Large Language Models

CSCI 601 471/671
NLP: Self-Supervised Models

https://self-supervised.cs.jhu.edu/sp2023/

JOHNS HOPKINS
UNIVERSITY

[Slide credit: Chris Tanner, Jacob Devlin and many others]

Logistics Update

- The midterm:
  - will be on March 7 during class time.
  - I will not be here; Adam (TA) will run the show.
  - it will be on paper
  - Based on the ideas you have seen in homework and lectures. If you understand them, you're set!
  - Scope HW 1-5 and lectures until Feb 23

- Post questions only on Piazza (no direct DM to course staff).

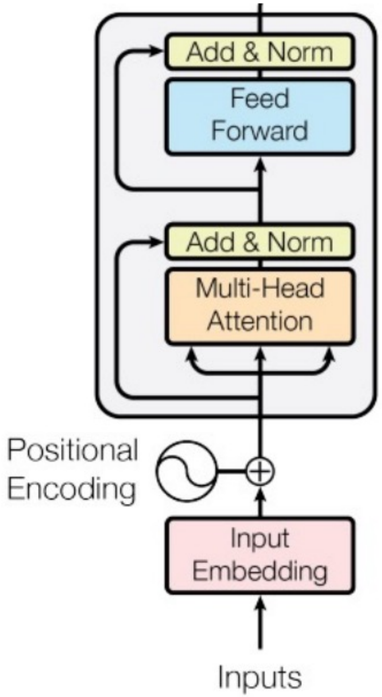
- Since had less HW than expected:
  - (1) Semi-weekly assignments (60%), → now 50%
  - (2) midterm exam (20%), → now 20%
  - (3) a final project (20%) → now 30%

Recap: Attention Block

Given input $\mathbf{x}$:

$$Q = \mathbf{W}^q \mathbf{x}$$
$$K = \mathbf{W}^k \mathbf{x}$$
$$V = \mathbf{W}^v \mathbf{x}$$

$$\text{Attention}(\mathbf{x}) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{h}}\right)V$$



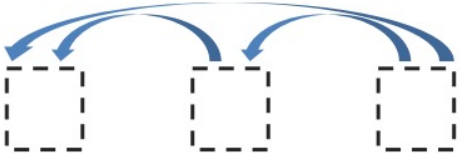
[Attention Is All You Need, Vaswani et al. 2017]

Recap: Transformer [Vaswani et al. 2017]

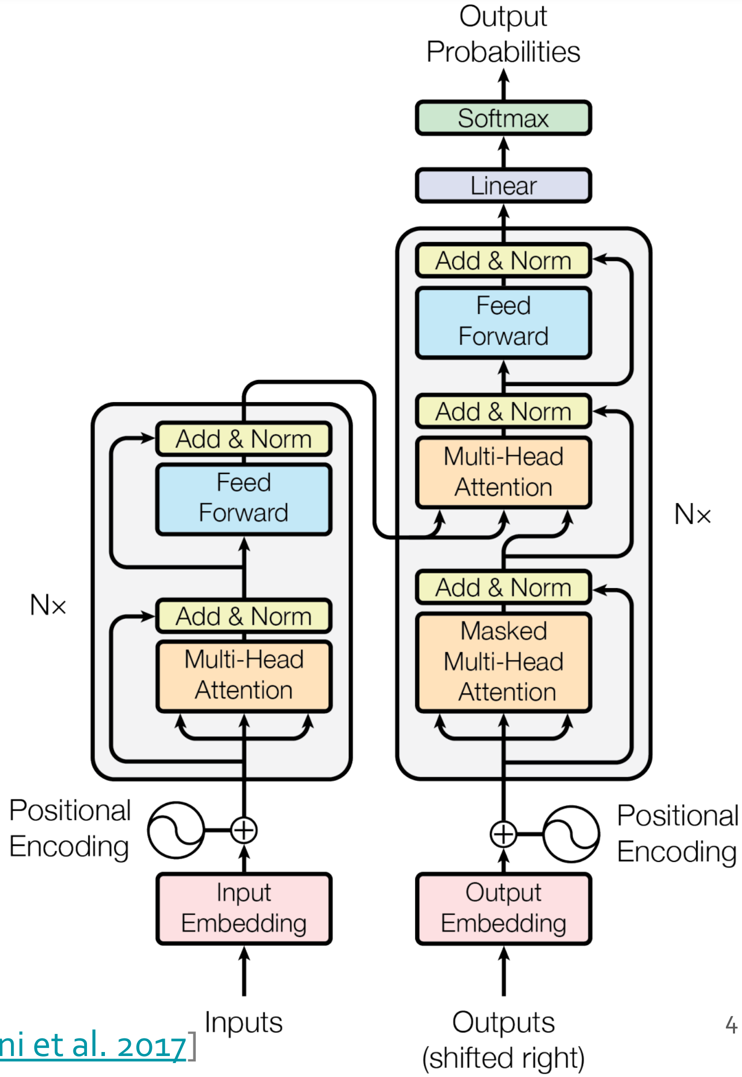
- An encoder-decoder architecture
- 3 forms of attention

Encoder-Decoder Attention

Encoder Self-Attention



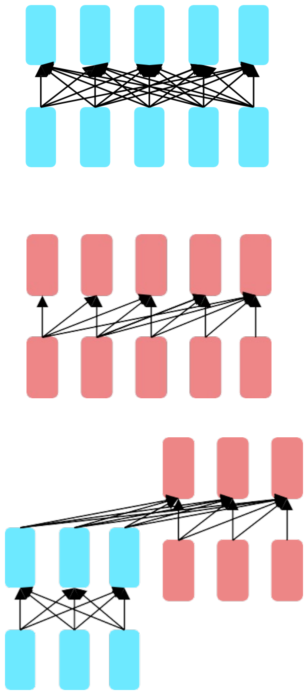
MaskedDecoder Self-Attention



[Attention Is All You Need, Vaswani et al. 2017]

Impact of Transformers

- A building block for a variety of LMs



Encoders

Decoders

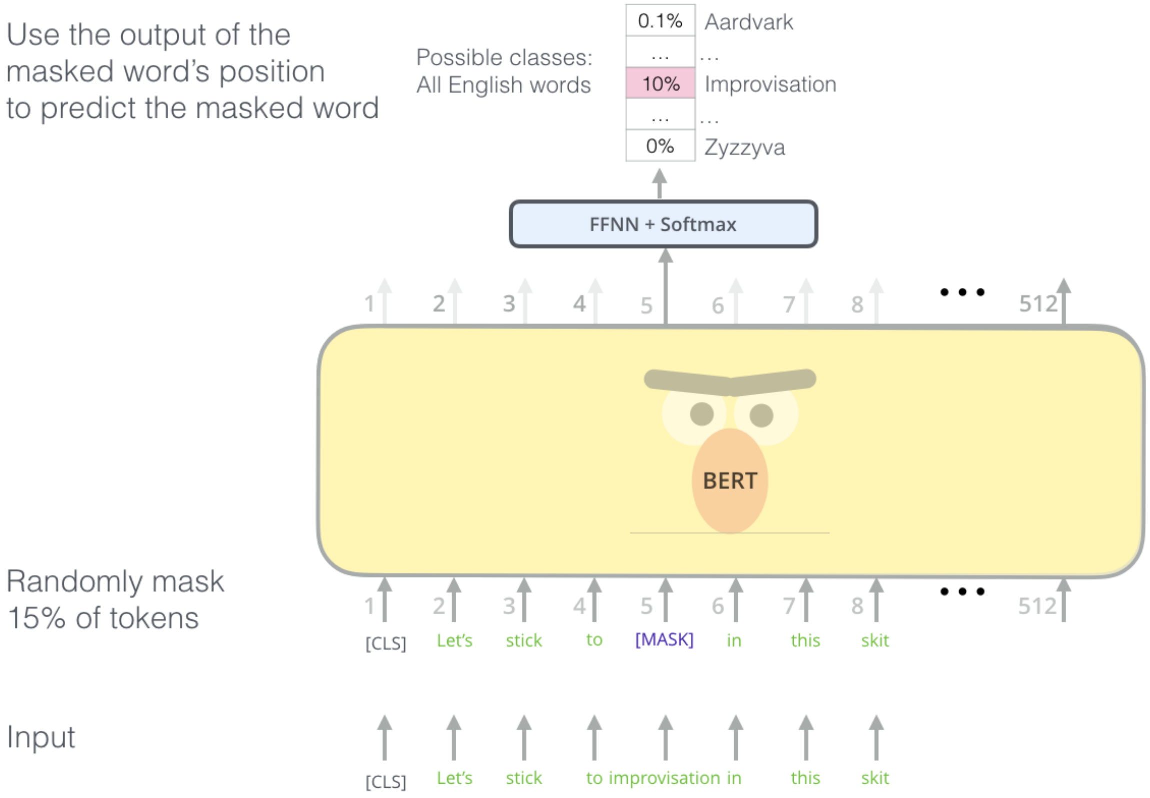
Encoder-
Decoders

BERT

Bidirectional Encoder Representations from Transformers

BERT: Pre-training Objective (1): Masked Tokens

- Randomly mask 15% of the tokens and train the model to predict them.

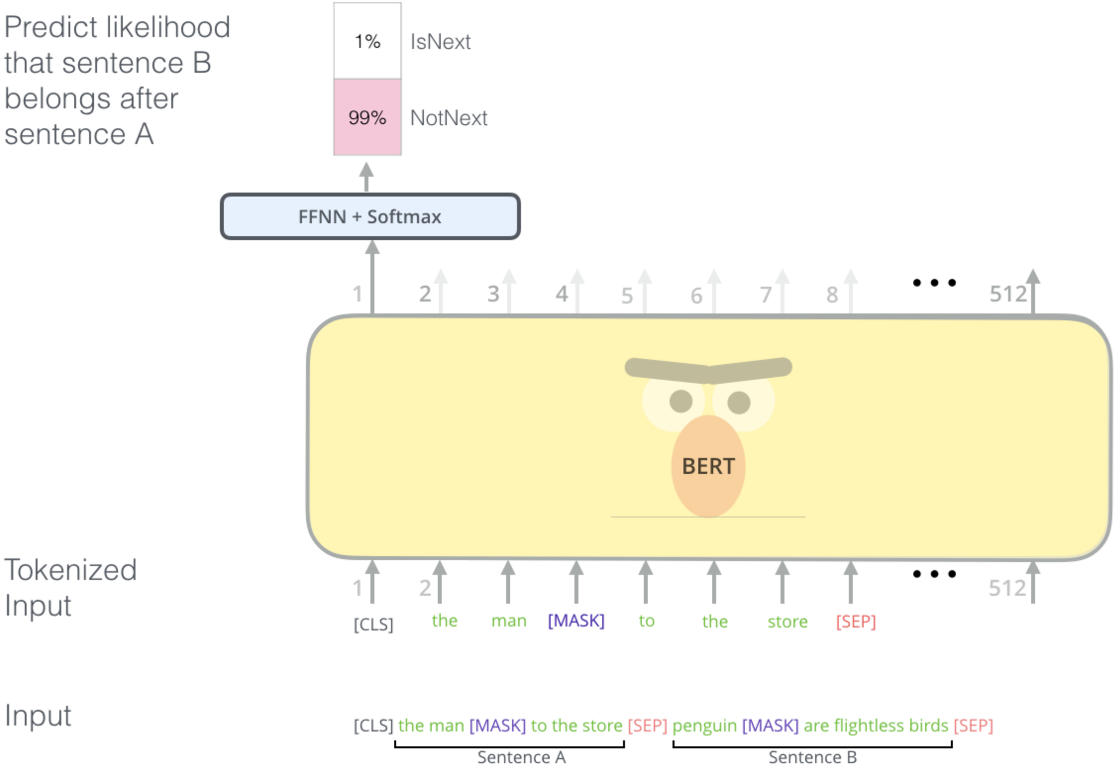


[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

BERT: Pre-training Objective (2): Sentence Ordering

- Predict sentence ordering

- 50% correct ordering, and
  50% random incorrect ones



[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

Text generation using BERT

BERT has a Mouth, and It Must Speak:
BERT as a Markov Random Field Language Model

Alex Wang
New York University
alexwang@nyu.edu

Kyunghyun Cho
New York University
Facebook AI Research
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Mask-Predict: Parallel Decoding of
Conditional Masked Language Models

Marjan Ghazvininejad* **Omer Levy*** **Yinhan Liu*** **Luke Zettlemoyer**
Facebook AI Research
Seattle, WA

Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis--Hastings

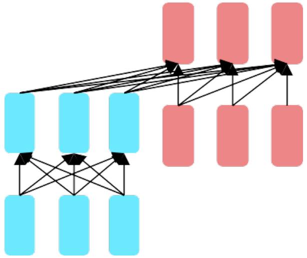
Kartik Goyal, Chris Dyer, Taylor Berg-Kirkpatrick

src	Der Abzug der franzsischen Kampftruppen wurde am 20. November abgeschlossen .
$t = 0$	The departure of the French combat completed completed on 20 November .
$t = 1$	The departure of French combat troops was completed on 20 November .
$t = 2$	The withdrawal of French combat troops was completed on November 20th .

Leveraging Pre-trained Checkpoints for Sequence Generation Tasks

Sascha Rothe, Shashi Narayan, Aliaksei Severyn

BART/T5



Encoder-
Decoders

T5: Text-To-Text Transfer Transformer (2019)

- An encoder-decoder architecture

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel* CRAFFEL@GMAIL.COM
Noam Shazeer* NOAM@GOOGLE.COM
Adam Roberts* ADAROB@GOOGLE.COM
Katherine Lee* KATHERINELEE@GOOGLE.COM
Sharan Narang SHARANNARANG@GOOGLE.COM
Michael Matena MMATENA@GOOGLE.COM
Yanqi Zhou YANQIZ@GOOGLE.COM
Wei Li MWEILI@GOOGLE.COM
Peter J. Liu PETERJLIU@GOOGLE.COM
Google, Mountain View, CA 94043, USA

- But it's more than just a model paper
- The paper conducts an in-depth analysis of various parameters of model design

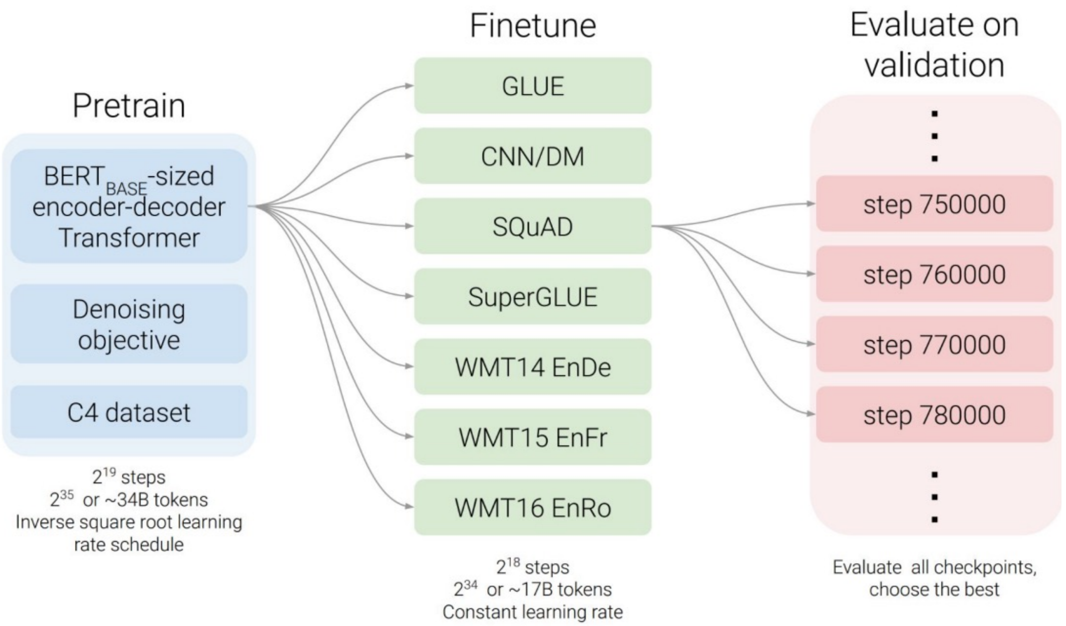
Various Aspects of Model Design

- Architectures
- Pre-training masking
- Pre-training dataset
- Scale of the pre-training

- **Goal:** Understand the first order effect of each design choice by altering it while keeping other choices fixed.

Experimental Setup

- Decide a default model
  - Encoder-decoder architecture
  - Noising objective
  - ....
- Evaluate a design axis, fixing the rest of the parameters



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`
- Deshuffling
  - **Input:** `party me for your to. last fun you inviting week Thanks.`
  - **Output:** `Thank you for inviting me to your party last week`
- IID noise, replace spans
  - **Input:** `Thank you <X> me to your party <X> week`
  - **Output:** `<X> for inviting <Y> last <Z>`
- IID noise, drop tokens
  - **Input:** `Thank you me to your party week .`
  - **Output:** `for inviting last`

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.

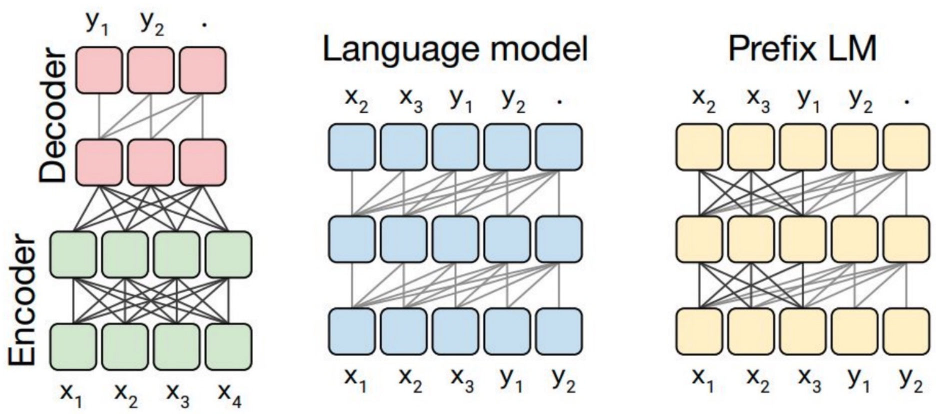
Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Prefix language modeling	80.69	18.94	77.99	65.27	**26.86**	39.73	**27.49**
Deshuffling	73.17	18.59	67.61	58.47	26.11	39.30	25.62
BERT-style (Devlin et al., 2018)	82.96	19.17	**80.65**	69.85	26.78	**40.03**	27.41
★ Replace corrupted spans	83.28	**19.24**	**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**

Objectives: Experiments

- Performance of the i.i.d. corruption objective with different corruption rates
- Takeaway:
  - Little corruption rate may prevent effective learning.
  - Larger corruption rate leads to downstream performance degradation.
  - Larger corruption rate also leads to longer targets, slowing down training.

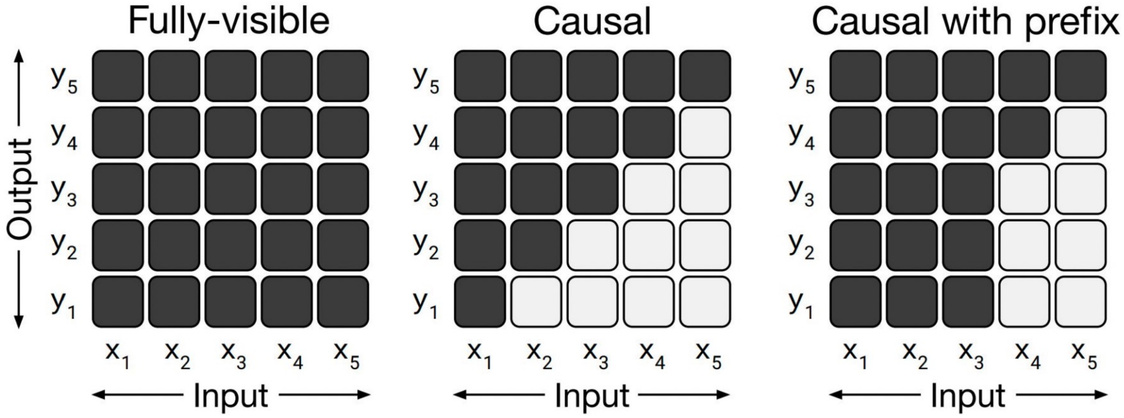
Corruption rate	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
10%	**82.82**	19.00	**80.38**	69.55	**26.87**	39.28	**27.44**
★ 15%	**83.28**	19.24	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
25%	**83.00**	**19.54**	**80.96**	70.48	**27.04**	**39.83**	**27.47**
50%	81.27	19.32	79.80	70.33	**27.01**	**39.90**	**27.49**

Architectures: Different Choices



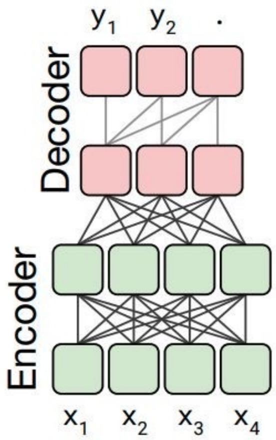
Architectures: Different Attention Masks

- **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

	Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★	Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**

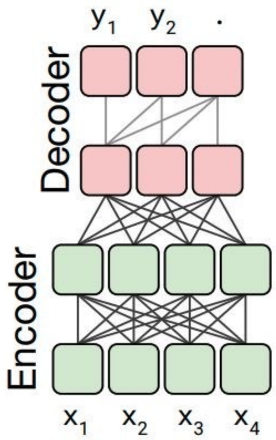


[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**

Input: Thank you for <X> me to your party <Y>.
Target: <X> inviting <Y> last week.

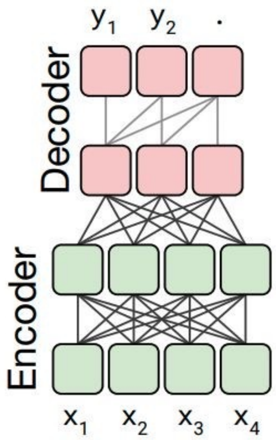


[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**

Number of parameters

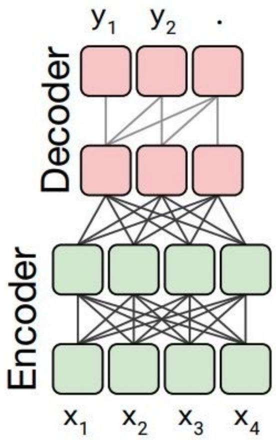


[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**

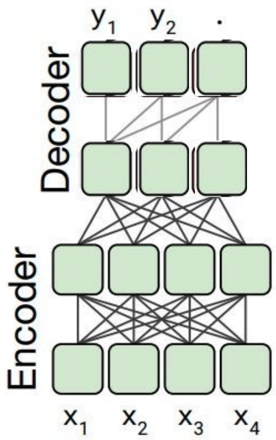
Number of flops



[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

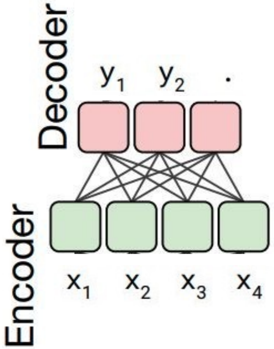
Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**



[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

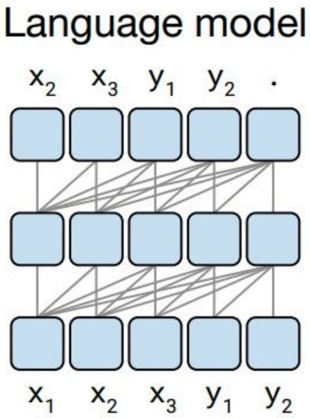
Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95



[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

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

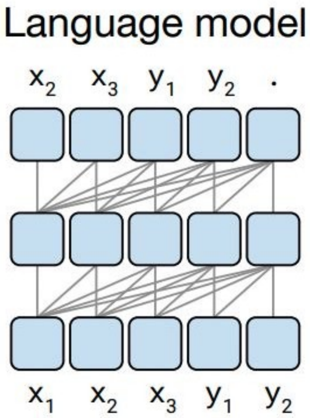


[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

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

Language model is decoder-only

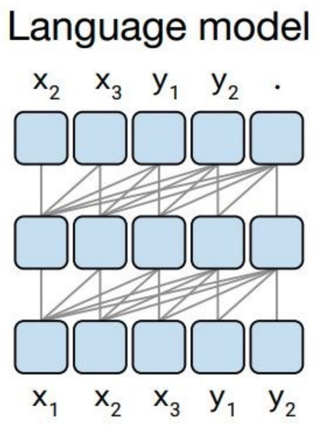


[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

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

LM looks at both input and target, while encoder only looks at input sequence and decoder looks at output sequence.

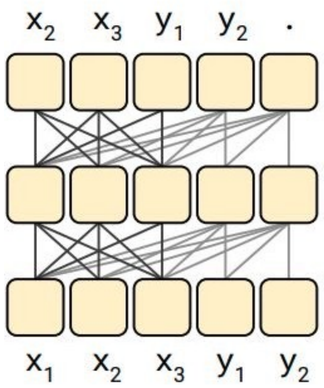


[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

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

Prefix LM



[slide credit:Abhishek Panigrahi, Victoria Graf]

Architectural Variants: Experiments

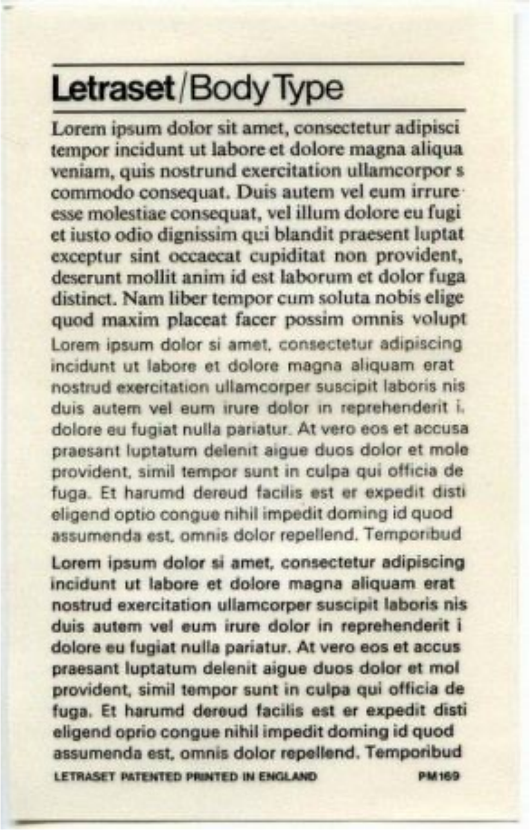
Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88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

- Takeaways:
  - Halving the number of layers in encoder and decoder hurts the performance.
  - Performance of Encoder and Decoder with shared parameters is better than decoder only LM and prefix LM.

[slide credit:Abhishek Panigrahi, Victoria Graf]

C4: The Data

- C4: Colossal Clean Crawled Corpus
    - Web-extracted text
    - English language only (langdetect)
    - 750GB
- Retain:
    - Sentences with terminal punctuation marks
    - Pages with at least 5 sentences, sentences with at least 3 words
    - Deduplicate three sentence spans
- Remove:
    - References to Javascript
    - “Lorem ipsum” text — placeholder text commonly used to demonstrate the visual form of a document



Letraset/Body Type

Lorem ipsum dolor sit amet, consectetur adipisci tempor incidunt ut labore et dolore magna aliqua veniam, quis nostrund exercitation ullamcorpor s commodo consequat. Duis autem vel eum irrure esse molestiae consequat, vel illum dolore eu fugi et iusto odio dignissim qui blandit praesent luptat exceptur sint occaecat cupiditat non provident, deserunt mollit anim id est laborum et dolor fuga distinct. Nam liber tempor cum soluta nobis elige quod maxim placeat facer possim omnis volupt

Lorem ipsum dolor si amet, consectetur adipiscing incidunt ut labore et dolore magna aliquam erat nostrud exercitation ullamcorper suscipit laboris nis duis autem vel eum irure dolor in reprehenderit i dolore eu fugiat nulla pariatur. At vero eos et accusa praesant luptatum delenit aigue duos dolor et mole provident, simil tempor sunt in culpa qui officia de fuga. Et harumd dereud facilis est er expedit disti eligend optio congue nihil impedit doming id quod assumenda est, omnis dolor repellend. Temporibud

Lorem ipsum dolor si amet, consectetur adipiscing incidunt ut labore et dolore magna aliquam erat nostrud exercitation ullamcorper suscipit laboris nis duis autem vel eum irure dolor in reprehenderit i dolore eu fugiat nulla pariatur. At vero eos et accus praesant luptatum delenit aigue duos dolor et mol provident, simil tempor sunt in culpa qui officia de fuga. Et harumd dereud facilis est er expedit disti eligend oprio congue nihil impedit doming id quod assumenda est, omnis dolor repellend. Temporibud

LETRASET PATENTED PRINTED IN ENGLAND PM169

C4: The Data

Menu

Lemon

Introduction

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
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.

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
FAQ

Dried Lemons, $3.59/pound

Organic dried lemons from our farm in California.
Lemons are harvested and sun-dried for maximum flavor.
Good in soups and on popcorn.

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
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.
Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit.
Fusce quis blandit lectus.
Mauris at mauris a turpis tristique lacinia at nec ante.
Aenean in scelerisque tellus, a efficitur ipsum.
Integer justo enim, ornare vitae sem non, mollis fermentum lectus.
Mauris ultrices nisl at libero porta sodales in ac orci.

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

Slide adapted from Colin Raffel

C4: The Data

Menu

Lemon

Introduction

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
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.

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
FAQ

Dried Lemons, $3.59/pound

Organic dried lemons from our farm in California.
Lemons are harvested and sun-dried for maximum flavor.
Good in soups and on popcorn.

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
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.
Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit.
Fusce quis blandit lectus.
Mauris at mauris a turpis tristique lacinia at nec ante.
Aenean in scelerisque tellus, a efficitur ipsum.
Integer justo enim, ornare vitae sem non, mollis fermentum lectus.
Mauris ultrices nisl at libero porta sodales in ac orci.

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

Slide adapted from Colin Raffel

C4: The Data

- 750GB? What does that mean?

Data set	Size
★ C4	745GB
C4, unfiltered	6.1TB
RealNews-like	35GB
WebText-like	17GB
Wikipedia	16GB
Wikipedia + TBC	20GB

Play with the data: `https://c4-search.apps.allenai.org/`

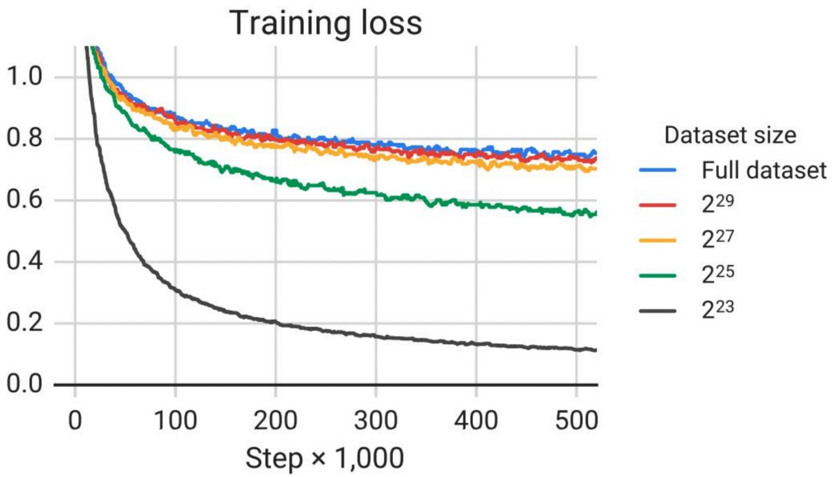
Pre-training Data: Experiment

- 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
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

Pre-training Data: Experiment

- Effect of data repetitions
- Takeaways:
  - (Table) Performance degrades as the information content shrinks.
  - (Figure) The model memorizes the pre-training data, with smaller/repeated datasets.



Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full data set	0	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
2^{29}	64	**82.87**	**19.19**	**80.97**	**72.03**	**26.83**	**39.74**	**27.63**
2^{27}	256	82.62	**19.20**	79.78	69.97	**27.02**	**39.71**	27.33
2^{25}	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81

Resulting Models: T5

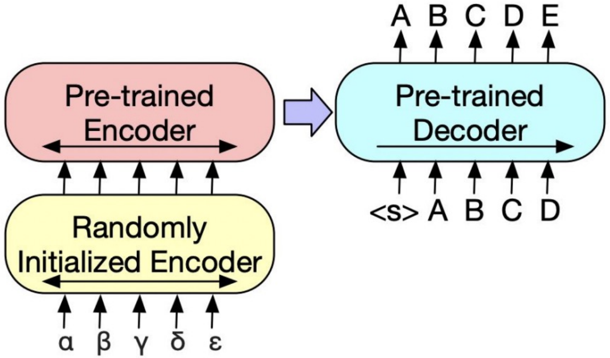
Model	Parameters	No. of layers	d_{model}	d_{ff}	d_{kv}	No. of heads
Small	60M	6	512	2048	64	8
Base	220M	12	768	3072	64	12
Large	770M	24	1024	4096	64	16
3B	3B	24	1024	16384	128	32
11B	11B	24	1024	65536	128	128

Model	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Previous best	89.4	20.30	95.5	84.6	**33.8**	**43.8**	**38.5**
T5-Small	77.4	19.56	87.24	63.3	26.7	36.0	26.8
T5-Base	82.7	20.34	92.08	76.2	30.9	41.2	28.0
T5-Large	86.4	20.68	93.79	82.3	32.0	41.5	28.1
T5-3B	88.5	21.02	94.95	86.4	31.8	42.6	28.2
T5-11B	**89.7**	**21.55**	**95.64**	**88.9**	32.1	43.4	28.1

https://huggingface.co/t5-base

BART (Lewis et al. 2020)

- Similar Architecture as T5.
  - Performs competitive to RoBERTa and XLNet on discriminative/classification tasks.
  - Outperformed existing methods on generative tasks (question answering, and summarization).
  - Improved results on machine translation with fine-tuning on target language.



BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer
Facebook AI
{mikelewis,yinhanliu,naman}@fb.com

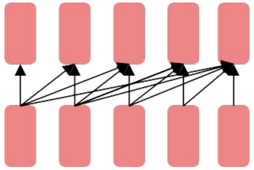
BART

```
from transformers import BartTokenizer, BartForConditionalGeneration

tokenizer = BartTokenizer.from_pretrained("facebook/bart-large")
model = BartForConditionalGeneration.from_pretrained("facebook/bart-large")

TXT = "The sun is <mask> ."
input_ids = tokenizer([TXT], return_tensors="pt")["input_ids"]
logits = model(input_ids).logits

masked_index = (input_ids[0] == tokenizer.mask_token_id).nonzero().item()
probs = logits[0, masked_index].softmax(dim=0)
values, predictions = probs.topk(5)

tokenizer.decode(predictions).split()
```

<u>Result:</u>

```
['located', 'at', 'approximately', 'also', 'about']
```

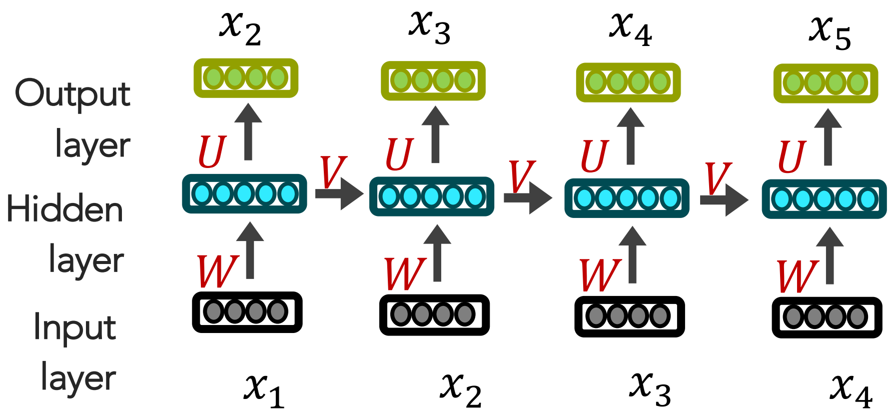
GPT



Decoders

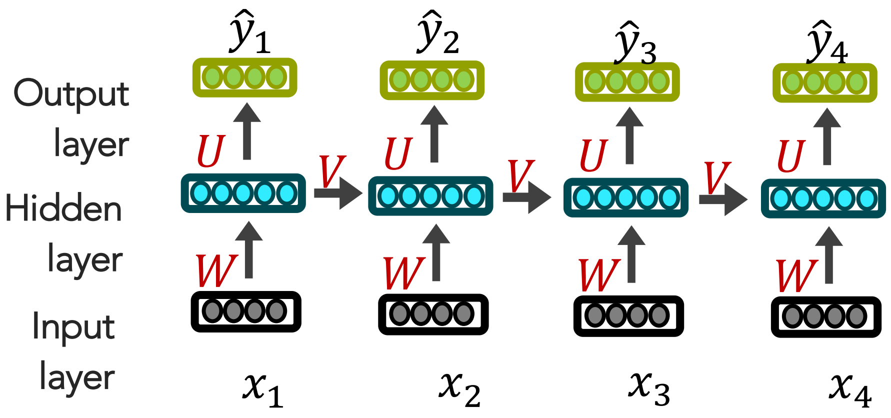
Terminology: Causal or Auto-regressive Model

Language Modelling



Auto-regressive

1-to-1 tagging/classification



Non-Auto-regressive

GPT

Generative Pre-trained Transformer

GPT-2: A Big Language Model (2019)

Language Models are Unsupervised Multitask Learners

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

GPT: An Auto-Regressive LM (2018)

Improving Language Understanding by Generative Pre-Training

Alec Radford, OpenAI, alec@openai.com
Karthik Narasimhan, OpenAI, karthikn@openai.com
Tim Salimans, OpenAI, tim@openai.com
Ilya Sutskever, OpenAI, ilyasu@openai.com

GPT-2

- GPT-2 uses only Transformer Decoders (no Encoders) to generate new sequences from scratch or from a starting sequence

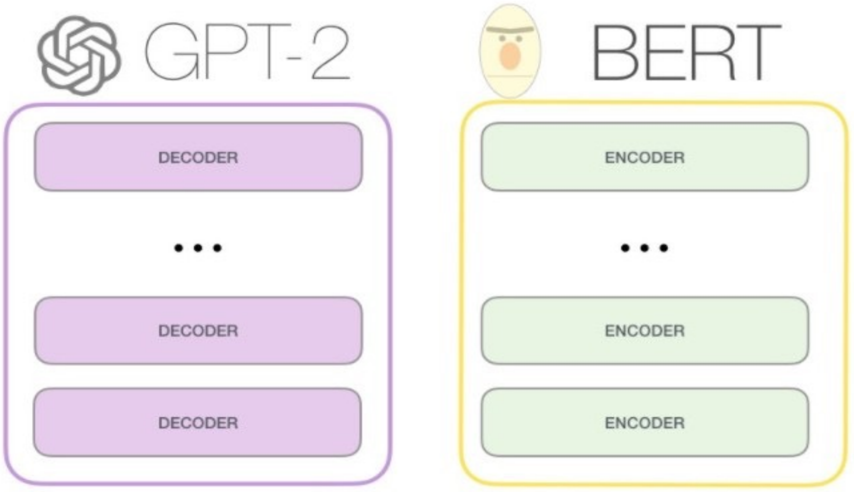


Image by http://jalammar.github.io/illustrated-gpt2/

GPTo-2: Next Word Prediction

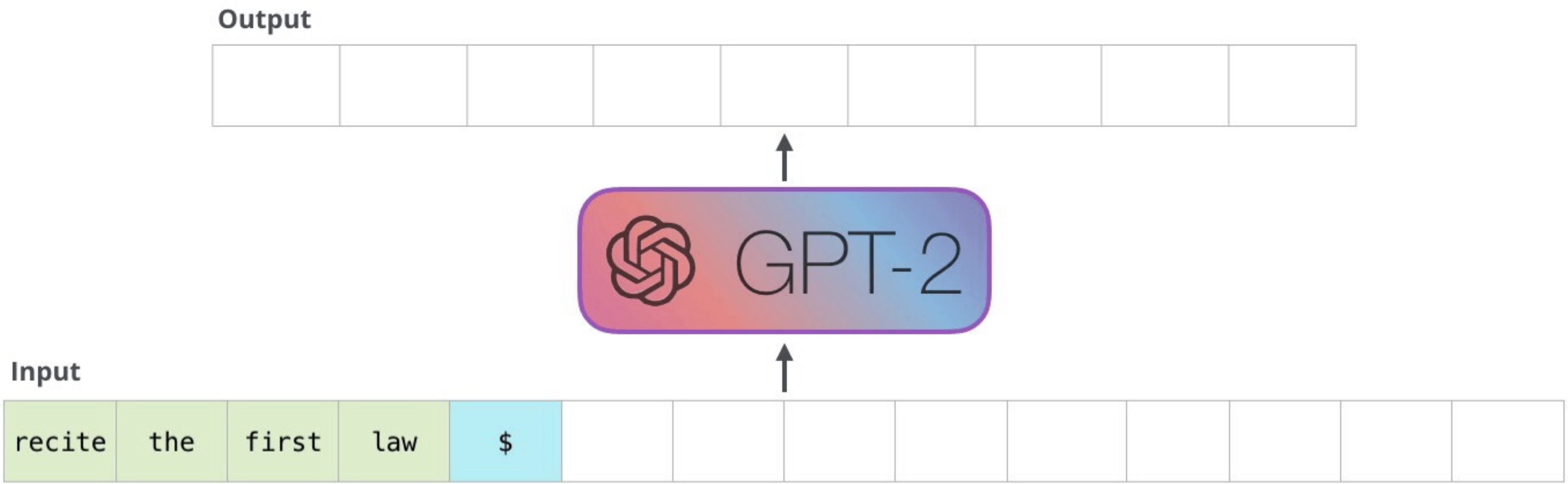


Image by http://jalammar.github.io/illustrated-gpt2/

GPT-2

- As it processes each subword, it masks the “future” words and conditions on and attends to the previous words

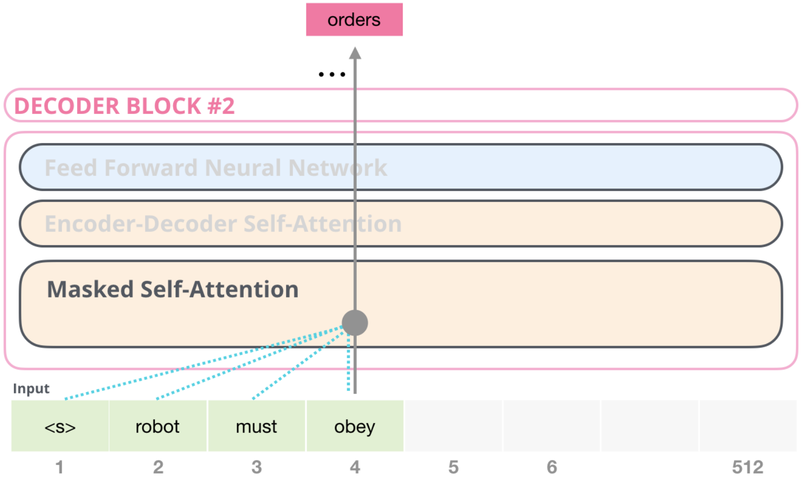


Image by http://jalammar.github.io/illustrated-gpt2/

GPT-2

- As it processes each subword, it masks the “future” words and conditions on and attends to the previous words

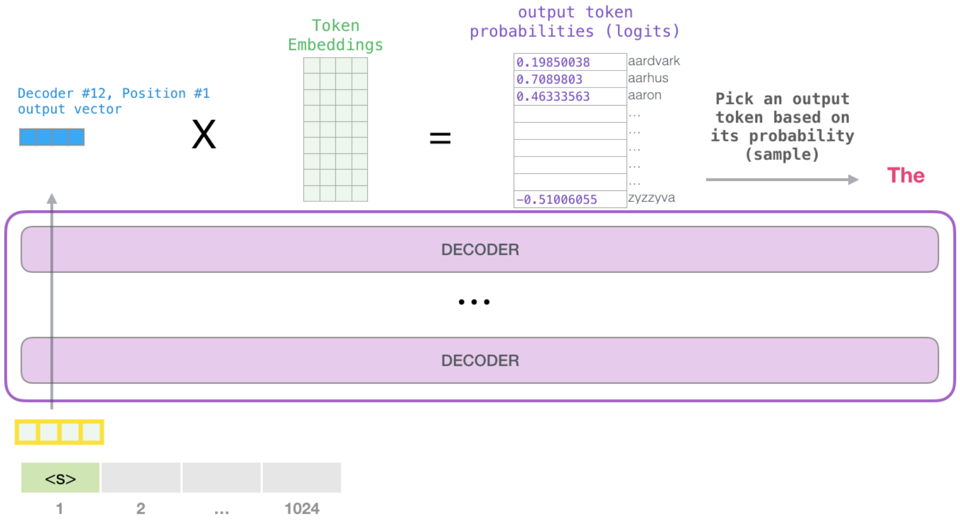
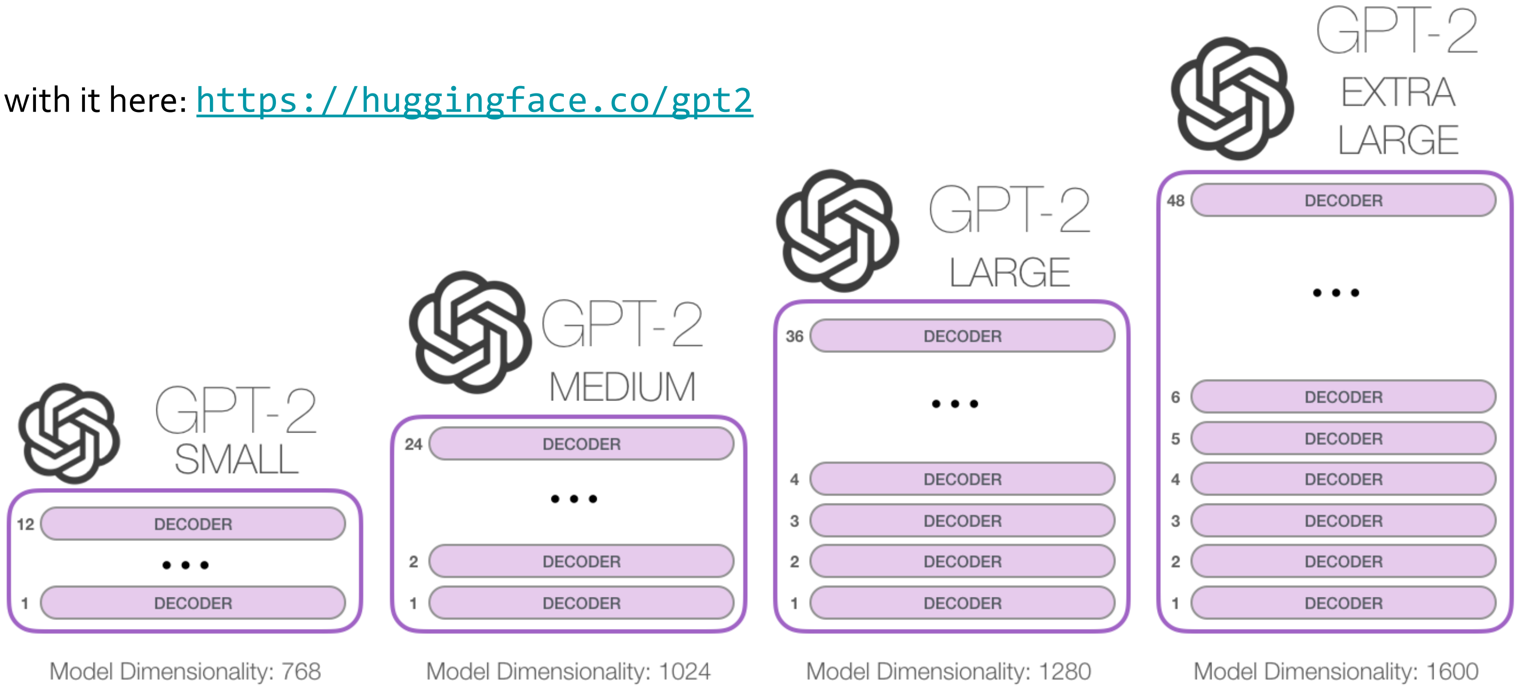


Image by http://jalammar.github.io/illustrated-gpt2/

GPT2: Model Sizes

Play with it here: https://huggingface.co/gpt2



117M parameters 345M 762M 1542M

[Image by http://jalammar.github.io/illustrated-gpt2/]

GPT2: Some Results

Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	56.25	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). Other language model results are from (Dai et al., 2019).

Article: Prehistoric man sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave in modern day France 36,000 years ago. Now, with the help of cutting-edge technology, those works of art in the Chauvet-Pont-d'Arc Cave have been reproduced to create the biggest replica cave in the world. The manmade cavern named the Caverne du Pont-d'Arc has been built a few miles from the original site in Vallon-Pont-D'arc in Southern France and contains 1,000 painstakingly-reproduced drawings as well as around 450 bones and other features... Cavemen and women sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave 36,000 years ago and now a replica has been created (pictured) ...	
GPT-2: The original site in Vallon-Pont-D'arc in Southern France is a Unesco World Heritage site and is the oldest known and the best preserved cave decorated by man. The replica cave was built a few miles from the original site in Vallon-Pont-D'Arc in Southern France. The cave contains images of 14 different species of animals including woolly rhinoceros, mammoths, and big cats.	**Reference:** Cave mimics famous Caverne du Pont-d'Arc in France, the oldest cave decorated by man and the best preserved. The replica contains all 1,000 paintings which include 425 such as a woolly rhinoceros and mammoths. Minute details were copied using 3D modelling and anamorphic techniques, often used to shoot widescreen images. The modern cave also includes replica paw prints of bears, bones and details preserved in the original cave.

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

- Has Layer normalization in between each sub-block (as we've already seen)
- 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]

GPT-3: A Very Large Language Model (2020)

- More layers & parameters
- Bigger dataset
- Longer training
- Larger embedding/hidden dimension
- Larger context window

[Slide credit: Sbhya Chhabria & Michael Tang]

Size Comparisons

- BERT-Base model has 12 transformer blocks, 12 attention heads,
  - 110M parameters!
- BERT-Large model has 24 transformer blocks, 16 attention heads,
  - 340M parameters!
- GPT-2 is trained on 40GB of text data (8M webpages)!
  - 1.5B parameters!
- GPT-3 is an even bigger version of GPT-2, but isn't open-source
  - 175B parameters!

```
Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church
has agreed to a historic split - one that is expected to end in the
creation of a new denomination, one that will be "theologically and
socially conservative," according to The Washington Post. The majority of
delegates attending the church's annual General Conference in May voted to
strengthen a ban on the ordination of LGBTQ clergy and to write new rules
that will "discipline" clergy who officiate at same-sex weddings. But
those who opposed these measures have a new plan: They say they will form a
separate denomination by 2020, calling their church the Christian Methodist
denomination.
The Post notes that the denomination, which claims 12.5 million members, was
in the early 20th century the "largest Protestant denomination in the U.S.,"
but that it has been shrinking in recent decades. The new split will be the
second in the church's history. The first occurred in 1968, when roughly
10 percent of the denomination left to form the Evangelical United Brethren
Church. The Post notes that the proposed split "comes at a critical time
for the church, which has been losing members for years," which has been
"pushed toward the brink of a schism over the role of LGBTQ people in the
church." Gay marriage is not the only issue that has divided the church. In
2016, the denomination was split over ordination of transgender clergy, with
the North Pacific regional conference voting to ban them from serving as
clergy, and the South Pacific regional conference voting to allow them.
```

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

GPT3: Try it yourself!

https://beta.openai.com/playground