Large Language Models

CSCI 601 471/671
NLP: Self-Supervised Models

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

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

- **The midterm:**
  - will be on March 7 during class time.
  - 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 $\mathbf{x}$:

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

\[
\text{Attention}(\mathbf{x}) = \text{softmax} \left( \frac{Q K^T}{\sqrt{h}} \right) V
\]

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

- An **encoder-decoder** architecture
- 3 forms of attention

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

- A building block for a variety of LMs
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.

Use the output of the masked word’s position to predict the masked word.

Possible classes: All English words

0.1% Aardvark

... Improvisation

... Zyzzyva

FFNN + Softmax

Randomly mask 15% of tokens

Input

[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
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*  
Facebook AI Research  
Seattle, WA

Omer Levy*  
Yinhan Liu*  
Luke Zettlemoyer

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

Kartik Goyal, Chris Dyer, Taylor Berg-Kirkpatrick

Leveraging Pre-trained Checkpoints for Sequence Generation Tasks

Sascha Rothe, Shashi Narayan, Aliaksei Severyn

<table>
<thead>
<tr>
<th>src</th>
<th>Der Abzug der franzischen Kampfruppen wurde am 20. November abgeschlossen.</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = 0</td>
<td>The departure of the French combat completed completed on 20 November.</td>
</tr>
<tr>
<td>t = 1</td>
<td>The departure of French combat troops was completed on 20 November.</td>
</tr>
<tr>
<td>t = 2</td>
<td>The withdrawal of French combat troops was completed on November 20th.</td>
</tr>
</tbody>
</table>
BART/T5
T5: Text-To-Text Transfer Transformer (2019)

- An encoder-decoder architecture
- But it’s more than just a model paper
- The paper conducts an in-depth analysis of various parameters of model design

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

Colin Raffel*  
Noam Shazeer*  
Adam Roberts*  
Katherine Lee*  
Sharan Narang  
Michael Matena  
Yanqi Zhou  
Wei Li  
Peter J. Liu

*Google, Mountain View, CA 94043, USA

* Corresponding author.
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.

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

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

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Objective</th>
<th>Params</th>
<th>Cost</th>
<th>GLUE</th>
<th>CNNDM</th>
<th>SQuAD</th>
<th>SGLUE</th>
<th>EnDe</th>
<th>EnFr</th>
<th>EnRo</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ Encoder-decoder</td>
<td>Denoising</td>
<td>$2P$ M</td>
<td>$M$</td>
<td>83.28</td>
<td>19.24</td>
<td>80.88</td>
<td>71.36</td>
<td>26.98</td>
<td>39.82</td>
<td>27.65</td>
</tr>
</tbody>
</table>

[slide credit: Abhishek Panigrahi, Victoria Graf]
Architectural Variants: Experiments

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Objective</th>
<th>Params</th>
<th>Cost</th>
<th>GLUE</th>
<th>CNNDM</th>
<th>SQuAD</th>
<th>SGLUE</th>
<th>EnDe</th>
<th>EnFr</th>
<th>EnRo</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ Encoder-decoder</td>
<td>Denoising</td>
<td>2P</td>
<td>M</td>
<td>83.28</td>
<td>19.24</td>
<td>80.88</td>
<td>71.36</td>
<td>26.98</td>
<td>39.82</td>
<td>27.65</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Objective</th>
<th>Params</th>
<th>Cost</th>
<th>GLUE</th>
<th>CNNDM</th>
<th>SQuAD</th>
<th>SGLUE</th>
<th>EnDe</th>
<th>EnFr</th>
<th>EnRo</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ Encoder-decoder</td>
<td>Denoising</td>
<td>2P</td>
<td>M</td>
<td>83.28</td>
<td>19.24</td>
<td>80.88</td>
<td>71.36</td>
<td>26.98</td>
<td>39.82</td>
<td>27.65</td>
</tr>
</tbody>
</table>

[slide credit: Abhishek Panigrahi, Victoria Graf]
## Architectural Variants: Experiments

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Objective</th>
<th>Params</th>
<th>Cost</th>
<th>GLUE</th>
<th>CNNDM</th>
<th>SQuAD</th>
<th>SGLUE</th>
<th>EnDe</th>
<th>EnFr</th>
<th>EnRo</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ Encoder-decoder</td>
<td>Denoising</td>
<td>$2P$</td>
<td>$M$</td>
<td>83.28</td>
<td>19.24</td>
<td>80.88</td>
<td>71.36</td>
<td>26.98</td>
<td>39.82</td>
<td>27.65</td>
</tr>
</tbody>
</table>

[slide credit: Abhishek Panigrahi, Victoria Graf]
## Architectural Variants: Experiments

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Objective</th>
<th>Params</th>
<th>Cost</th>
<th>GLUE</th>
<th>CNNDM</th>
<th>SQuAD</th>
<th>SGLUE</th>
<th>EnDe</th>
<th>EnFr</th>
<th>EnRo</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ Encoder-decoder</td>
<td>Denoising</td>
<td>$2P$</td>
<td>$M$</td>
<td>83.28</td>
<td>19.24</td>
<td>80.88</td>
<td>71.36</td>
<td>26.98</td>
<td>39.82</td>
<td>27.65</td>
</tr>
<tr>
<td>Enc-dec, shared</td>
<td>Denoising</td>
<td>$P$</td>
<td>$M$</td>
<td>82.81</td>
<td>18.78</td>
<td>80.63</td>
<td>70.73</td>
<td>26.72</td>
<td>39.03</td>
<td>27.46</td>
</tr>
</tbody>
</table>

[slide credit: Abhishek Panigrahi, Victoria Graf]
# Architectural Variants: Experiments

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

[slide credit: Abhishek Panigrahi, Victoria Graf]
## Architectural Variants: Experiments

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

[slide credit: Abhishek Panigrahi, Victoria Graf]
## Architectural Variants: Experiments

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

Language model is decoder-only

---

[slide credit: Abhishek Panigrahi, Victoria Graf]
**Architectural Variants: Experiments**

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

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

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

Prefix LM

[slide credit: Abhishek Panigrahi, Victoria Graf]
Architectural Variants: Experiments

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

- **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.
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
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.
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.

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.


```
function Ball(r) {
  this.radius = r;
  this.area = pi * r ** 2;
  this.show = function(){
    drawCircle(r);
  }
}
```
C₄: The Data

- 750GB? What does that mean?

<table>
<thead>
<tr>
<th>Data set</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ C₄</td>
<td>745GB</td>
</tr>
<tr>
<td>C₄, unfiltered</td>
<td>6.1TB</td>
</tr>
<tr>
<td>RealNews-like</td>
<td>35GB</td>
</tr>
<tr>
<td>WebText-like</td>
<td>17GB</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>16GB</td>
</tr>
<tr>
<td>Wikipedia + TBC</td>
<td>20GB</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Size</th>
<th>GLUE</th>
<th>CNNDM</th>
<th>SQuAD</th>
<th>SGLUE</th>
<th>EnDe</th>
<th>EnFr</th>
<th>EnRo</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ C4</td>
<td>745GB</td>
<td>83.28</td>
<td>19.24</td>
<td>80.88</td>
<td>71.36</td>
<td>26.98</td>
<td>39.82</td>
<td>27.65</td>
</tr>
<tr>
<td>C4, unfiltered</td>
<td>6.1TB</td>
<td>81.46</td>
<td>19.14</td>
<td>78.78</td>
<td>68.04</td>
<td>26.55</td>
<td>39.34</td>
<td>27.21</td>
</tr>
<tr>
<td>RealNews-like</td>
<td>35GB</td>
<td>83.83</td>
<td>19.23</td>
<td>80.39</td>
<td>72.38</td>
<td>26.75</td>
<td>39.90</td>
<td>27.48</td>
</tr>
<tr>
<td>WebText-like</td>
<td>17GB</td>
<td>84.03</td>
<td>19.31</td>
<td>81.42</td>
<td>71.40</td>
<td>26.80</td>
<td>39.74</td>
<td>27.59</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>16GB</td>
<td>81.85</td>
<td>19.31</td>
<td>81.29</td>
<td>68.01</td>
<td>26.94</td>
<td>39.69</td>
<td>27.67</td>
</tr>
<tr>
<td>Wikipedia + TBC</td>
<td>20GB</td>
<td>83.65</td>
<td>19.28</td>
<td>82.08</td>
<td>73.24</td>
<td>26.77</td>
<td>39.63</td>
<td>27.57</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Number of tokens</th>
<th>Repeats</th>
<th>GLUE</th>
<th>CNNDM</th>
<th>SQuAD</th>
<th>SGLUE</th>
<th>EnDe</th>
<th>EnFr</th>
<th>EnRo</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ Full data set</td>
<td>0</td>
<td>83.28</td>
<td>19.24</td>
<td>80.88</td>
<td>71.36</td>
<td>26.98</td>
<td>39.82</td>
<td>27.65</td>
</tr>
<tr>
<td>$2^{29}$</td>
<td>64</td>
<td>82.87</td>
<td>19.19</td>
<td>80.97</td>
<td>72.03</td>
<td>26.83</td>
<td>39.74</td>
<td>27.63</td>
</tr>
<tr>
<td>$2^{27}$</td>
<td>256</td>
<td>82.62</td>
<td>19.20</td>
<td>79.78</td>
<td>69.97</td>
<td>27.02</td>
<td>39.71</td>
<td>27.33</td>
</tr>
<tr>
<td>$2^{25}$</td>
<td>1,024</td>
<td>79.55</td>
<td>18.57</td>
<td>76.27</td>
<td>64.76</td>
<td>26.38</td>
<td>39.56</td>
<td>26.80</td>
</tr>
<tr>
<td>$2^{23}$</td>
<td>4,096</td>
<td>76.34</td>
<td>18.33</td>
<td>70.92</td>
<td>59.29</td>
<td>26.37</td>
<td>38.84</td>
<td>25.81</td>
</tr>
</tbody>
</table>
## Resulting Models: T5

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>No. of layers</th>
<th>$d_{model}$</th>
<th>$d_{ff}$</th>
<th>$d_{kv}$</th>
<th>No. of heads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>60M</td>
<td>6</td>
<td>512</td>
<td>2048</td>
<td>64</td>
<td>8</td>
</tr>
<tr>
<td>Base</td>
<td>220M</td>
<td>12</td>
<td>768</td>
<td>3072</td>
<td>64</td>
<td>12</td>
</tr>
<tr>
<td>Large</td>
<td>770M</td>
<td>24</td>
<td>1024</td>
<td>4096</td>
<td>64</td>
<td>16</td>
</tr>
<tr>
<td>3B</td>
<td>3B</td>
<td>24</td>
<td>1024</td>
<td>16384</td>
<td>128</td>
<td>32</td>
</tr>
<tr>
<td>11B</td>
<td>11B</td>
<td>24</td>
<td>1024</td>
<td>65536</td>
<td>128</td>
<td>128</td>
</tr>
</tbody>
</table>

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

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.
```python
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()
```

**Result:** `['located', 'at', 'approximately', 'also', 'about']`
GPT Decoders
Terminology: Causal or Auto-regressive Model

Language Modelling

Auto-regressive

Non-Auto-regressive

1-to-1 tagging/classification

Output layer
Hidden layer
Input layer

Output layer
Hidden layer
Input layer

\[ x_1 \]
\[ W \]
\[ U \]
\[ x_2 \]
\[ x_3 \]
\[ x_4 \]
\[ x_5 \]

\[ \hat{y}_1 \]
\[ W \]
\[ U \]
\[ \hat{y}_2 \]
\[ \hat{y}_3 \]
\[ \hat{y}_4 \]

\[ V \]
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 ² ³  Ilya Sutskever ² ³


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  ilyas@openai.com
GPT-2

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

GPTo-2: Next Word Prediction

Input
recite the first law $ 

Output

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.

GPT2: Model Sizes

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

GPT-2 SMALL
Model Dimensionality: 768
117M parameters

GPT-2 MEDIUM
Model Dimensionality: 1024
345M

GPT-2 LARGE
Model Dimensionality: 1280
762M

GPT-2 EXTRA LARGE
Model Dimensionality: 1600
1542M

[Image by http://jalammar.github.io/illustrated-gpt2/]
### GPT2: Some Results

#### Language Models are Unsupervised Multitask Learners

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

*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)  

<table>
<thead>
<tr>
<th><strong>GPT-2:</strong></th>
<th><strong>Reference:</strong> 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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
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