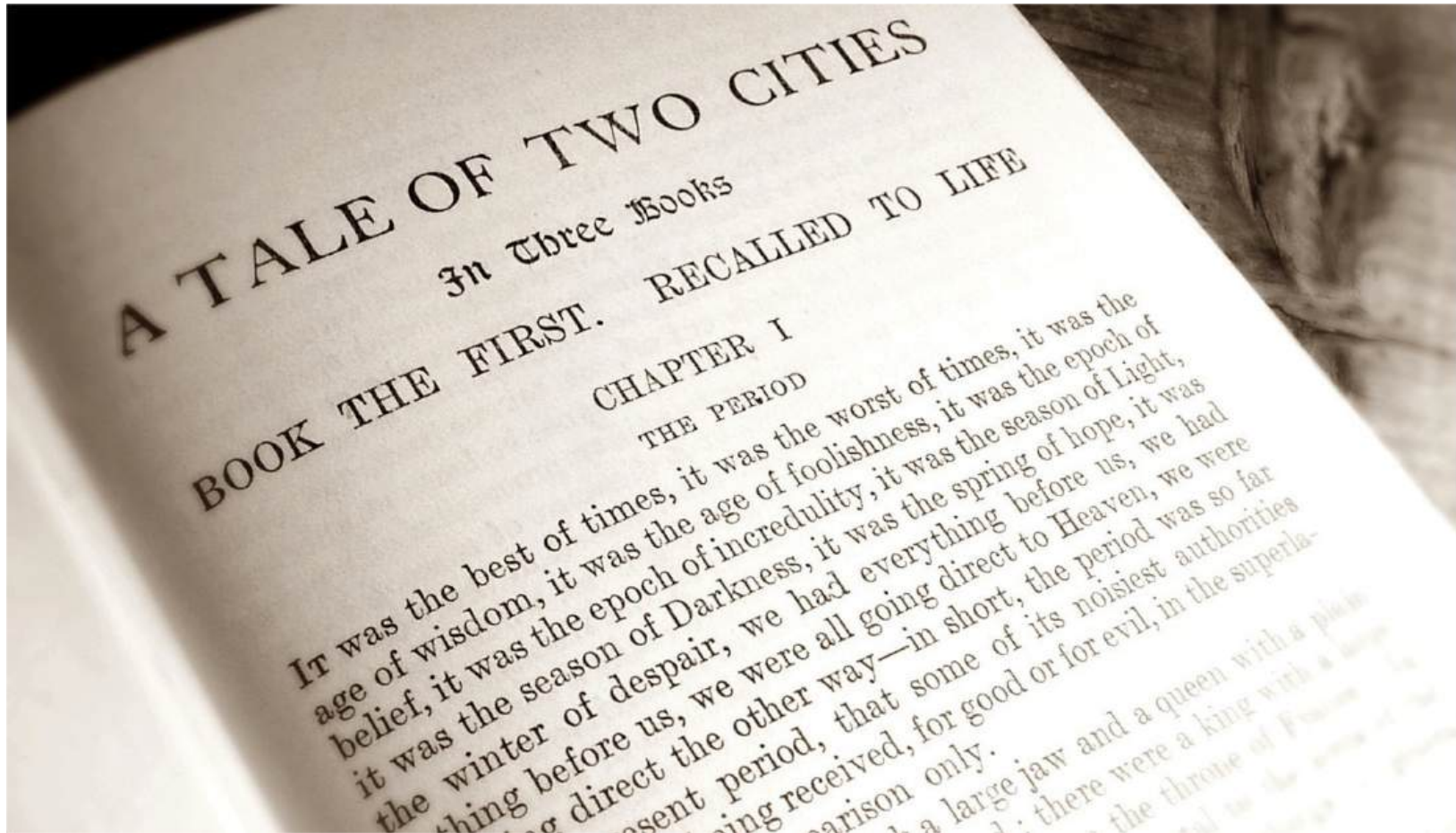
<https://www.youtube.com/watch?v=UZDiGooFs54>



## Outline

1. Recap on ANNs and CNNs
2. The problem
3. Self-Attention
4. Transformers
5. Impact

## What about text data?



Or other sequential data?

<http://introtodeeplearning.com/>



# A Sequence Modeling Problem: Predict the Next Word

"This morning I took my cat for a ???

given these words

predict the  
next word

# A Sequence Modeling Problem: Predict the Next Word

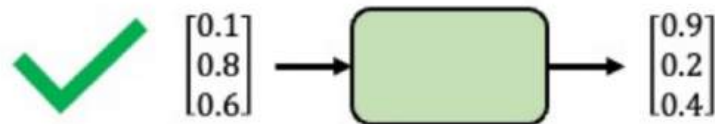
"This morning I took my cat for a ???

given these words      predict the next word

## Representing Language to a Neural Network



*Neural networks cannot interpret words*



*Neural networks require numerical inputs*



# A Sequence Modeling Problem: Predict the Next Word

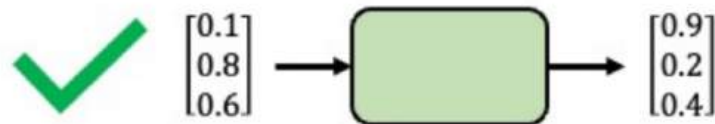
$f(\text{"This morning I took my cat for a"}) = ???$

given these words      predict the next word

## Representing Language to a Neural Network



*Neural networks cannot interpret words*



*Neural networks require numerical inputs*

# A Sequence Modeling Problem: Predict the Next Word

"This morning I took my cat for a walk."

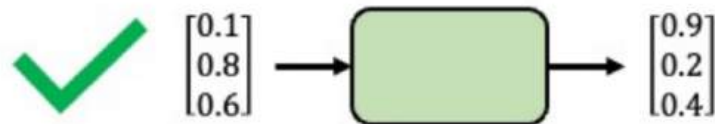
given these words

predict the  
next word

## Representing Language to a Neural Network



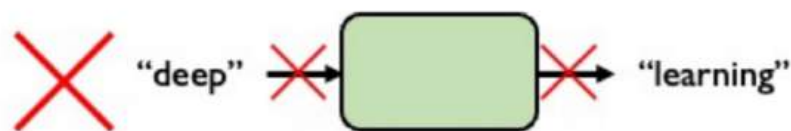
*Neural networks cannot interpret words*



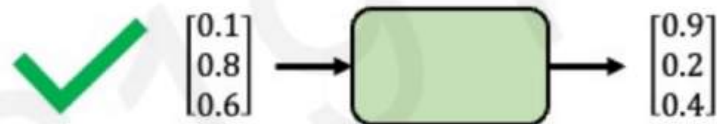
*Neural networks require numerical inputs*



# Encoding Language for a Neural Network



Neural networks cannot interpret words



Neural networks require numerical inputs

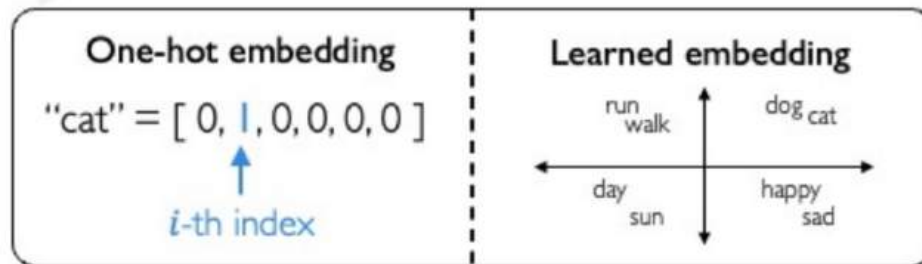
Embedding: transform indexes into a vector of fixed size.

this cat for  
my took  
a | walk  
morning

**1. Vocabulary:**  
Corpus of words

a → 1  
cat → 2  
... → ...  
walk → N

**2. Indexing:**  
Word to index



**3. Embedding:**  
Index to fixed-sized vector

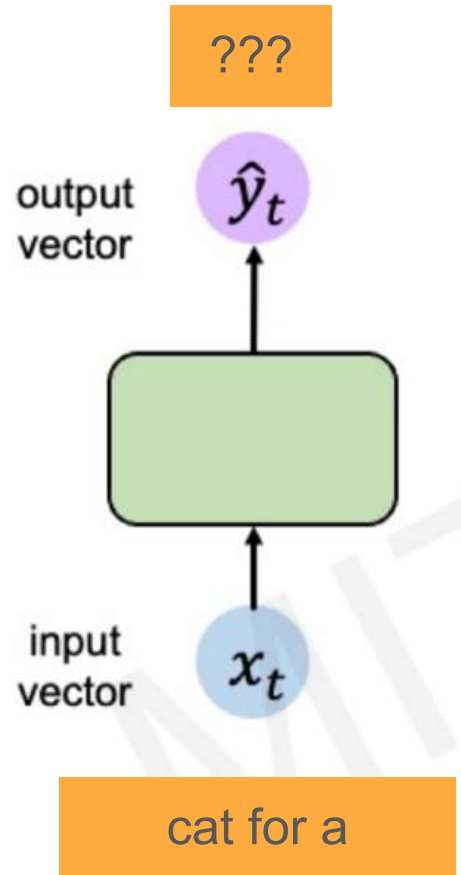
# A Sequence Modeling Problem: Predict the Next Word

"This morning I took my cat for a ???

given these words                      predict the next word

To model sequences, we need to:

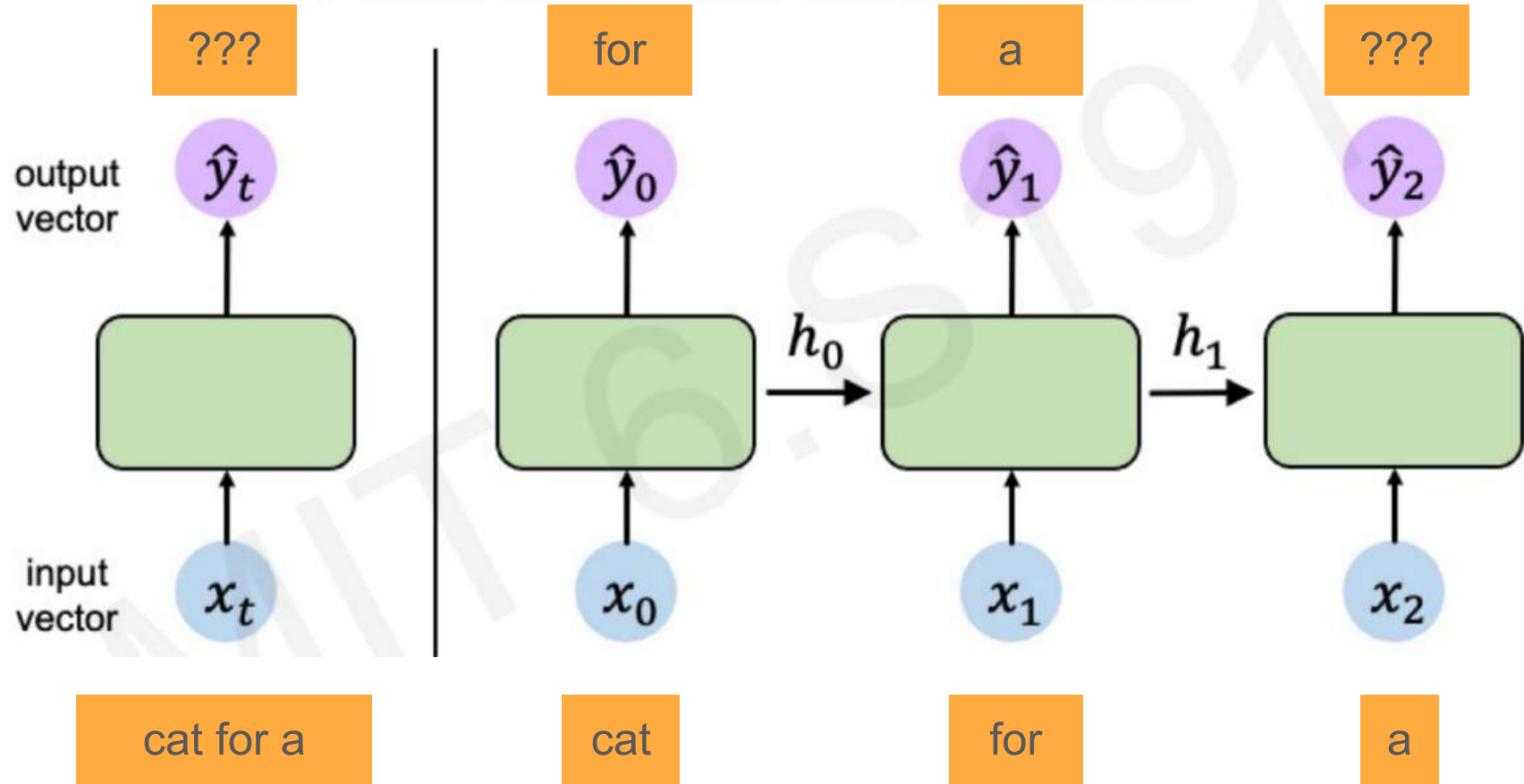
1. Handle **variable-length** sequences
2. Track **long-term** dependencies
3. Maintain information about **order**
4. **Share parameters** across the sequence



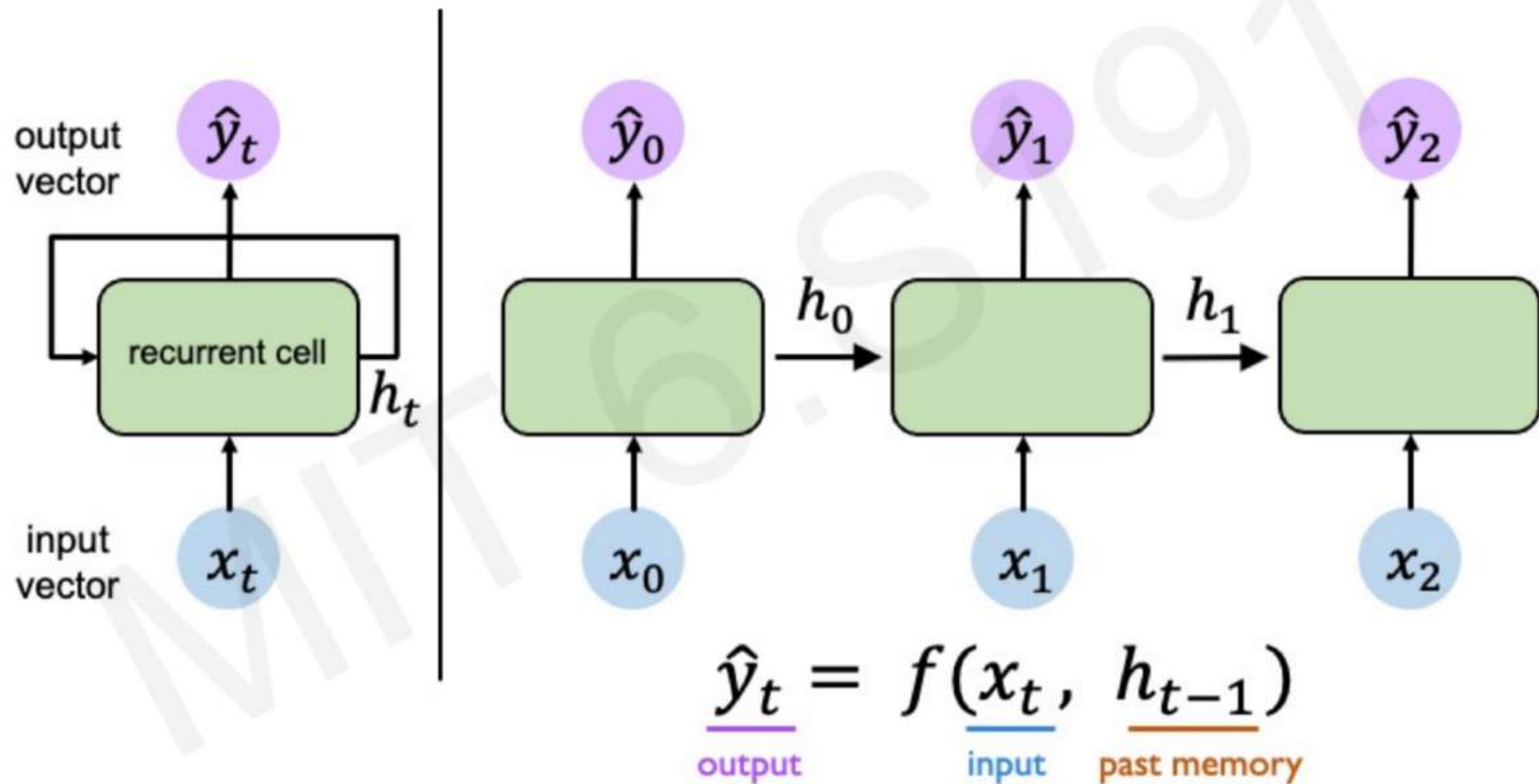
### Fixed context networks

- Simple but very limited accuracy
- Quickly runs out of training data for longer windows

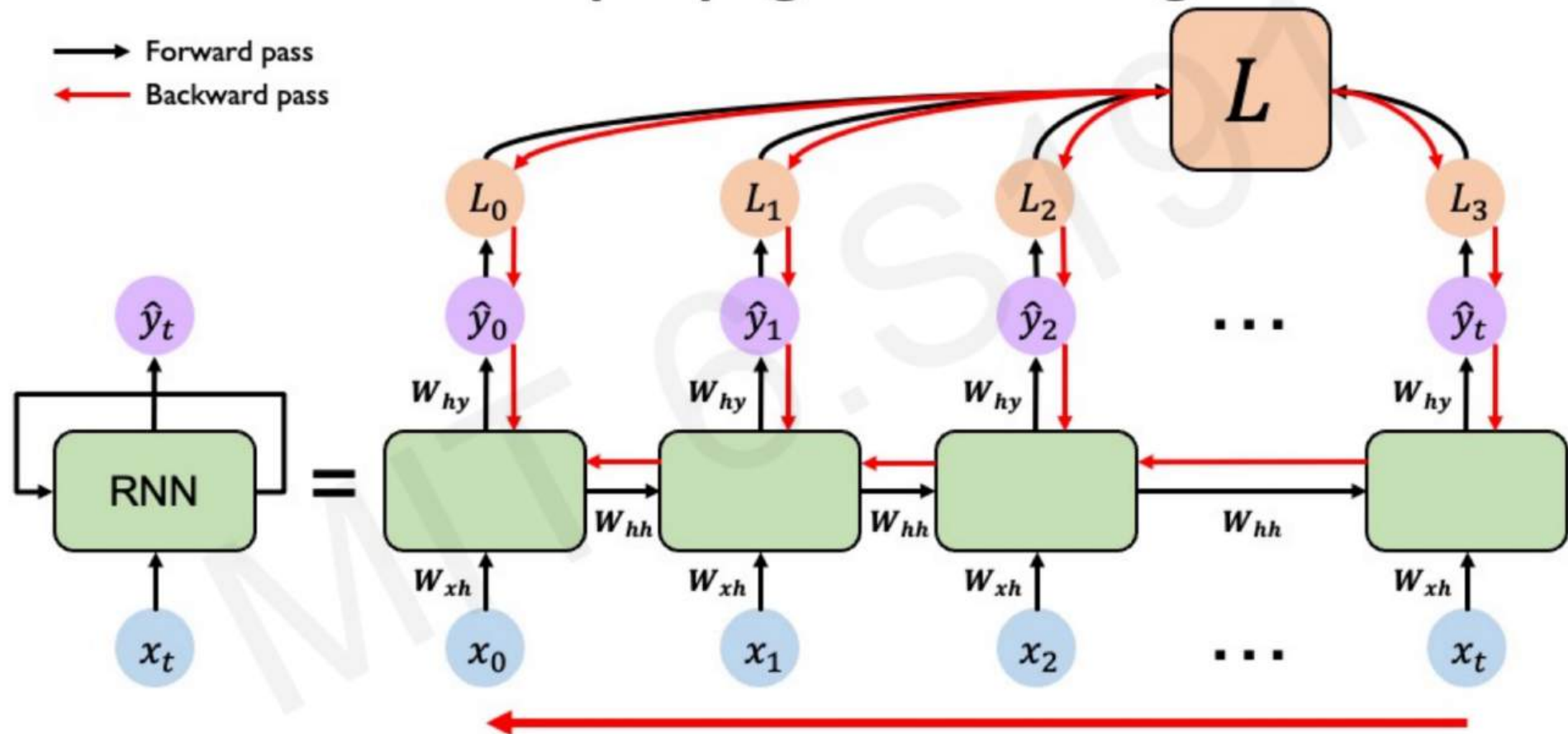
# Neurons with Recurrence



# Neurons with Recurrence



# RNNs: Backpropagation Through Time





# The Problem of Long-Term Dependencies

Why are vanishing gradients a problem?

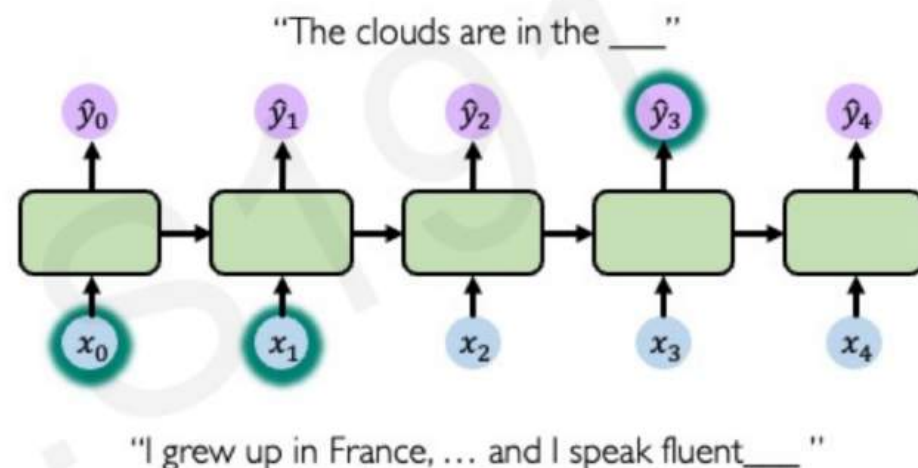
Multiply many **small numbers** together



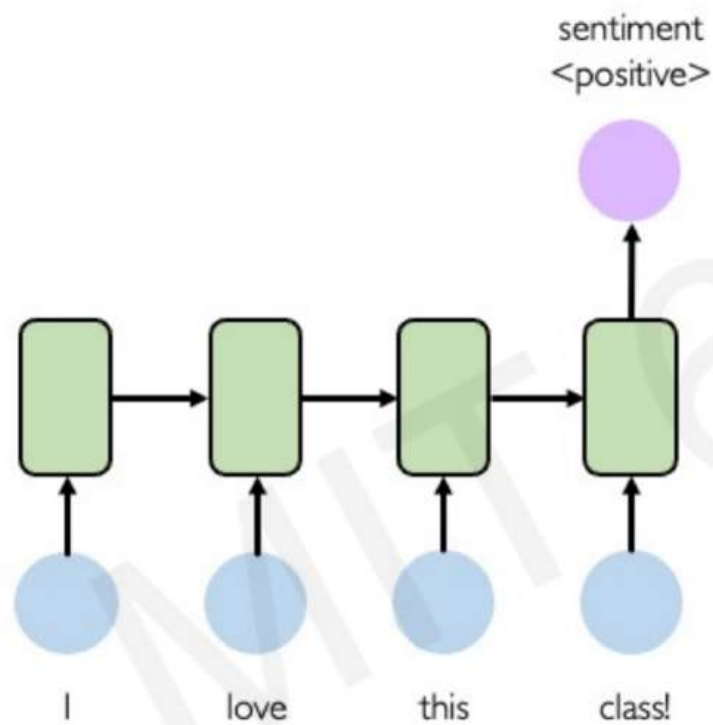
Errors due to further back time steps have smaller and smaller gradients






Bias parameters to capture short-term dependencies



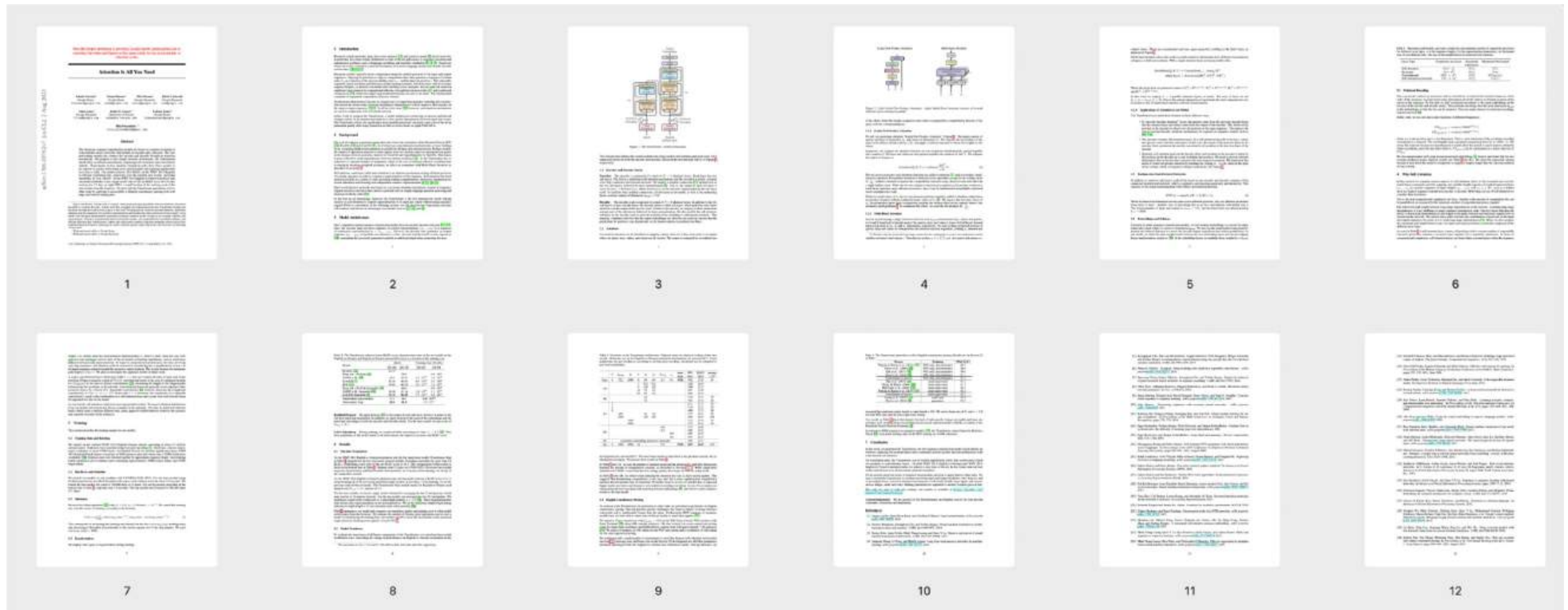
# Limitations of Recurrent Models



## Limitations of RNNs

-  Encoding bottleneck
-  Slow, no parallelization
-  Not long memory

# The paper



Vaswani et al (2017) arXiv:1706.03762

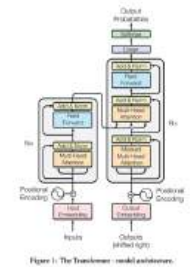
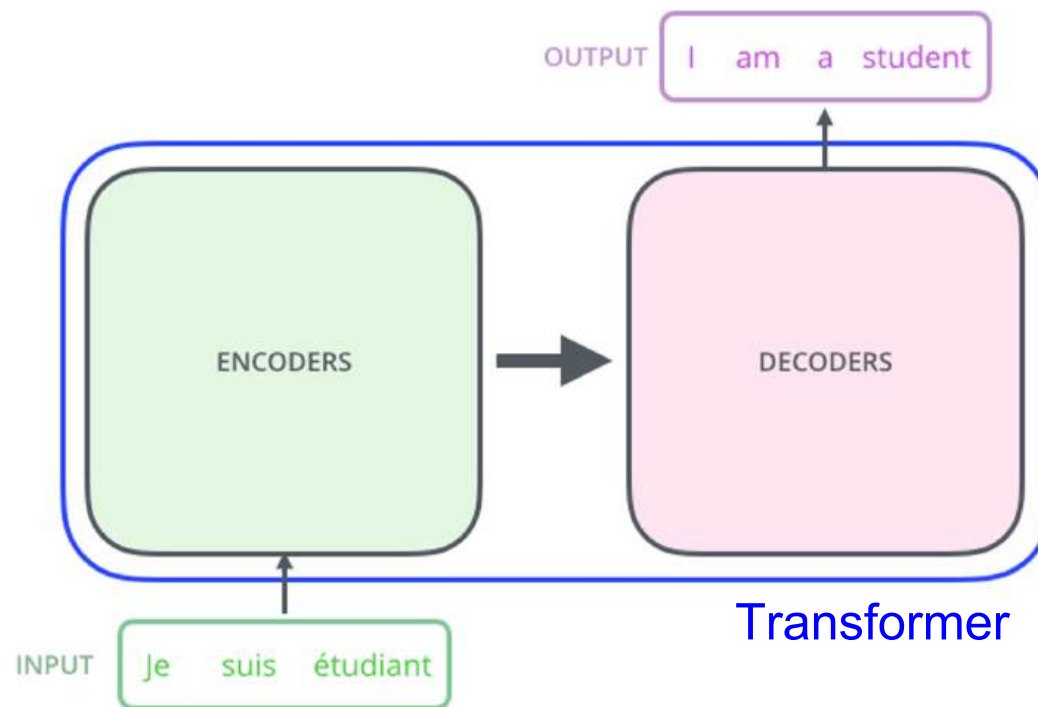
## The application



Trained with a huge collection of pairs of sentences in french & english  
Measure accuracy using “BLEU” scores to gold standard (higher is better)  
Use a “reasonable” amount of GPU time for training (<1 week on 8 GPUs)

<https://jalammar.github.io/illustrated-transformer/>

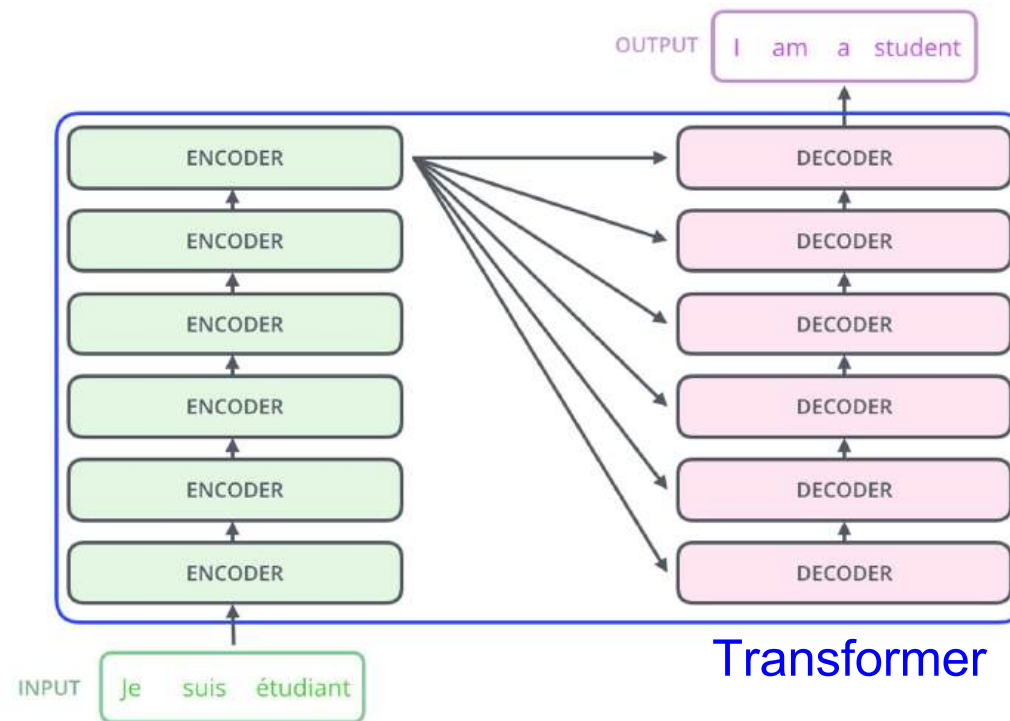
# The architecture



Transformer is composed of an input encoder followed by a output decoder

<https://jalammar.github.io/illustrated-transformer/>

## The architecture



Encoder transforms sentences in language A into a highly abstract embedded space  
Decoders progressively decodes highly abstract embedded space into language B

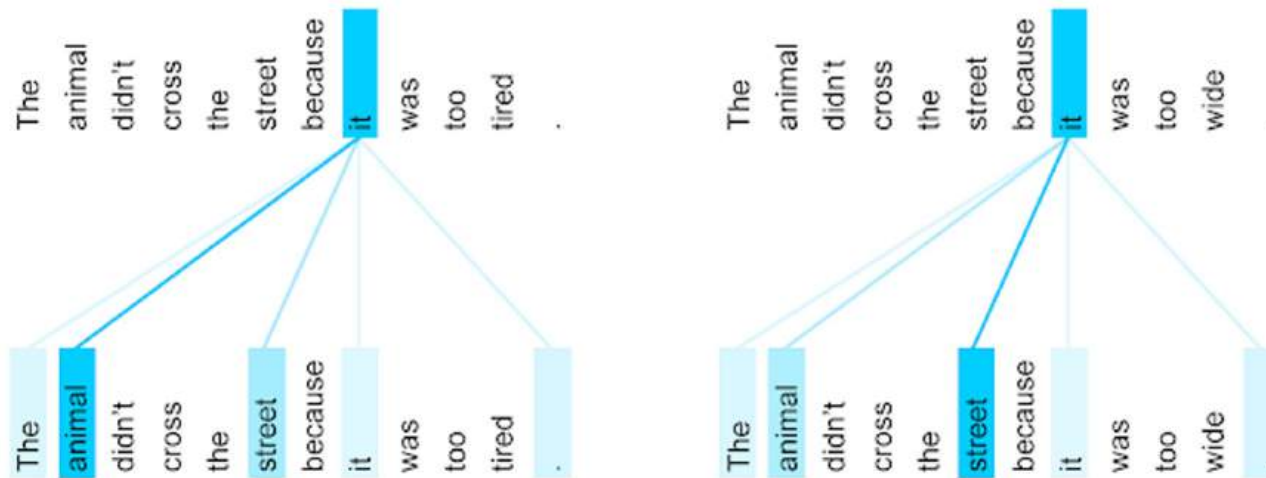
<https://jalammar.github.io/illustrated-transformer/>



## Coreference resolution: self attention

The animal didn't cross the street because it was too tired.  
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.  
L'animal n'a pas traversé la rue parce qu'elle était trop large.



<https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/>