Pre-Training Strategies

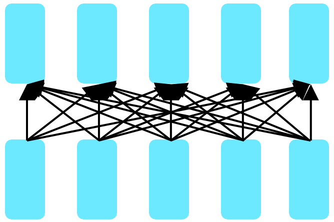
Encoder-Decoder and Decoder-only Models

Tanmoy Chakraborty

Associate Professor, IIT Delhi

https://tanmoychak.com/

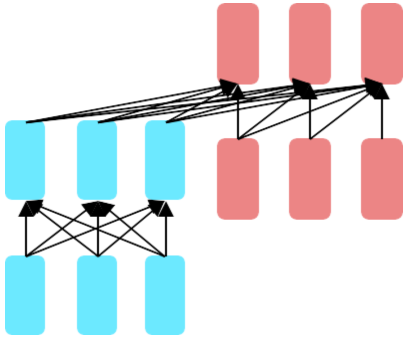
Pre-Training for Different Types of Architectures



Encoder-only

BERT (already discussed)

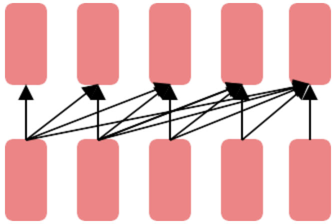
- Gets bi-directional context – can condition on future!
- How do we train them to build strong representations?



Encoder-Decoder

BART, T5

- Good parts of decoders and encoders?
- What's the best way to pretrain them?



Decoder-only

GPT, Llama

- Language models!
- Nice to generate from; can't condition on future words

Pre-Training Encoder-Decoder Models

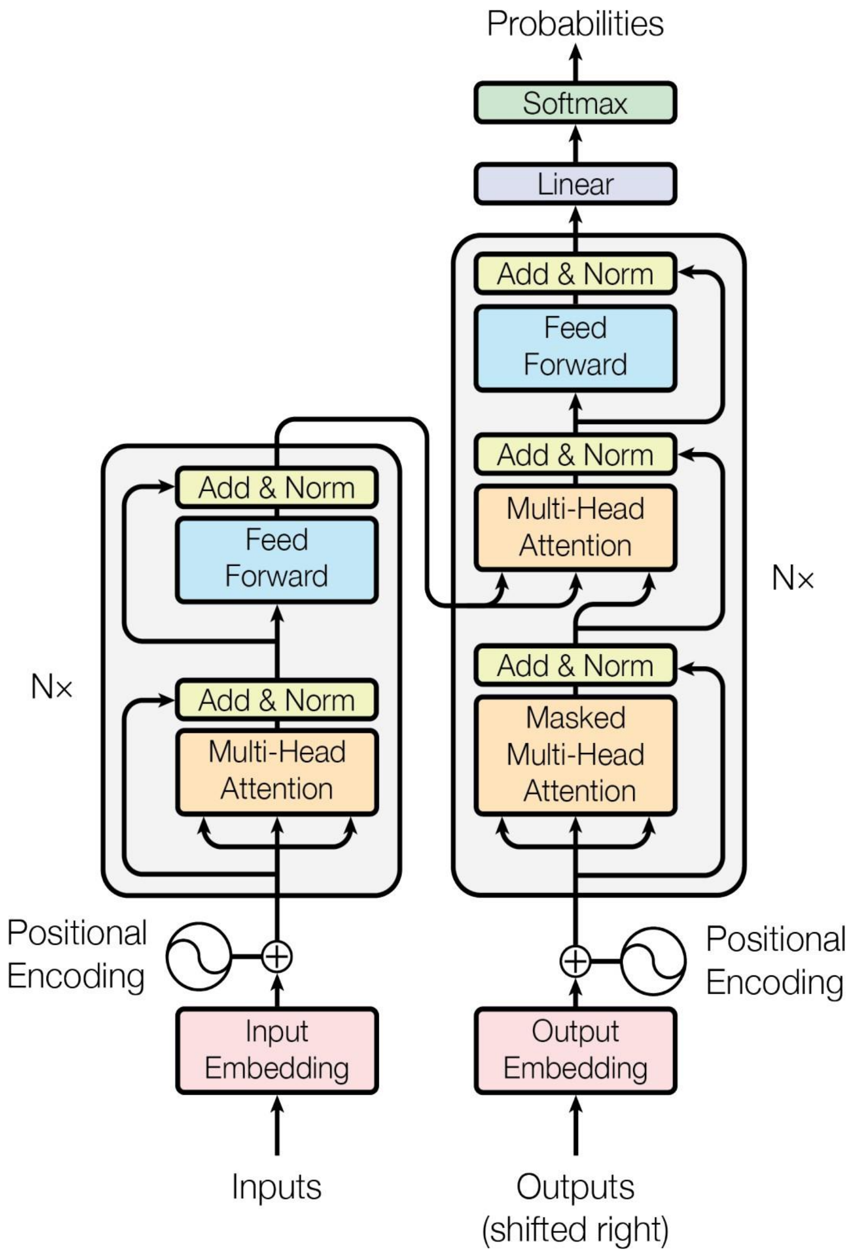
BART and T5

Pre-Training Encoder-Decoder Models

- Masked LMs: trained bidirectionally but with masking
- How can we pre-train a model for P($\mathbf{y}$ | $\mathbf{x}$)?
- Why was BERT effective?
    - Predicting a mask requires some kind of text “understanding”.
- What would it take to do the same for sequence prediction?

Recall: Encoder-Decoder Architecture

- Standard Transformer Architecture
- Decoder attends back to the input. But the input doesn't change, so this just needs to be encoded once.



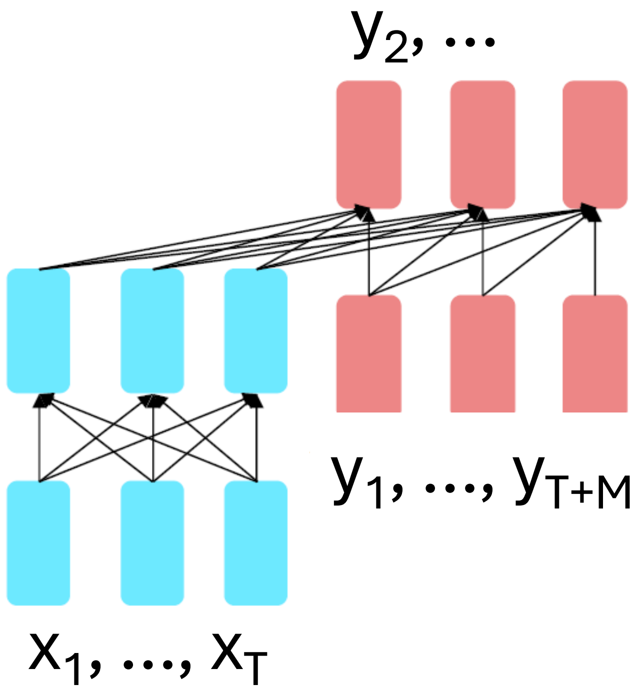
Pre-Training Encoder-Decoder Models

- For **encoder-decoders**, we could do something like **language modeling**, but where a prefix of every input is provided to the encoder and is not predicted.

$$h_1, \ldots, h_T = \textit{Encoder}\,(x_1, \ldots, x_T)$$

$$h_{T+1}, \ldots, h_{T+M} = \textit{Decoder}\,(y_1, \ldots, y_{i-1}, h_1, \ldots, h_T)$$

$$P(y_i \mid y_{<i}, h_{1:T}) = \textit{Softmax}(Wh_i + b)$$

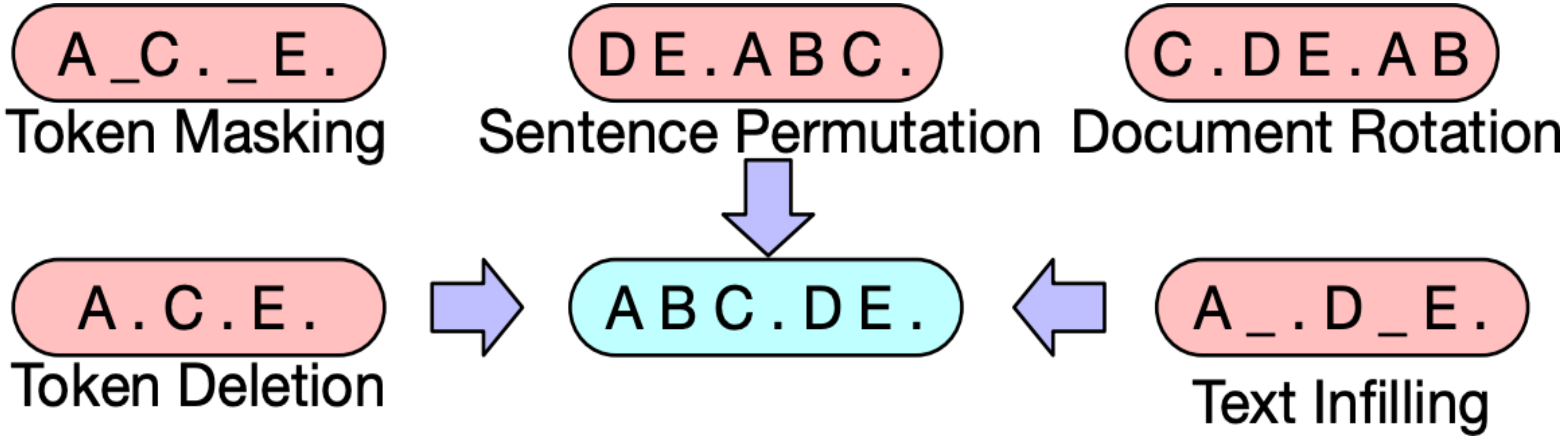


The **encoder** portion benefits from bidirectional context; the **decoder** portion is used to train the whole model through language modeling.

Pre-Training Encoder-Decoder Models

- How can we pre-train a model for P(**y** | **x**)?
- **Requirements:**
  1. should use unlabeled data
  2. should force a model to attend from **y** back to **x**

Pre-Training BART (Bidirectional and Auto-Regressive Transformers)



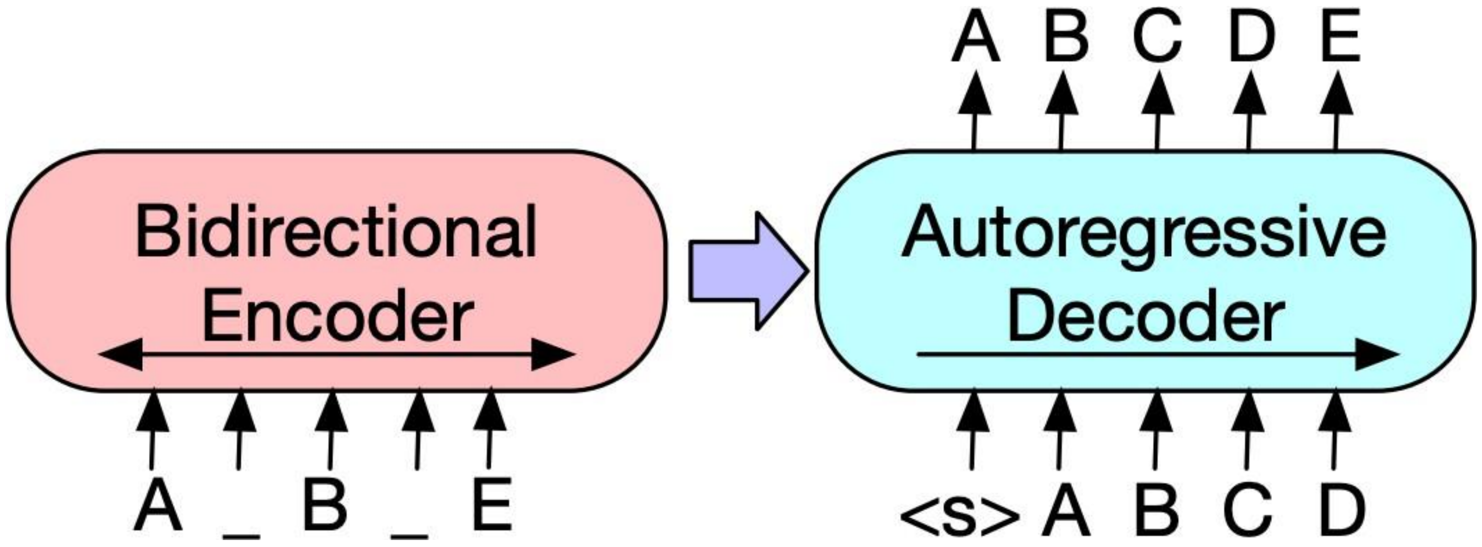
Infilling is longer spans than masking

- Several possible strategies for corrupting a sequence are explored in the BART paper.

Lewis et al. (2019), "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension"

Pre-Training BART

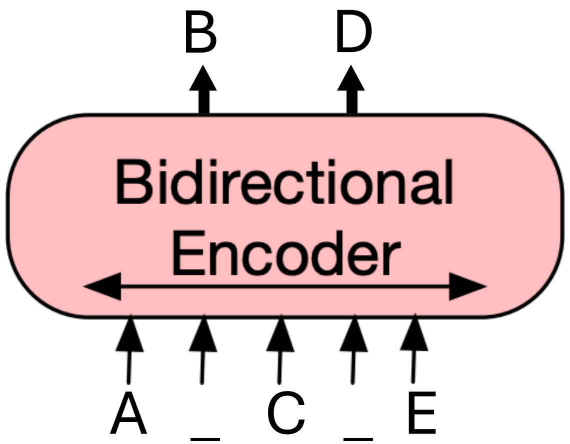
- Sequence-to-sequence Transformer trained on this data: permute/make/delete tokens, then predict full sequence autoregressively.

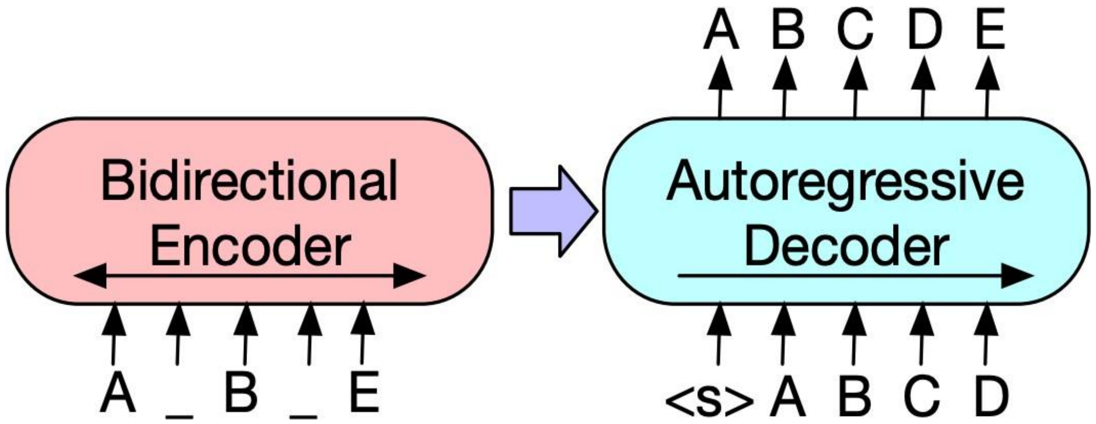


Lewis et al. (2019), “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”

BERT vs. BART

- **BERT:** only an encoder, trained with masked language modeling objective. Cannot generate text or do Seq2Seq tasks (in standard form).





- **BART:** consists of both an encoder and a decoder. Can also use just the encoder wherever we would use BERT.

BART for Summarization

- **Pre-train** on the BART task: take random chunks of text, noise them according to the schemes described, and try to “decode” the clean text
- **Fine-tune** on a summarization dataset: a news article is the input and a summary of that article is the output (usually 1-3 sentences depending on the dataset)
- Can achieve good results even with **few summaries to fine-tune on**, compared to basic seq2seq models which require 100k+ examples to do well

BART for Summarization: Output Example

PG&E stated it scheduled the blackouts in response to forecasts for high winds amid dry conditions. The aim is to reduce the risk of wildfires. Nearly 800 thousand customers were scheduled to be affected by the shutoffs which were expected to last through at least midday tomorrow.

↓

Power has been turned off to millions of customers in California as part of a power shutoff plan.

Lewis et al. (2019)

March 25, 2025

Google OpenAI

GEMINI 2.5

Gemini 2.5 models are **thinking models**, capable of reasoning through their thoughts before responding

Combines a significantly base model with **improved post-training**

Gemini 2.5 Pro **tops the LMArena leaderboard** by a significant margin.

Gemini 2.5 Pro scores **18.8% on Humanity's last exam** and leads in math and science benchmarks

OpenAI 4o Image Generation

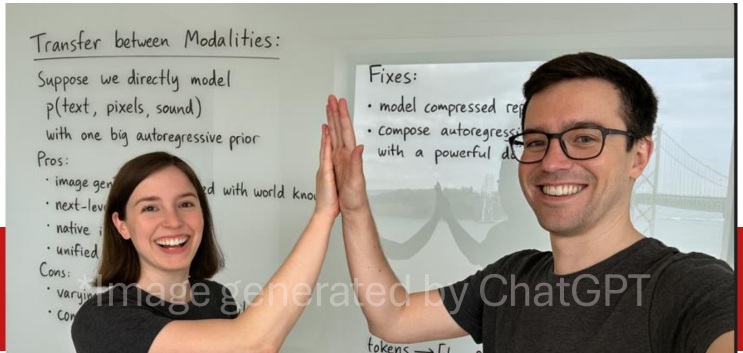
Image generation that is not only beautiful, but **useful**.

GPT-4o image generation excels at **accurately rendering text**, precisely **following prompts**, and **leveraging 4o's inherent knowledge base** and chat context

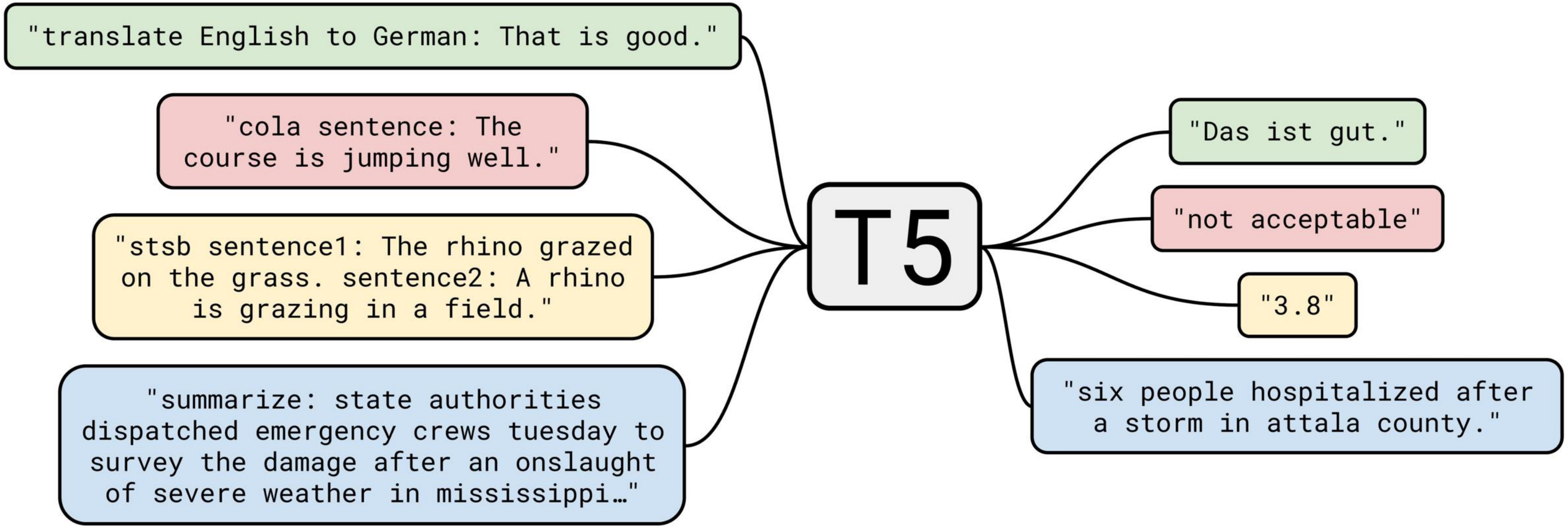
Trained models on the **joint distribution of online images and text**, and learnt how images relate to language and to each other

Two major announcements:

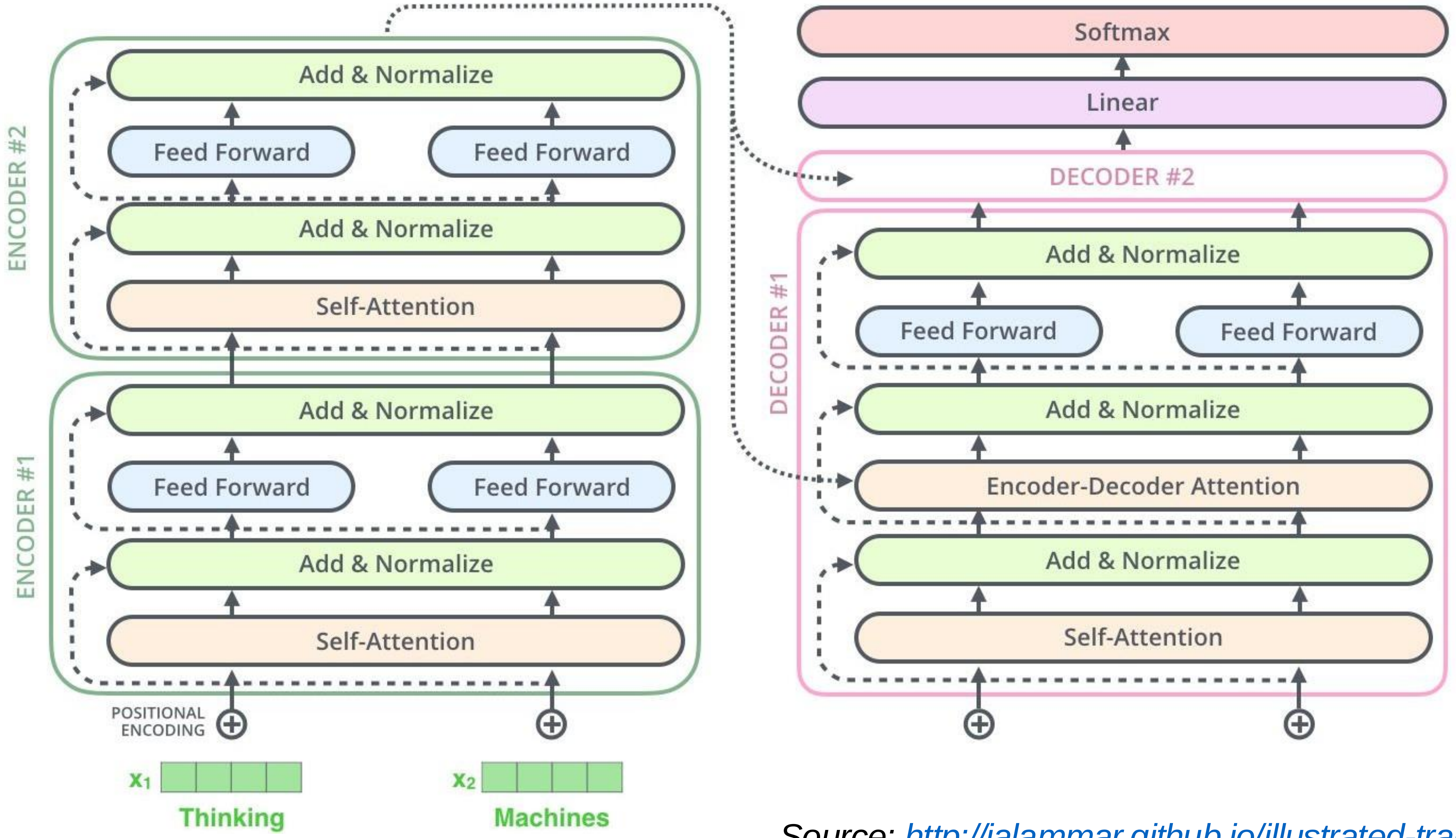
1) **Gemini 2.5 by Google - Advanced Reasoning**
2) **OpenAI 4o Image Generation - Mind-blowing Image generation**



T5: Text-to-Text Transfer Transformer



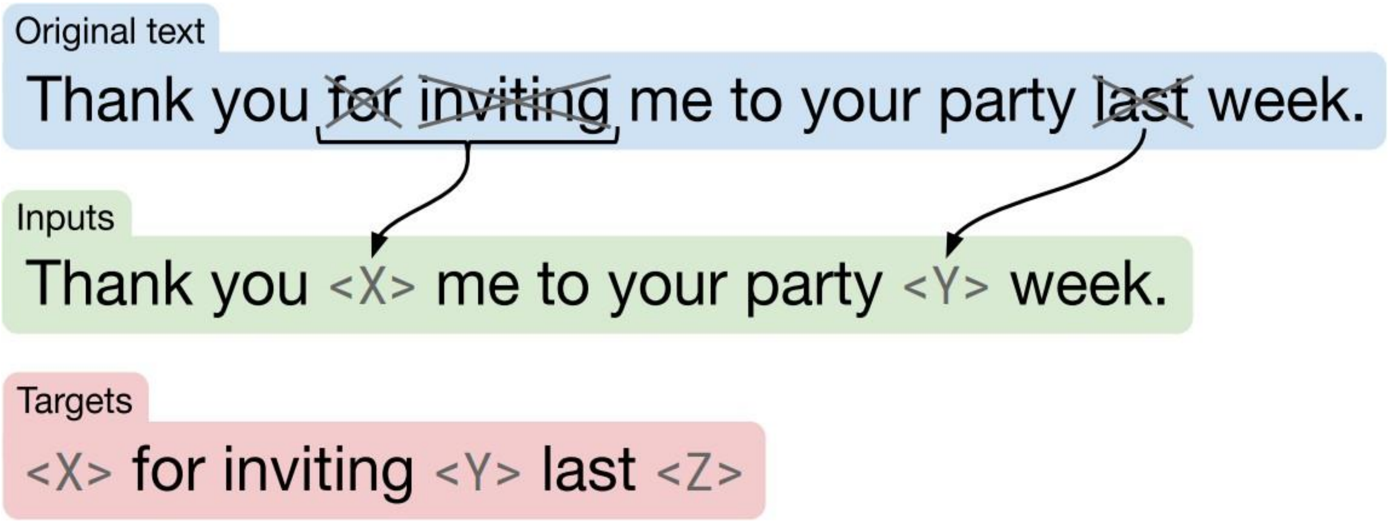
Raffel et al. (2019), "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"



Source: http://jalammar.github.io/illustrated-transformer/

Pre-Training T5

- **Pre-training:** similar denoising scheme to BART (they were released within a week of each other in fall 2019)
- **Input:** text with gaps ; **Output:** a series of phrases to fill those gaps.



Replace different-length spans from the input with unique placeholders; decode out the spans that were removed!

Raffel et al. (2019)

pital and largest city of the u
ma. the county seat of oklah
ty ranks 27th among united
tion. the population grew fol
s, with the population estim
ed to 643,648 as of july 201
oklahoma city metropolitan
of 1,358,452,[9] and the
hawnee combined statistic
of 1,459,758 residents,[9]
oma's largest metropolitan a

city limits extend into cana

running man was c
variety"; a genre of
environment.[1] the
complete missions
race.[2] the show h
familiar reality-vari
games. it has garn
comeback program
of the program, aft
family outing in feb

the show has beco
asia, and has gained online
hallyu fans, having been fansubbed into various
languages, such as english, spanish, portuguese,
french, italian, thai, vietnamese, chinese, ...

county,[8] the c
cities in popula
the 2010 census
to have increas
as of 2015, the
had a populatio
oklahoma city's
had a populati
making it oklah

oklahoma city's

the signing of the treaty formally ended the seven years' war, known as the french and indian war in the north american theatre,[1] and marked the beginning of an era of british dominance outside europe.[2] great britain and france each returned much of the territory that they had captured during the war, but great britain gained much of france's possessions in north america. additionally, great britain agreed to protect roman catholicism in the new world...

s a small hand-propelled vehicle,
one wheel, designed to be
ed by a single person using two
ar, or by a sail to push the
rrow by wind. the term
made of two words: "wheel" and
" is a derivation of the old
which was a device used for

is designed to distribute the
d between the wheel and the
operator, so enabling the convenient carriage of
heavier and bulkier loads than would be possible
were the weight carried entirely by the operator.
as such it is a second-class lever...

were the weight c
as such it is a second-class lever...

== piano

the piano is an acoustic, stringed musical instrument invented in italy by bartolomeo cristofori around the year 1700 (the exact year is uncertain), in which the strings are struck by hammers. it is played using a keyboard,[1] which is a row of keys (small levers) that the performer presses down or strikes with the fingers and thumbs of both hands to cause the hammers to strike the strings.

the word piano is a shortened form of pianoforte, the italian term for the early 1700s versions of the instrument, which in turn derives from gravicembalo col piano e forte[2] and fortepiano. the italian musical terms piano and forte indicate

== lemon

species of
lant family
y north

== wheelbarrow

a wheelbarrow is a small hand-propelled vehicle, usually with just one wheel, designed to be pushed and guided by a single person using two handles at the rear, or by a sail to push the ancient wheelbarrow by wind. the term "wheelbarrow" is made of two words: "wheel" and "barrow." "barrow" is a derivation of the old english "bearwe" which was a device used for carrying loads.

paris, also known as the treaty
gned on 10 february 1763 by th
great britain, france and spain,
greement, after great britain's v
nd spain during the seven year

f the treaty formally ended the
nown as the french and indian
rican theatre,[1] and marked
an era of british dominance ou
eat britain and france each ret
territory that they had captured
ar, but great britain gained muc
sessions in north america.

Common Crawl Web Extracted Text

Menu

Lemon

Introduction

The lemon, Citrus Limon (l.) Osb
species of small evergreen tree
flowering plant family rutaceae
The tree's ellipsoidal yellow frui
culinary and non-culinary purpos
throughout the world, primarily
which has both culinary and cle
The juice of the lemon is about
citric acid, with a ph of around 2
a sour taste.

Article

The origin of the lemon is unknown, though
lemons are thought to have first grown in
Assam (a region in northeast India),
northern Burma or China.
A genomic study of the lemon indicated it
was a hybrid between bitter orange (sour
orange) and citron.

Please enable JavaScript to use our site.

Home
Products
Shipping
Contact

The tree's ellipsoidal yellow fruit is used for
culinary and non-culinary purposes
throughout the world, primarily for its juice,
which has both culinary and cleaning uses.
The juice of the lemon is about 5% to 6%
citric acid, with a ph of around 2.2, giving it
a sour taste.

Lorem ipsum dolor sit amet, consectetur
adipiscing elit.
Curabitur in tempus quam. In mollis et ante
at consectetur.
Aliquam erat volutpat.
Donec at lacinia est.
erdum
eget
e lacinia at
fficitur
em non,
a sodales in

```
function Ball(r) {
   this.radius  = r;
   this.area = pi * r ** 2;
   this.show = function(){
      drawCircle(r);
   }
}
```

- Removed lines that didn’t end in a terminal punctuation mark.
- Language classifier to retain only English text
- Removed texts which look like placeholder texts
- Removed anything which look like code
- Removed duplicated texts

TensorFlow Resources More Search Language GitHub Sign in

Datasets v1.3.2

Overview Catalog Guide API

TensorFlow > Resources > Datasets v1.3.2 > Catalog

c4 (Manual download)

Contents

A colossal, cleaned version of Common Crawl's web crawl corpus.

Pretrain

$\text{BERT}_{\text{BASE}}$-sized encoder-decoder Transformer

Denoising objective

C4 dataset

Finetune

GLUE

Pretrain

$\text{BERT}_{\text{BASE}}$-sized encoder-decoder Transformer

Denoising objective

C4 dataset

GLUE Benchmark

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

Pretrain

$\text{BERT}_{\text{BASE}}$-sized encoder-decoder Transformer

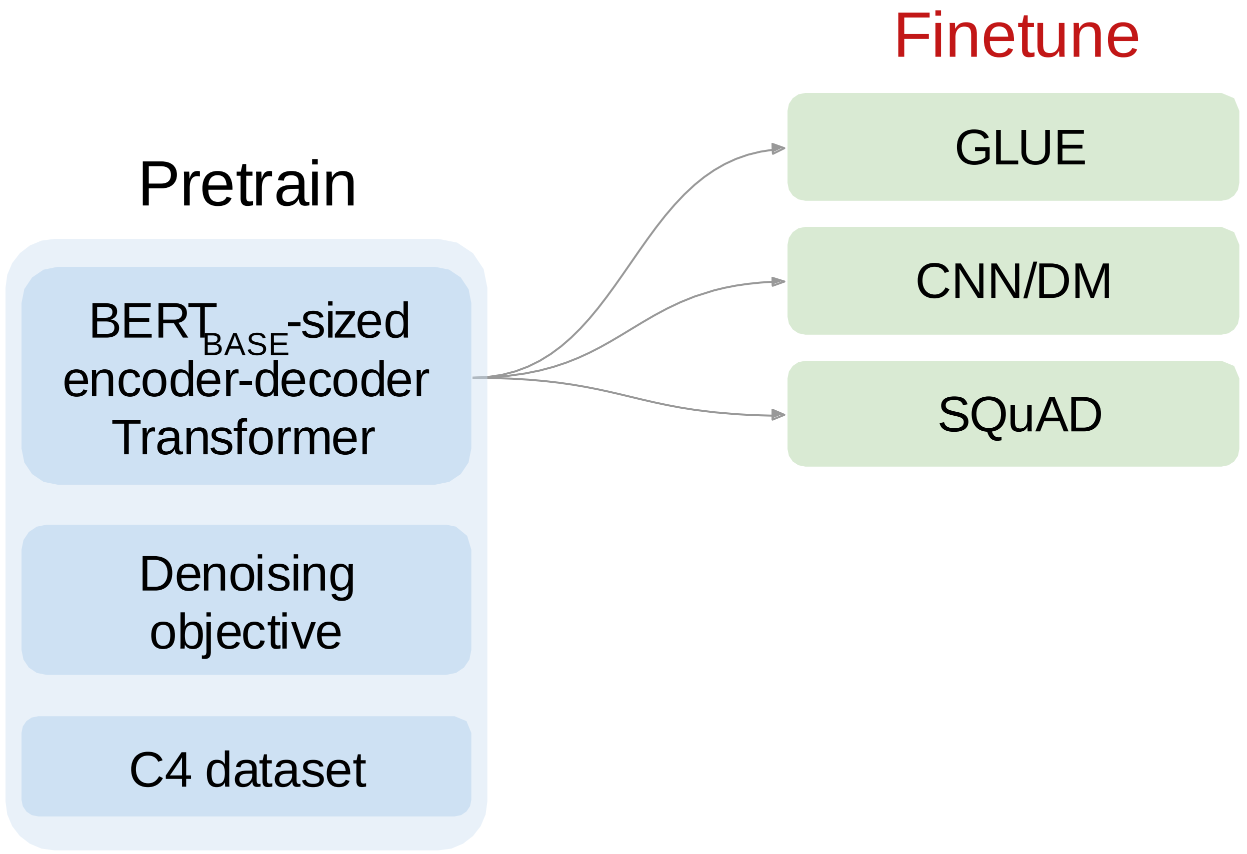
Denoising objective

C4 dataset

Finetune

GLUE

CNN/DM


Finetune
Pretrain
BERT$_{BASE}$-sized encoder-decoder Transformer
Denoising objective
C4 dataset
GLUE
CNN/DM
SQuAD

Finetune

GLUE

CNN/DM

SQuAD

SuperGLUE

Pretrain

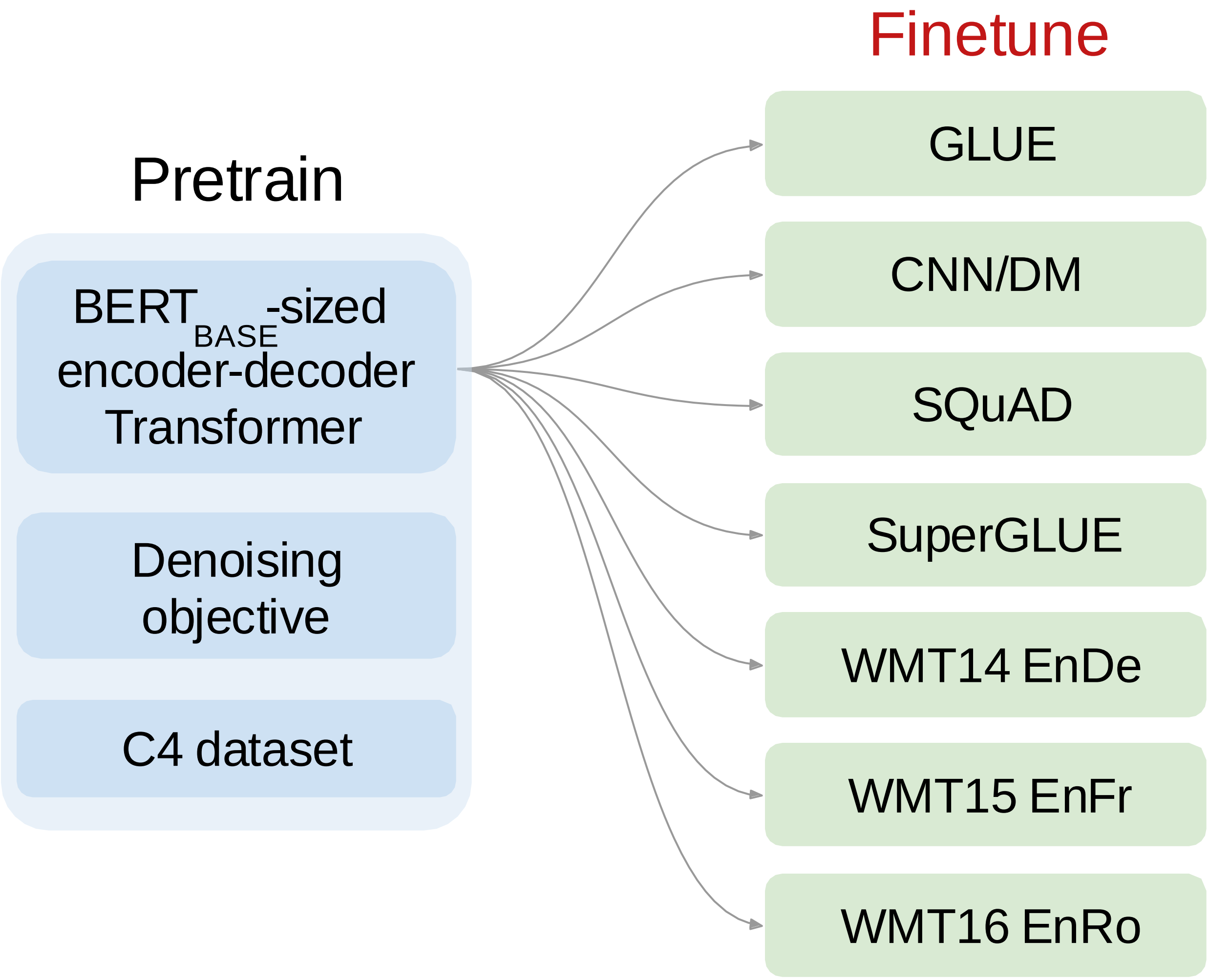
BERT$_{\text{BASE}}$-sized encoder-decoder Transformer

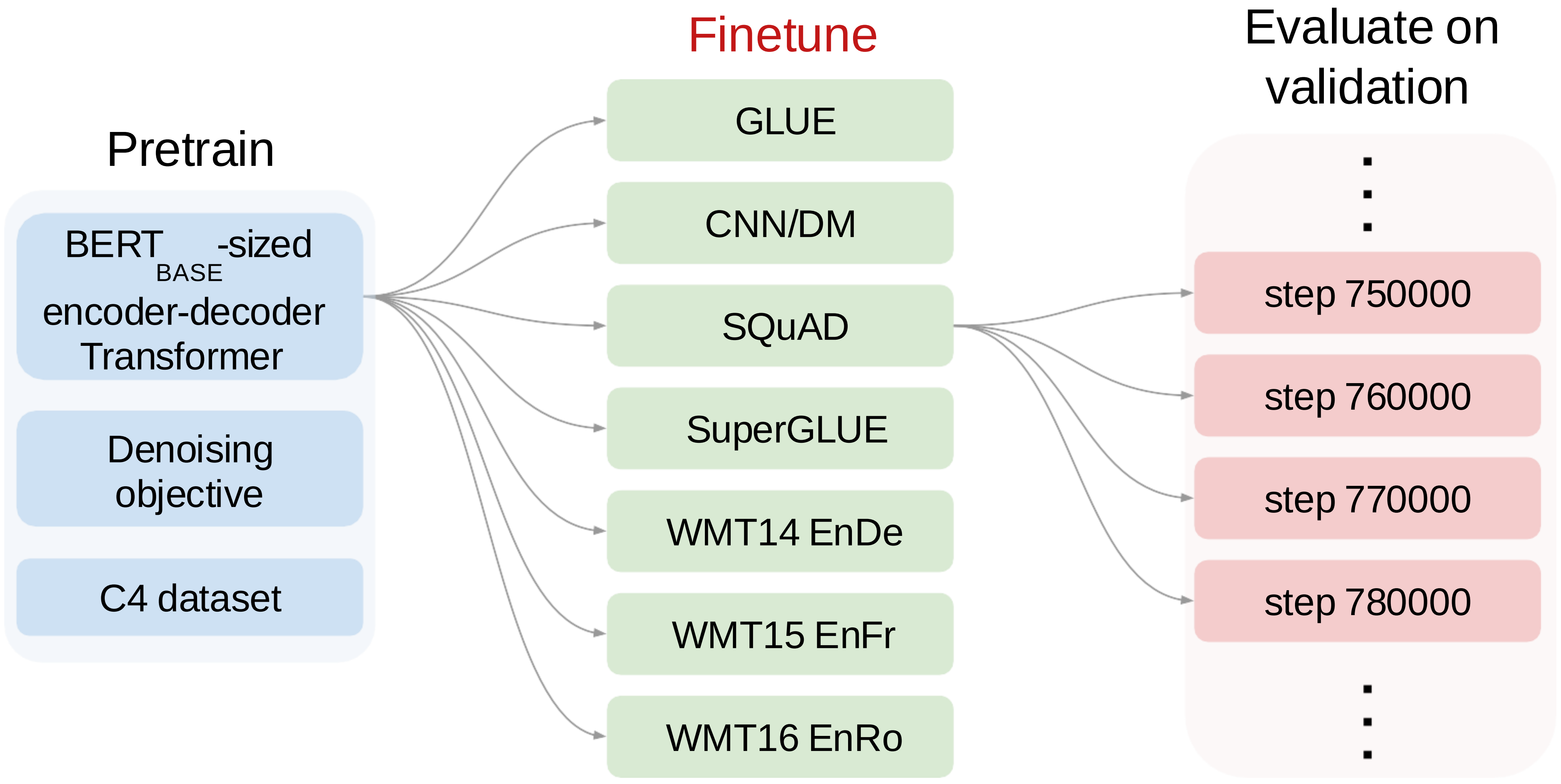
Denoising objective

C4 dataset

SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy


Finetune
Pretrain
BERT$_{BASE}$-sized encoder-decoder Transformer
Denoising objective
C4 dataset
GLUE
CNN/DM
SQuAD
SuperGLUE
WMT14 EnDe
WMT15 EnFr
WMT16 EnRo


Pretrain
BERT$_{BASE}$-sized encoder-decoder Transformer
Denoising objective
C4 dataset
Finetune
GLUE
CNN/DM
SQuAD
SuperGLUE
WMT14 EnDe
WMT15 EnFr
WMT16 EnRo
Evaluate on validation
step 750000
step 760000
step 770000
step 780000

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	**39.77**	24.04

Star denotes baseline

Comparable to BERT

Bold = 1 std. dev. of max

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	**39.77**	24.04

Big training set

No pre-training is dramatically worse, except EnFr!

C4: The Data

- C4: Colossal Clean Crawled Corpus
  - Web-extracted text
  - English language only
  - 750GB

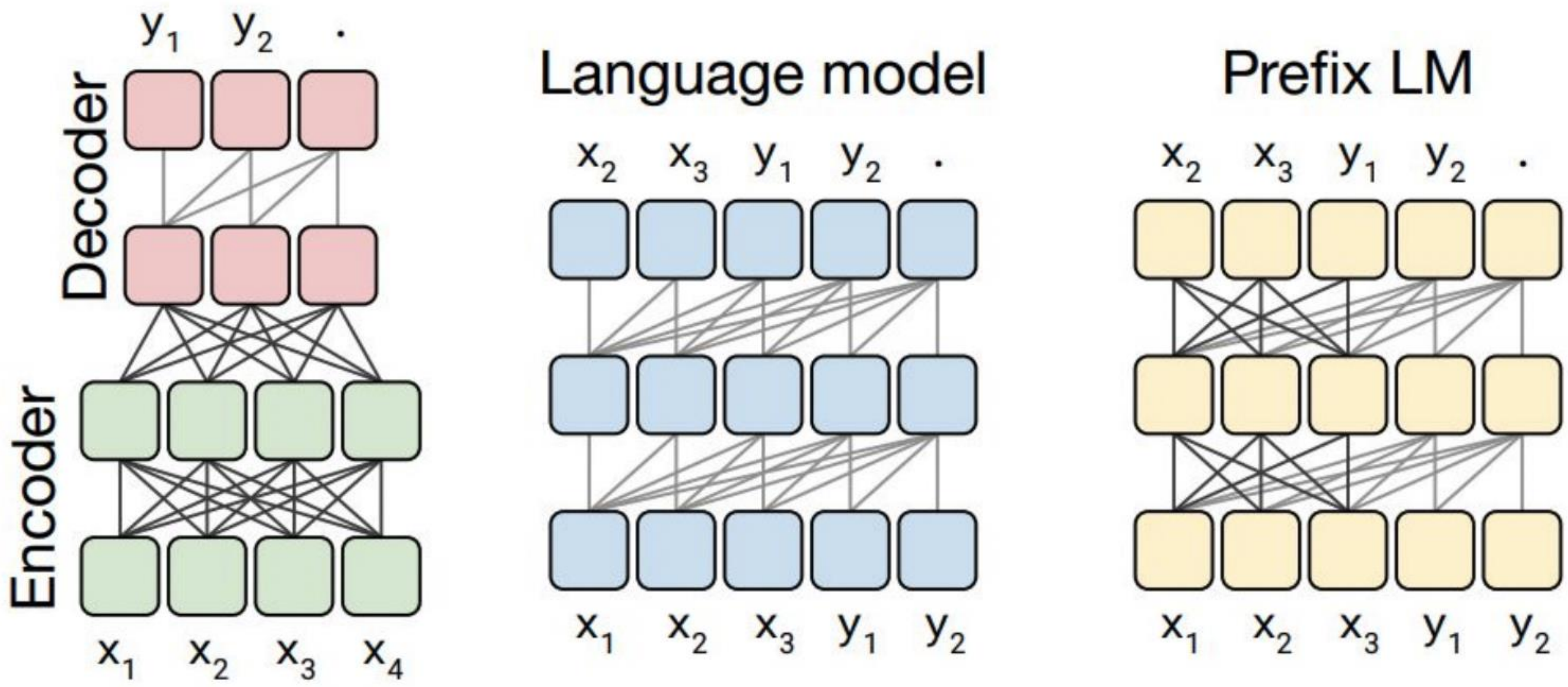
	Data set	Size
★	C4	745GB
	C4, unfiltered	6.1TB

Pre-Training Data: Experiment

- Takeaway:
  - Clean and compact data is better than large, but noisy data.
  - Pre-training on in-domain data helps.

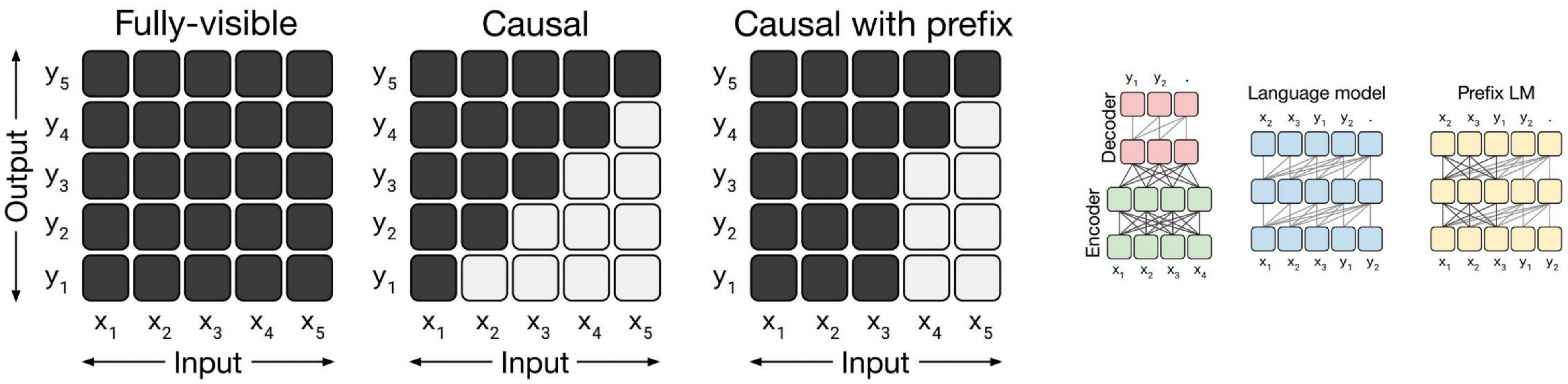
Data set	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21

Architectures: Different Choices



Architectures: Different Attention Masks

- **Fully visible mask** allows the self attention mechanism to attend to the full input.
- A **causal mask** doesn't allow output elements to look into the future.
- **Causal mask with prefix** allows to fully-visible masking on a portion of input.



Architectural Variants: Experiments

Evaluated for classification tasks.

Architecture	Objective	Params	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Encoder-decoder	Denoising	$2P$	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	81.82	18.61	78.94	68.11	26.43	37.98	27.39

Takeaways:

1. Halving the number of layers in encoder and decoder hurts the performance.
2. Performance of Encoder-Decoder with shared params is almost on-par with prefix LM.

T5: Pre-Training Objectives

- **Prefix language modeling**
  - **Input:** Thank you for inviting
  - **Output:** me to your party last week
- **BERT-style denoising**
  - **Input:** Thank you <M> <M> me to your party apple week
  - **Output:** Thank you for inviting me to your party last week
- **De-shuffling**
  - **Input:** party me for your to. last fun you inviting week Thanks.
  - **Output:** Thank you for inviting me to your party last week
- **Replace spans**
  - **Input:** Thank you <X> me to your party <X> week
  - **Output:** <X> for inviting <Y> last <Z>
- **Drop tokens**
  - **Input:** Thank you me to your party week .
  - **Output:** for inviting last

Pre-Training Objectives: Experiments

- All the variants perform similarly
- "Replace corrupted spans" and "Drop corrupted tokens" are more appealing because **target sequences are shorter, speeding up training**.

Assuming Enc-Dec architecture.
Evaluated for classification tasks.

Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Prefix language modeling	80.69	18.94	77.99	65.27	**26.86**	39.73	**27.49**
Deshuffling	73.17	18.59	67.61	58.47	26.11	39.30	25.62
BERT-style (Devlin et al., 2018)	82.96	19.17	**80.65**	69.85	26.78	**40.03**	27.41
★ Replace corrupted spans	83.28	**19.24**	**80.88**	**71.36**	**26.98**	39.82	**27.65**
Drop corrupted tokens	**84.44**	**19.31**	**80.52**	68.67	**27.07**	39.76	**27.82**

Pre-Training Decoder-only Models

GPT and Llama

Recall: Probabilistic Language Models

$$\underbrace{P(X)}$$

Sentence/Document

A generative model that calculates the probability of language

Auto-regressive Language Models

$$P(X) = \prod_{i=1}^{I} P(\underbrace{x_i}_{\text{Next Token}} \mid \underbrace{x_1, \ldots, x_{i-1}}_{\text{Context}})$$

Next Token Prediction

- This is essentially **classification**!
  - We can think of neural language models as neural classifiers. They classify prefix of a text into |V| classes, where the classes are vocabulary tokens.

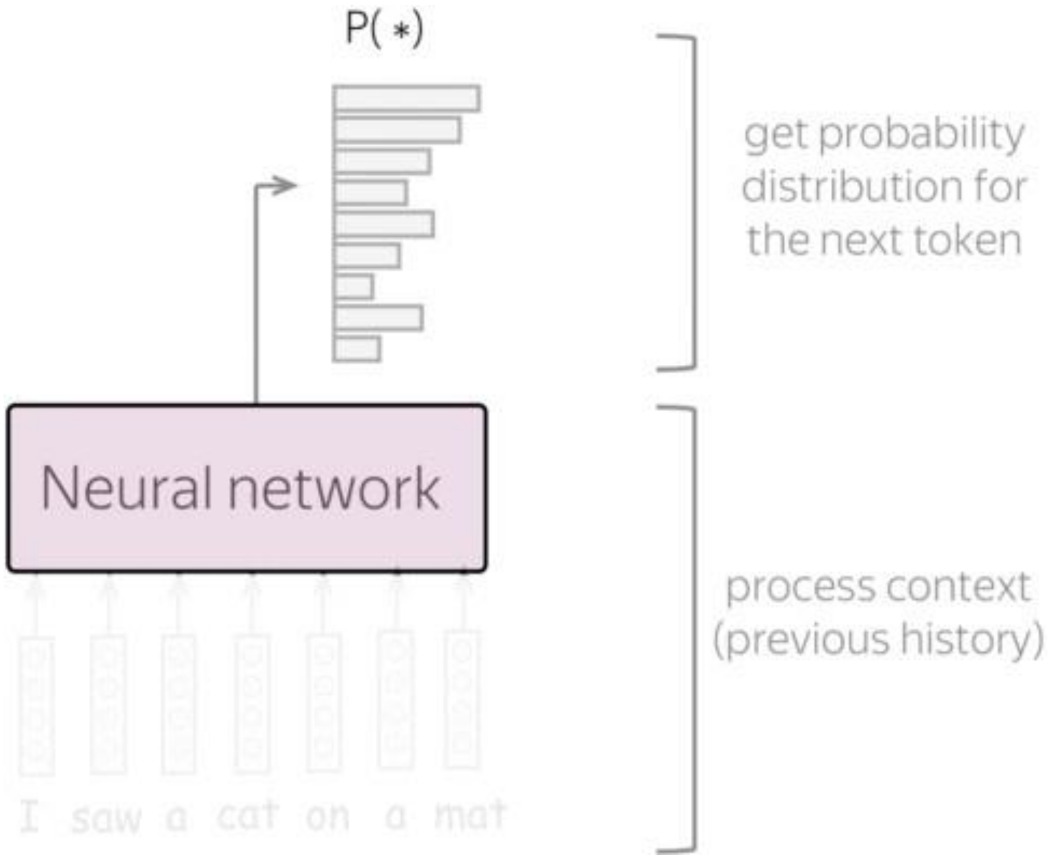
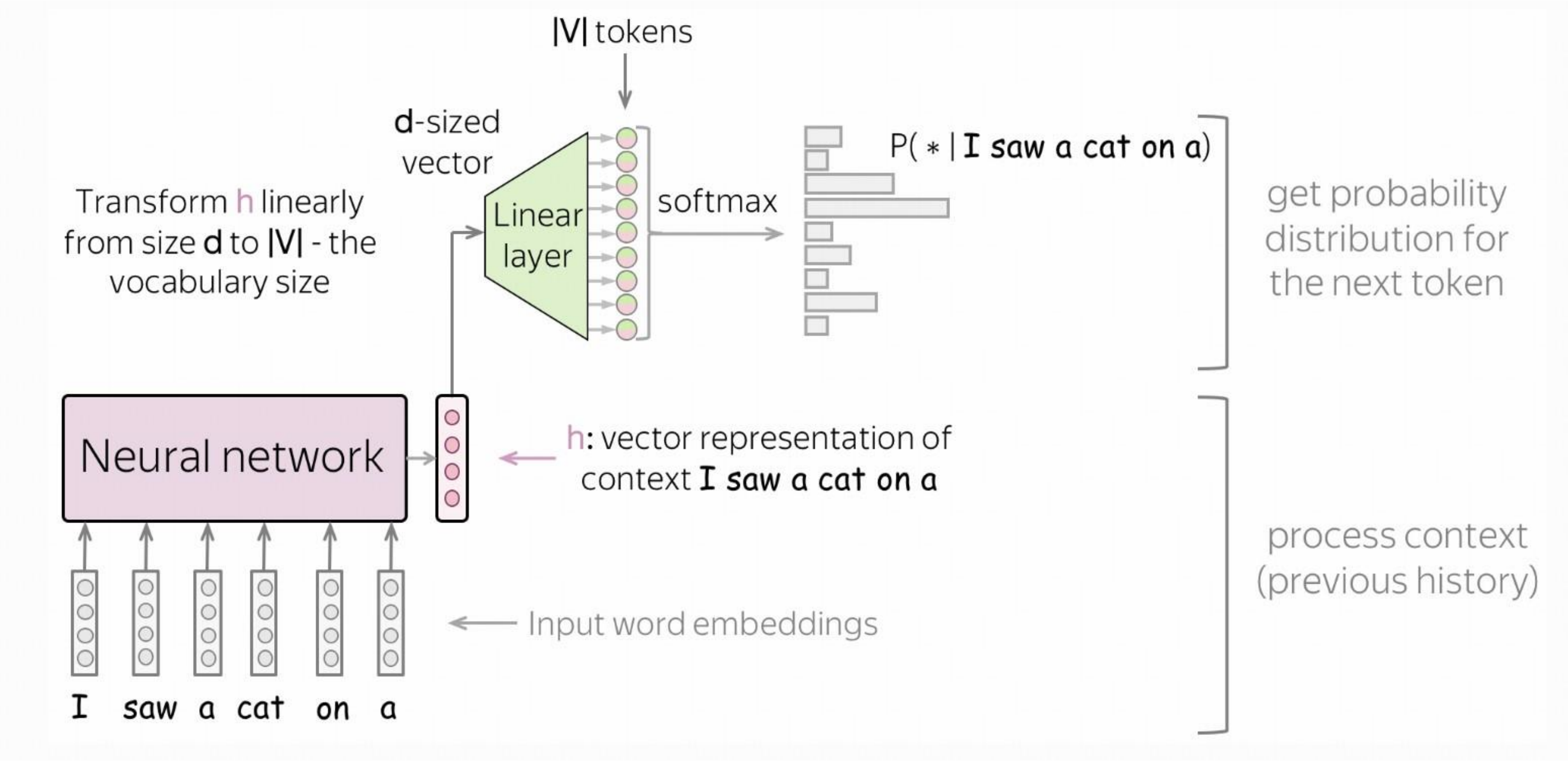


Image Credit: https://lena-voita.github.io/nlp_course/language_modeling.html

Next Token Prediction



- Feed word embedding for previous (context) words into a network.
- Get vector representation of context from the network.
- From this vector representation, predict a probability distribution for the next token.

Image Credit: https://lena-voita.github.io/nlp_course/language_modeling.html

Encoders vs. Decoders

- BERT is a Transformer **Encoder**: bidirectional attention, trained with masked language modelling.

$$P\,(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

- GPT and many other Transformer language models (e.g., LLaMA) are **Decoders**: unidirectional attention, **trained to predict the next token**.

$$P\,(x_i \mid x_1, \ldots, x_{i-1})$$

Generative Pre-trained Transformer (GPT)

- 2018's GPT was a big success in pretraining a decoder!
- **Transformer decoder with 12 layers, 117M parameters**.
- 768-dimensional hidden states, 3072-dimensional feed-forward hidden layers.
- Byte-pair encoding with 40,000 merges
  - Trained on BooksCorpus: over 7000 unique books.
- Contains long spans of contiguous text, for learning long-distance dependencies.

Radford et al. (2018), "Improving Language Understanding by Generative Pre-Training"

GPT-2

GPT-2 is identical to GPT-1, but:

- Has Layer normalization in between each sub-block
- Vocab extended to 50,257 tokens and context size increased from 512 to 1024
- Data: 8 million docs from the web (Common Crawl), minus Wikipedia

---

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

Increasingly Convincing Generations by GPT-2

- We discussed how we can sample sentences from auto-regressive LMs for text generation.
  - This is how pre-trained decoders are used **in their capacities as language models.**
- **GPT-2,** a larger version (1.5B) of GPT trained on more data, was shown to produce relatively convincing samples of natural language.

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

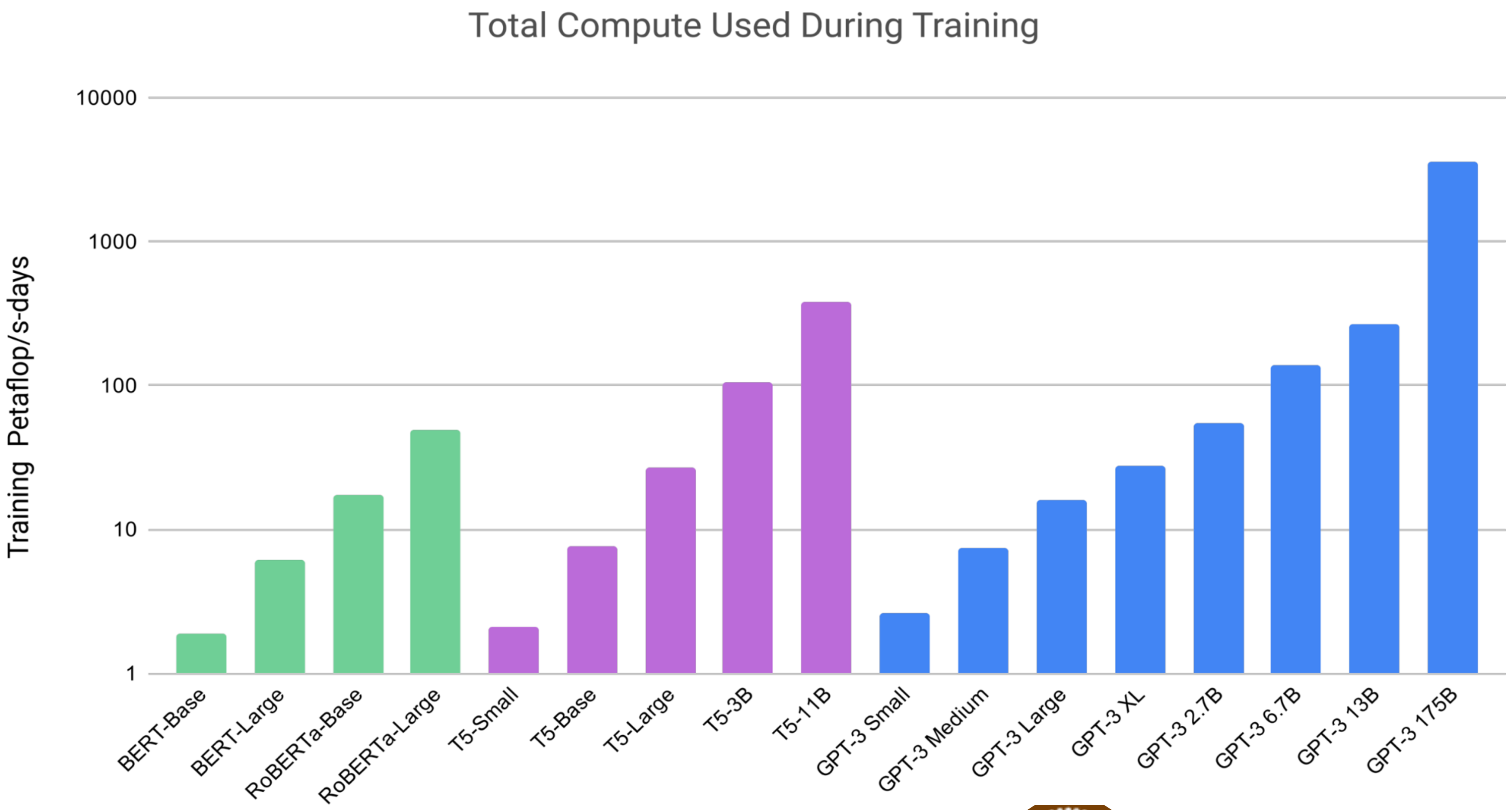
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pre-Training Cost (with GCP/AWS)

- **BERT:** Base $500, Large $7000
- **GPT-2** (as reported in other work): $25,000
- This is for a single pre-training run…developing new pre-training techniques may require many runs.
- Fine-tuning these models can typically be done with a single GPU (but may take 1-3 days for medium-sized datasets).

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

GPT-3



- **175B parameter model**
  - 96 layers, 96 heads, 12k-dim vectors
- Trained on Microsoft Azure, estimated to cost roughly **$10M**

Comparison: GPT-1, 2, 3

Model	Parameters	Layers	Training Data	Key Advancement
GPT-1	117M	12	BooksCorpus	First large-scale Transformer for NLP
GPT-2	1.5B	48	WebText	Zero-shot learning, larger training data
GPT-3	175B	96	Common Crawl + others	In-context learning, emergent behaviors

GPT-4

Model	Usage
davinci-002	$0.0020 / 1K tokens

Model	Input	Output
gpt-4	$0.03 / 1K tokens	$0.06 / 1K tokens

- Transformer-based
  - The rest is .... mystery!
  - If we're going based on costs, GPT-4 is ~15-30 times costlier than GPT3. That should give you an idea how its likely size!

- Note, these language models involve more than just pre-training.
  - Pre-training provides the foundation based on which we build the model.
  - We will discuss the later stages next week.

Llama: A Family of Open-Source LLMs from Meta AI

- **Llama-1 + Llama-2**

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2: **Model sizes, architectures, and optimization hyper-parameters.**

- Models have mostly gotten smaller since GPT-3, but haven't changed much.
- **Tokenizer: Byte-Pair Encoding (BPE)** [Recall: we have already discussed this algorithm in lecture on 'Tokenization Strategies']
- **Rotary positional encodings**, a few other small architecture changes
- **Optimized mix of pre-training data**: Common Crawl, GitHub, Wikipedia, Books, etc.

Next Week: How to Make Pre-Trained LMs Work?

- Instruction Tuning
  - Finetune the pre-trained model to follow instructions
- Prompting and In-context Learning
  - Give few examples of the task that you want the model to solve
- Reinforcement Learning from Human Feedback (RLHF)
  - Train the LMs to align their outputs with human preferences
  - Also called ‘Preference Optimization’