What can MIR learn from transfer learning in NLP?

Colin Raffel

NLP4MusA (First Workshop on NLP for Music and Audio)
What is transfer learning and what makes it so useful?

How should we do transfer learning in NLP?

How can we apply these ideas to MIR?
The cabs ___ the same rates as those ___ by horse-drawn cabs and were ___ quite popular, ___ the Prince of Wales (the ___ King Edward VII) travelled in ___. The cabs quickly ___ known as "hummingbirds" for ___ noise made by their motors and their distinctive black and ___ livery. Passengers ___ ___ the interior fittings were ___ when compared to ___ cabs but there ___ some complaints ___ the ___ lighting made them too ___ to those outside ___.

charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were that, internal, conspicuous, cab

This movie is terrible! The acting is bad and I was bored the entire time. There was no plot and nothing interesting happened. I was really surprised since I had very high expectations. I want 103 minutes of my life back!

negative
SQuAD Exact Match score (validation set)

Source: https://paperswithcode.com/sota/question-answering-on-squad11-dev
Transfer learning

No transfer learning

Source: https://paperswithcode.com/sota/question-answering-on-squad11-dev
Lots of stuff!
Text-to-Text Transfer Transformer

T5
translate English to German: That is good.

Das ist gut.
"cola sentence: The course is jumping well."
T5
"not acceptable"
"stsbt sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi...

"six people hospitalized after a storm in attala county."
translate English to German: That is good.

cola sentence: The course is jumping well.

summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi...

stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field.

six people hospitalized after a storm in attala county.

Das ist gut.

not acceptable

3.8
running man was one of the many ''lemon'' on a beach in the capital and largest city of the united kingdom, the county seat of oklahoma city. oklahoma city ranks 27th among united states cities in population, with the population estimated to have increased to 643,648 as of july 2017.[5] oklahoma city metropolitan area went from 1,376,349 residents in 2010 to 1,385,452,[9] and the county, with the population estimated to have increased to 643,648 as of july 2017, has a density of 3,169.5 people per square mile (1,221.7/km²), which is higher than the state average of 1,549,758 residents.[9] the oklahoma city metropolitan area is the largest metropolitan area in oklahoma, and the city limits extend into canada county.[8]

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...
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.

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.

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.

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

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.


function Ball(r) {
  this.radius = r;
  this.area = pi * r ** 2;
  this.show = function(){
    drawCircle(r);
  }
}
A colossal, cleaned version of Common Crawl's web crawl corpus.
Thank you for inviting me to your party last week.
Thank you for inviting me to your party last week.
Thank you for inviting me to your party last week.
Thank you for inviting me to your party last week.

Inputs:
Thank you <X> me to your party <Y> week.

Targets:
<X> for inviting <Y> last <Z>
Pretrain

- $\text{BERT}_{\text{BASE}}$-sized encoder-decoder Transformer
- Denoising objective
- C4 dataset

- $2^{19}$ steps
- $2^{35}$ or ~34B tokens
- Inverse square root learning rate schedule
Pretrain

- BERT\textsubscript{BASE}\textsubscript{-sized} encoder-decoder Transformer
- Denoising objective
- C4 dataset

2\textsuperscript{19} steps
2\textsuperscript{35} or ~34B tokens
Inverse square root learning rate schedule

Finetune

GLUE

2\textsuperscript{18} steps
2\textsuperscript{34} or ~17B tokens
Constant learning rate
Pretrain

- BERT\textsubscript{BASE}-sized encoder-decoder Transformer
- Denoising objective
- C4 dataset

Finetune

- GLUE
- CNN/DM

2\textsuperscript{19} steps
2\textsuperscript{35} or \sim 34B tokens
Inverse square root learning rate schedule

2\textsuperscript{18} steps
2\textsuperscript{34} or \sim 17B tokens
Constant learning rate
Pretrain

BERT\textsubscript{BASE}-sized encoder-decoder Transformer

Denoising objective

C4 dataset

Finetune

GLUE

CNN/DM

SQuAD

$2^{19}$ steps

$2^{35}$ or ~34B tokens

Inverse square root learning rate schedule

$2^{18}$ steps

$2^{34}$ or ~17B tokens

Constant learning rate
Pretrain

- BERT\textsubscript{BASE}-sized encoder-decoder Transformer
- Denoising objective
- C4 dataset

Finetune

- GLUE
- CNN/DM
- SQuAD
- SuperGLUE

2\textsuperscript{19} steps
2\textsuperscript{35} or ~34B tokens
Inverse square root learning rate schedule

2\textsuperscript{18} steps
2\textsuperscript{34} or ~17B tokens
Constant learning rate
Pretrain

- BERT\textsubscript{BASE}-sized encoder-decoder Transformer
- Denoising objective
- C4 dataset

2^{19} steps
2^{35} or ~34B tokens
Inverse square root learning rate schedule

Finetune

- GLUE
- CNN/DM
- SQuAD
- SuperGLUE
- WMT14 EnDe
- WMT15 EnFr
- WMT16 EnRo

2^{18} steps
2^{34} or ~17B tokens
Constant learning rate
Pretrain

- BERT\textsubscript{BASE}\textsuperscript{-sized} encoder-decoder Transformer
- Denoising objective
- C4 dataset

2^{19} steps
2^{35} or ~34B tokens
Inverse square root learning rate schedule

Finetune

- GLUE
- CNN/DM
- SQuAD
- SuperGLUE
- WMT14 EnDe
- WMT15 EnFr
- WMT16 EnRo

Evaluate on validation

- step 750000
- step 760000
- step 770000
- step 780000

Evaluate all checkpoints, choose the best

\[2^{18}\] steps
2^{34} or ~17B tokens
Constant learning rate
<table>
<thead>
<tr>
<th>Setting 1</th>
<th>Setting 2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLUE</td>
<td>CNNDM</td>
<td>SQuAD</td>
</tr>
</tbody>
</table>

*Downstream task performance*
<table>
<thead>
<tr>
<th></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>Baseline average</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>Baseline standard deviation</td>
<td>0.235</td>
<td>0.065</td>
<td>0.343</td>
<td>0.416</td>
<td>0.112</td>
<td>0.090</td>
<td>0.108</td>
</tr>
<tr>
<td>No pre-training</td>
<td>66.22</td>
<td>17.60</td>
<td>50.31</td>
<td>53.04</td>
<td>25.86</td>
<td>39.77</td>
<td>24.04</td>
</tr>
<tr>
<td>Star denotes baseline</td>
<td>GLUE</td>
<td>CNNDM</td>
<td>SQuAD</td>
<td>SGLUE</td>
<td>EnDe</td>
<td>EnFr</td>
<td>EnRo</td>
</tr>
<tr>
<td>----------------------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Baseline average</td>
<td><strong>83.28</strong></td>
<td><strong>19.24</strong></td>
<td><strong>80.88</strong></td>
<td><strong>71.36</strong></td>
<td><strong>26.98</strong></td>
<td><strong>39.82</strong></td>
<td><strong>27.65</strong></td>
</tr>
<tr>
<td>Baseline standard deviation</td>
<td>0.235</td>
<td>0.065</td>
<td>0.343</td>
<td>0.416</td>
<td>0.112</td>
<td>0.090</td>
<td>0.108</td>
</tr>
<tr>
<td>No pre-training</td>
<td>66.22</td>
<td>17.60</td>
<td>50.31</td>
<td>53.04</td>
<td>25.86</td>
<td><strong>39.77</strong></td>
<td>24.04</td>
</tr>
</tbody>
</table>

Bold = 1 std. dev. of max

Big training set
<table>
<thead>
<tr>
<th>Architecture</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>$2P$</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>$P$</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>$P$</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>$P$</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>$P$</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>

- **Encoder-decoder**
  - $2P$: 2 parameters
  - $M$: Medium cost
- **Enc-dec, shared**
  - $P$: 1 parameter
  - $M$: Medium cost
- **Enc-dec, 6 layers**
  - $P$: 1 parameter
  - $M/2$: Medium/2 cost
- **Language model**
  - $P$: 1 parameter
  - $M$: Medium cost
- **Prefix LM**
  - $P$: 1 parameter
  - $M$: Medium cost

![Diagram](image-url)
High-level approaches
- Language modeling
- BERT-style
- Deshuffling

Corruption strategies
- Mask
- Replace spans
- Drop

Corruption rate
- 10%
- 15%
- 25%
- 50%

Corrupted span length
- 2
- 3
- 5
- 10

<table>
<thead>
<tr>
<th>Span length</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>★ Baseline (i.i.d.)</td>
<td><strong>83.28</strong></td>
<td>19.24</td>
<td>80.88</td>
<td><strong>71.36</strong></td>
<td><strong>26.98</strong></td>
<td>39.82</td>
<td>27.65</td>
</tr>
<tr>
<td>2</td>
<td>83.54</td>
<td>19.39</td>
<td>82.09</td>
<td><strong>72.20</strong></td>
<td>26.76</td>
<td>39.99</td>
<td>27.63</td>
</tr>
<tr>
<td>3</td>
<td>83.49</td>
<td>19.62</td>
<td><strong>81.84</strong></td>
<td>72.53</td>
<td>26.86</td>
<td>39.65</td>
<td>27.62</td>
</tr>
<tr>
<td>5</td>
<td>83.40</td>
<td>19.24</td>
<td>82.05</td>
<td><strong>72.23</strong></td>
<td>26.88</td>
<td>39.40</td>
<td>27.53</td>
</tr>
<tr>
<td>10</td>
<td>82.85</td>
<td>19.33</td>
<td>81.84</td>
<td>70.44</td>
<td>26.79</td>
<td>39.49</td>
<td>27.69</td>
</tr>
<tr>
<td>Dataset</td>
<td>Size</td>
<td>GLUE</td>
<td>CNNDM</td>
<td>SQuAD</td>
<td>SGLUE</td>
<td>EnDe</td>
<td>EnFr</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<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>
</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>
</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>
</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>
</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>
</tr>
<tr>
<td>Wikipedia + TEC</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>
</tr>
</tbody>
</table>

- Much better on ReCoRD
- Much worse on CoLA
- Order of magnitude smaller
- Much better on MultiRC
<table>
<thead>
<tr>
<th>Number of tokens</th>
<th>Repeats</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>Full dataset</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>

**Training loss**

- **Dataset size**
  - Full dataset
  - $2^{29}$
  - $2^{27}$
  - $2^{25}$
  - $2^{23}$
<table>
<thead>
<tr>
<th>Mixing strategy</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>Baseline (pre-train/fine-tune)</td>
<td><strong>83.28</strong></td>
<td>19.24</td>
<td><strong>80.88</strong></td>
<td>71.36</td>
<td><strong>26.98</strong></td>
<td><strong>39.82</strong></td>
<td><strong>27.65</strong></td>
</tr>
<tr>
<td>Equal</td>
<td>76.13</td>
<td>19.02</td>
<td>76.51</td>
<td>63.37</td>
<td>23.89</td>
<td>34.31</td>
<td>26.78</td>
</tr>
<tr>
<td>Examples-proportional, $K = 2^{16}$</td>
<td>80.45</td>
<td>19.04</td>
<td>77.25</td>
<td>69.95</td>
<td>24.35</td>
<td>34.99</td>
<td>27.10</td>
</tr>
<tr>
<td>Examples-proportional, $K = 2^{17}$</td>
<td>81.56</td>
<td>19.12</td>
<td>77.00</td>
<td>67.91</td>
<td>24.36</td>
<td>35.00</td>
<td>27.25</td>
</tr>
<tr>
<td>Examples-proportional, $K = 2^{18}$</td>
<td>81.67</td>
<td>19.07</td>
<td>78.17</td>
<td>67.94</td>
<td>24.57</td>
<td>35.19</td>
<td>27.39</td>
</tr>
<tr>
<td>Examples-proportional, $K = 2^{19}$</td>
<td>81.42</td>
<td><strong>19.24</strong></td>
<td>79.78</td>
<td>67.30</td>
<td>25.21</td>
<td>36.30</td>
<td><strong>27.76</strong></td>
</tr>
<tr>
<td>Examples-proportional, $K = 2^{20}$</td>
<td>80.80</td>
<td><strong>19.24</strong></td>
<td><strong>80.36</strong></td>
<td>67.38</td>
<td>25.66</td>
<td>36.93</td>
<td><strong>27.68</strong></td>
</tr>
<tr>
<td>Examples-proportional, $K = 2^{21}$</td>
<td>79.83</td>
<td>18.79</td>
<td>79.50</td>
<td>65.10</td>
<td>25.82</td>
<td>37.22</td>
<td>27.13</td>
</tr>
<tr>
<td>Temperature-scaled, $T = 2$</td>
<td>81.90</td>
<td><strong>19.28</strong></td>
<td>79.42</td>
<td>69.92</td>
<td>25.42</td>
<td>36.72</td>
<td>27.20</td>
</tr>
<tr>
<td>Temperature-scaled, $T = 4$</td>
<td>80.56</td>
<td><strong>19.22</strong></td>
<td>77.99</td>
<td>69.54</td>
<td>25.04</td>
<td>35.82</td>
<td>27.45</td>
</tr>
<tr>
<td>Temperature-scaled, $T = 8$</td>
<td>77.21</td>
<td>19.10</td>
<td>77.14</td>
<td>66.07</td>
<td>24.55</td>
<td>35.35</td>
<td>27.17</td>
</tr>
<tr>
<td>Training strategy</td>
<td>GLUE</td>
<td>CNNDM</td>
<td>SQuAD</td>
<td>SGLUE</td>
<td>EnDe</td>
<td>EnFr</td>
<td>EnRo</td>
</tr>
<tr>
<td>--------------------------------------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>★ Unsupervised pre-training + fine-tuning</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>Multi-task training</td>
<td>81.42</td>
<td>19.24</td>
<td>79.78</td>
<td>67.30</td>
<td>25.21</td>
<td>36.30</td>
<td>27.76</td>
</tr>
<tr>
<td>Multi-task pre-training + fine-tuning</td>
<td>83.11</td>
<td>19.12</td>
<td>80.26</td>
<td>71.03</td>
<td>27.08</td>
<td>39.80</td>
<td>28.07</td>
</tr>
<tr>
<td>Leave-one-out multi-task training</td>
<td>81.98</td>
<td>19.05</td>
<td>79.97</td>
<td>71.68</td>
<td>26.93</td>
<td>39.79</td>
<td>27.87</td>
</tr>
<tr>
<td>Supervised multi-task pre-training</td>
<td>79.93</td>
<td>18.96</td>
<td>77.38</td>
<td>65.36</td>
<td>26.81</td>
<td>40.13</td>
<td>28.04</td>
</tr>
</tbody>
</table>

- Unsupervised Task
- Task A
- Task B
- Task C

- Unsupervised Task
- Task A
- Task B
- Task C

- Unsupervised Task
- Task A
- Task B
- Task C

- Unsupervised Task
- Task A
- Task B
- Task C

- Unsupervised Task
- Task A
- Task B
- Task C
<table>
<thead>
<tr>
<th>Scaling strategy</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>★ Baseline</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>1× size, 4× training steps</td>
<td>85.33</td>
<td>19.33</td>
<td>82.45</td>
<td>74.72</td>
<td>27.08</td>
<td>40.66</td>
<td>27.93</td>
</tr>
<tr>
<td>1× size, 4× batch size</td>
<td>84.60</td>
<td>19.42</td>
<td>82.52</td>
<td>74.64</td>
<td>27.07</td>
<td>40.60</td>
<td>27.84</td>
</tr>
<tr>
<td>2× size, 2× training steps</td>
<td><strong>86.18</strong></td>
<td>19.66</td>
<td><strong>84.18</strong></td>
<td>77.18</td>
<td>27.52</td>
<td><strong>41.03</strong></td>
<td>28.19</td>
</tr>
<tr>
<td>4× size, 1× training steps</td>
<td><strong>85.91</strong></td>
<td>19.73</td>
<td><strong>83.86</strong></td>
<td><strong>78.04</strong></td>
<td>27.47</td>
<td>40.71</td>
<td>28.10</td>
</tr>
<tr>
<td>4× ensembled</td>
<td>84.77</td>
<td><strong>20.10</strong></td>
<td>83.09</td>
<td>71.74</td>
<td><strong>28.05</strong></td>
<td>40.53</td>
<td><strong>28.57</strong></td>
</tr>
<tr>
<td>4× ensembled, fine-tune only</td>
<td>84.05</td>
<td>19.57</td>
<td>82.36</td>
<td>71.55</td>
<td>27.55</td>
<td>40.22</td>
<td>28.09</td>
</tr>
</tbody>
</table>
Encoder-decoder architecture

Span prediction objective

C4 dataset

Multi-task pre-training

Bigger models trained longer

<table>
<thead>
<tr>
<th>Architecture</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>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>27.05</td>
</tr>
<tr>
<td>Enc-dec, shared</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>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>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>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>

<table>
<thead>
<tr>
<th>Span length</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>Baseline (i.i.d.)</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>27.05</td>
</tr>
<tr>
<td>2</td>
<td>83.54</td>
<td>19.39</td>
<td>82.09</td>
<td>72.20</td>
<td>26.76</td>
<td>39.99</td>
<td>27.63</td>
</tr>
<tr>
<td>3</td>
<td>83.49</td>
<td>19.62</td>
<td>81.84</td>
<td>72.53</td>
<td>26.86</td>
<td>39.65</td>
<td>27.62</td>
</tr>
<tr>
<td>5</td>
<td>83.40</td>
<td>19.24</td>
<td>82.05</td>
<td>72.23</td>
<td>26.88</td>
<td>39.40</td>
<td>27.53</td>
</tr>
<tr>
<td>10</td>
<td>82.85</td>
<td>19.33</td>
<td>81.84</td>
<td>70.44</td>
<td>26.79</td>
<td>39.49</td>
<td>27.69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</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>27.05</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>27.07</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>

<table>
<thead>
<tr>
<th>Training strategy</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>★ Unsupervised pre-training + fine-tuning</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>27.05</td>
</tr>
<tr>
<td>Multi-task training</td>
<td>81.42</td>
<td>19.24</td>
<td>79.78</td>
<td>67.30</td>
<td>26.23</td>
<td>36.30</td>
<td>27.76</td>
</tr>
<tr>
<td>Multi-task pre-training + fine-tuning</td>
<td>83.11</td>
<td>19.12</td>
<td>80.26</td>
<td>71.03</td>
<td>27.08</td>
<td>39.80</td>
<td>28.07</td>
</tr>
<tr>
<td>Leave-one-out multi-task training</td>
<td>81.08</td>
<td>19.05</td>
<td>79.97</td>
<td>71.68</td>
<td>26.93</td>
<td>39.79</td>
<td>27.87</td>
</tr>
<tr>
<td>Supervised multi-task pre-training</td>
<td>79.93</td>
<td>18.96</td>
<td>77.38</td>
<td>65.36</td>
<td>26.81</td>
<td>40.13</td>
<td>28.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scaling strategy</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>Baseline</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>27.05</td>
</tr>
<tr>
<td>1× size, 4× training steps</td>
<td>85.33</td>
<td>19.33</td>
<td>82.45</td>
<td>74.72</td>
<td>27.08</td>
<td>40.66</td>
<td>27.93</td>
</tr>
<tr>
<td>1× size, 4× batch size</td>
<td>84.60</td>
<td>19.42</td>
<td>82.52</td>
<td>74.64</td>
<td>27.07</td>
<td>40.60</td>
<td>27.84</td>
</tr>
<tr>
<td>2× size, 2× training steps</td>
<td>86.18</td>
<td>19.66</td>
<td>84.18</td>
<td>77.18</td>
<td>27.52</td>
<td>41.03</td>
<td>28.19</td>
</tr>
<tr>
<td>4× size, 1× training steps</td>
<td>85.91</td>
<td>19.73</td>
<td>83.86</td>
<td>78.04</td>
<td>27.47</td>
<td>40.71</td>
<td>28.10</td>
</tr>
<tr>
<td>4× ensembled</td>
<td>84.77</td>
<td>20.10</td>
<td>83.09</td>
<td>71.74</td>
<td>28.05</td>
<td>40.53</td>
<td>28.57</td>
</tr>
<tr>
<td>4× ensembled, fine-tune only</td>
<td>84.05</td>
<td>19.57</td>
<td>82.36</td>
<td>71.55</td>
<td>27.55</td>
<td>40.22</td>
<td>28.09</td>
</tr>
</tbody>
</table>
## Model size variants

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th># layers</th>
<th>(d_{\text{model}})</th>
<th>(d_{\text{ff}})</th>
<th>(d_{\text{kv}})</th>
<th># 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>
<tr>
<td>Model</td>
<td>GLUE Average</td>
<td>CNN/DM ROUGE-2-F</td>
<td>SQuAD EM</td>
<td>SuperGLUE Average</td>
<td>WMT EnDe BLEU</td>
<td>WMT EnFr BLEU</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>------------------</td>
<td>----------</td>
<td>-------------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Previous best</td>
<td>89.4</td>
<td>20.30</td>
<td>90.1</td>
<td>84.6</td>
<td>33.8</td>
<td>43.8</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>
</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>
</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>
</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>
</tr>
<tr>
<td>T5-11B</td>
<td><strong>90.3</strong></td>
<td><strong>21.55</strong></td>
<td><strong>91.26</strong></td>
<td><strong>89.3</strong></td>
<td>32.1</td>
<td>43.4</td>
</tr>
</tbody>
</table>

Back-translation beats English-only pre-training

Human score = 89.8
Code for the paper "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"

Released Model Checkpoints

We have released the following checkpoints for pre-trained models described in our paper:

- **T5-Small** (60 million parameters): gs://t5-data/pretrained_models/small
- **T5-Base** (220 million parameters): gs://t5-data/pretrained_models/base
- **T5-Large** (770 million parameters): gs://t5-data/pretrained_models/large
- **T5-3B** (3 billion parameters): gs://t5-data/pretrained_models/3B
- **T5-11B** (11 billion parameters): gs://t5-data/pretrained_models/11B

https://github.com/google-research/text-to-text-transfer-transformer
Copyright 2019 The T5 Authors

Licensed under the Apache License, Version 2.0 (the "License");

1 cell hidden

Fine-Tuning the Text-To-Text Transfer Transformer (T5) for Context-Free Trivia

Or: What does T5 know?

The following tutorial guides you through the process of fine-tuning a pre-trained T5 model, evaluating its accuracy, and using it for prediction, all on a free Google Cloud TPU.

http://tiny.cc/t5-colab
had a population of 1,358,452, and the state of Oklahoma. The capital and largest city of the United States, Oklahoma City ranks 27th among United States cities. As of July 2017, the city limits extend into Canadian, C addition, Great Britain agreed to protect Roman Catholicism in the New World... treaty of Paris, also known as the Treaty of Paris (1763), was signed on 10 February 1763 by the Kingdoms of Great Britain, France and Spain, with Portugal in agreement, after Great Britain's victory over France and Spain during the Seven Years' War, also known as the French and Indian War in America, and has gained online popularity among K-pop fans, having been fangirled into various languages, such as English, Spanish, Portuguese, French, Italian, Thai, Vietnamese, Chinese, etc. The show has become popular in other parts of the world, primarily for its juice, which has both culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The pulp and rind of the lemon, Citrus limon (L.) Osbeck, is a species of Rutaceae, native to South Asia, primarily North India. Its distinctive sour taste of lemon juice makes it a popular fruit, as such it is a second-class lever...
MIDI Data

Instrument 1: Piano

Instrument 2: Cello
1983 MIDI specification created
1985 America Online Founded
1988 MIDI file format defined
1991 General MIDI mapping defined
1993 MP3 format released
1994 USB specification created
1999 Initial release of Napster
“Prediction: The cost for 128 kilobytes of memory will fall below U$100 in the near future.”

Creative Computing magazine
December 1981, page 6
Lakh MIDI Dataset

Slide from Chris Donahue

NES MIDI → Events → Transformer-XL [Dai+ 2019]

NO_NOTEON_13, DT_732, NO_NOTEOFF, DT_2197, P1_NOTEON_76, DT_2, P1_NOTEON_70, DT_1463, P2_NOTEON_87, DT_1, P2_NOTEOFF, DT_148, ...
P1_NOTEON_61, DT_41, P2_NOTEON_85, DT_26, TR_NOTEON_39, DT_2, TR_NOTEON_43, DT_23, NO_NOTEON_9, DT_646, NO_NOTEON_13, DT_733
NES-MDB
(Donahue+ 2018)
46 hours

Lakh MIDI Dataset
9000+ hours
Goal: Pre-train on Lakh MIDI dataset, fine-tune on NES-MDB
Improving Automatic Jazz Melody Generation by Transfer Learning Techniques”, Hung et al. 2019
"Learning Music Helps You Read: Using Transfer to Study Linguistic Structure in Language Models", Papadimitriou & Jurafsky, 2020
"Transfer Learning for Music Classification and Regression Tasks", Choi et al. 2017

“Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”, Raffel, Shazeer, Roberts, Lee et al. 2019

“Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching”, Raffel 2016

“LakhNES: Improving multi-instrumental music generation with cross-domain pre-training”, Donahue et al. 2019

“Improving Automatic Jazz Melody Generation by Transfer Learning Techniques”, Hung et al. 2019

“Learning Music Helps You Read: Using Transfer to Study Linguistic Structure in Language Models”, Papadimitriou & Jurafsky, 2020

“Transfer Learning for Music Classification and Regression Tasks”, Choi et al. 2017


Thanks! Questions?