Simple and Efficient ways to Improve REALM
Open Domain QA

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Retrieval Augmented Language Model (REALM)

Retrieval Augmented Language Model (REALM)

The [MASK] at the top of the pyramid

\[ \text{[MASK]} = \text{pyramidion} \]

Neural Knowledge Retriever

Knowledge Augmented Encoder

Answer

What's the angle of an equilateral triangle?

60 degrees

Dense Passage Retrieval

What's the angle of an equilateral triangle?

In the familiar Euclidean geometry, an equilateral triangle is also equiangular.

What's the angle of an equilateral triangle?

Input Query

Neural Knowledge Retriever

Knowledge Augmented Encoder

Answer

60 degrees
REALM QA Finetuning

What's the angle of an equilateral triangle?

60 degrees
REALM QA Finetuning

What's the angle of an equilateral triangle?

60 degrees

Neural Knowledge Retriever

Knowledge Augmented Encoder

Answer

60 degrees

Input Query

$c = 5000$

$k = 5$

$s_{retr}(p_i, q) = h_q^\top h_p$
REALM QA Finetuning

What’s the angle of an equilateral triangle?

60 degrees

Neural Knowledge Retriever

Neural Knowledge Retriever

Knowledge Augmented Encoder

Supervision

Distant Sup - Does Passage Have Answer?

Gold Sup - Exact Span Match

Textual Knowledge Corpus

retrieve

c = 5000

$k = 5$

Answer

60 degrees
REALM QA Finetuning

Input Query

What’s the angle of an equilateral triangle?

Neural Knowledge Retriever

Knowledge Augmented Encoder

Answer

60 degrees

Supervision

Distant Sup - Does Passage Have Answer?

Gold Sup - Exact Span Match

Textual Knowledge Corpus

retrieve

c = 5000

Training = 12 GB GPU, 1 BS
Pre-Training = CC-News
## Bottlenecks in REALM

<table>
<thead>
<tr>
<th>Metric</th>
<th>NQ</th>
<th>WQ</th>
<th>CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test EM (Guu et al)</td>
<td>40.4</td>
<td>40.7</td>
<td>42.9</td>
</tr>
<tr>
<td>Test EM (Ours)</td>
<td>39.4</td>
<td>40.8</td>
<td>39.3</td>
</tr>
</tbody>
</table>
Bottlenecks in REALM

Reader Performance

Exact-Match Acc

- NQ
- WQ
- CT

EM
Upper Bound
Bottlenecks in REALM

Retriever Performance

- R@5
- R@10
- R@100
- R@1000
- R@5000

Recall

- NQ
- WQ
- CT
Training Scaling

- 1 12GB GPU → 8 TPU v3 core
- Batch Size = 1 Batch Size → 16 Batch Size
- TPU MIPS
  - TPU Exact Top-K
  - Efficient TPU Top-K - Binned Approximate
- Reader: k=5 → k=10
## Training Scaling

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Test Acc</th>
<th>Dev Acc</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>REALM</td>
<td>39.4</td>
<td>35.6</td>
<td>68.8</td>
</tr>
<tr>
<td>+Scale</td>
<td>42.8</td>
<td>37.9</td>
<td>69.5</td>
</tr>
</tbody>
</table>
Supervision

- **Supervision in REALM**
  - Reader - Span Match - Gold Label Supervision
  - Retriever - Has Answer - Distant Supervision

- **Has Answer - Simple Match if document has target answer**
  - Ambiguous and Noisy Signal
  - Unrelated Documents get positive signal

- **Gold Supervision** - expensive to obtain

- **Weak Supervision** - cheap and easily applicable to large datasets
Supervision

Q = Which president supported the creation of the Environmental Protection Agency (EPA)?

Ret Passage = Some historians say that President Richard Nixon’s southern strategy turned the southern United States into a republican stronghold, while others deem economic factors more important in the change.

Gold Passage = The Environmental Protection Agency (EPA) is an agency of the federal government of the United States created for the purpose of protecting human health and the environment. President Richard Nixon proposed the establishment of EPA and it began operation on December 2, 1970, after Nixon signed an executive order.
Supervision

- Gold Label Supervision for Retriever
  - Human Annotated Evidence Passages

- Natural Questions
  - Annotations for Candidate Passages - Long Answer
  - Relevant Passage with Answer Span

- Passages have small differences - Exact Match is restrictive
  - Passage with 50% word overlap with target passage is considered gold label
## Supervision

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Test Acc</th>
<th>Dev Acc</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>REALM</td>
<td>39.4</td>
<td>35.6</td>
<td>68.8</td>
</tr>
<tr>
<td>+Scale</td>
<td>42.8</td>
<td>37.9</td>
<td>69.5</td>
</tr>
<tr>
<td>+Scale+PS</td>
<td>43.2</td>
<td>38.6</td>
<td>69.9</td>
</tr>
</tbody>
</table>
Inference Scaling

- Scaling Reader to Process More Documents
  - Memory Constraints
  - Expensive - More Resources
  - Dedicated Architecture

- Read More Documents - Inference
  - Use extra memory from Optimization Storage
  - Increase No: Documents processed parallely by reader
# Inference Scaling

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Test Acc</th>
<th>Dev Acc</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>REALM</td>
<td>39.4</td>
<td>35.6</td>
<td>68.8</td>
</tr>
<tr>
<td>+Scale</td>
<td>42.8</td>
<td>37.9</td>
<td>69.5</td>
</tr>
<tr>
<td>+Scale+PS</td>
<td>43.2</td>
<td>38.6</td>
<td>69.9</td>
</tr>
<tr>
<td>+Scale+PS - 100 docs</td>
<td>44.8</td>
<td>38.6</td>
<td>69.9</td>
</tr>
</tbody>
</table>
Inference Scaling

Effect of no: reader documents on EM Acc
Cross-Document Passage Reranking

What's the angle of an equilateral triangle?

60 degrees
Cross-Document Passage Reranking

Document Reps

Question Rep $q$

Cross-Document Aware Rep

Question Aware Rep

Self-Attention Transformer Block

Cross-Attention Transformer Block

$\text{Ranking Score} = d''_i q^T$
## Cross-Document Passage Reranking

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Test Acc</th>
<th>Dev Acc</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>REALM</td>
<td>39.4</td>
<td>35.6</td>
<td>68.8</td>
</tr>
<tr>
<td>+Scale</td>
<td>42.8</td>
<td>37.9</td>
<td>69.5</td>
</tr>
<tr>
<td>+Scale+PS</td>
<td>43.2</td>
<td>38.6</td>
<td>69.9</td>
</tr>
<tr>
<td>+Scale+PS - 100 docs</td>
<td>44.8</td>
<td>38.6</td>
<td>69.9</td>
</tr>
<tr>
<td>+Scale+Rerank</td>
<td>42.3</td>
<td>37.4</td>
<td>67.5</td>
</tr>
</tbody>
</table>
# Cross-Document Passage Reranking

<table>
<thead>
<tr>
<th>Model</th>
<th>R@10</th>
<th>Dev EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>REALM</td>
<td>68.8</td>
<td>35.6</td>
</tr>
<tr>
<td>+Scale (Fixed Ret)</td>
<td>59.6</td>
<td>33.1</td>
</tr>
<tr>
<td>+Scale +Rerank (Fixed Ret)</td>
<td>67.9</td>
<td>35.8</td>
</tr>
<tr>
<td>+Scale +Rerank +PS (Fixed Ret)</td>
<td>67.5</td>
<td>37.1</td>
</tr>
</tbody>
</table>
## Cross-Document Passage Reranking

<table>
<thead>
<tr>
<th>Model</th>
<th>R@10</th>
<th>Dev EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>REALM</td>
<td>68.8</td>
<td>35.6</td>
</tr>
<tr>
<td>+Scale (Fixed Ret)</td>
<td>59.6</td>
<td>33.1</td>
</tr>
<tr>
<td>+Scale +Rerank (Fixed Ret)</td>
<td>67.9</td>
<td>35.8</td>
</tr>
<tr>
<td>+Scale +Rerank +PS (Fixed Ret)</td>
<td>67.5</td>
<td>37.1</td>
</tr>
<tr>
<td>+Scale (Trained Ret)</td>
<td>69.5</td>
<td>37.9</td>
</tr>
<tr>
<td>+Scale +Rerank (Trained Ret)</td>
<td>67.5</td>
<td>37.4</td>
</tr>
</tbody>
</table>
REALM++

- Training Setup Scaling
  - Distributed Training on TPUs
  - Increased Batch Size
  - Exact MIPS

- Gold Passage Supervision
  - Human Annotations on Evidence Passages

- Increased Reader Documents during Inference
  - Train with 10 docs, Predict with 100 docs
# REALM++ v/s Same size models

<table>
<thead>
<tr>
<th>Model</th>
<th>NQ</th>
<th>WQ</th>
<th>CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25+BERT (Lee et al., 2019)</td>
<td>26.5</td>
<td>17.7</td>
<td>21.3</td>
</tr>
<tr>
<td>ORQA (Lee et al., 2019)</td>
<td>33.3</td>
<td>36.4</td>
<td>30.1</td>
</tr>
<tr>
<td>REALM (Guu et al., 2019)</td>
<td>39.2</td>
<td>40.2</td>
<td>46.8</td>
</tr>
<tr>
<td>REALM$_{News}$ (Guu et al., 2019)</td>
<td>40.4</td>
<td>40.7</td>
<td>42.9</td>
</tr>
<tr>
<td>DPR (Karpukhin et al., 2020)</td>
<td>41.5</td>
<td>42.4</td>
<td>49.4</td>
</tr>
<tr>
<td>REALM++ (10 doc)</td>
<td>43.2</td>
<td>44.5</td>
<td>47.2</td>
</tr>
<tr>
<td>REALM++ (100 doc)</td>
<td>44.8</td>
<td>45.6</td>
<td>49.7</td>
</tr>
</tbody>
</table>
# REALM++ v/s Large models

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Size</th>
<th>NQ</th>
<th>WQ</th>
<th>CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>REALM</td>
<td>Base</td>
<td>39.2</td>
<td>40.2</td>
<td>46.8</td>
</tr>
<tr>
<td>REALM&lt;sub&gt;News&lt;/sub&gt;</td>
<td>Base</td>
<td>40.4</td>
<td>40.7</td>
<td>42.9</td>
</tr>
<tr>
<td>DPR</td>
<td>Base</td>
<td>41.5</td>
<td>42.4</td>
<td>49.4</td>
</tr>
<tr>
<td>REALM++ (10 doc)</td>
<td>Base</td>
<td>43.2</td>
<td>44.5</td>
<td>47.2</td>
</tr>
<tr>
<td>REALM++ (100 doc)</td>
<td>Base</td>
<td>44.8</td>
<td>45.6</td>
<td>49.7</td>
</tr>
<tr>
<td>RAG&lt;sub&gt;Large&lt;/sub&gt;</td>
<td>Large</td>
<td>44.5</td>
<td>45/5</td>
<td>52.2</td>
</tr>
<tr>
<td>ReConsider&lt;sub&gt;Large&lt;/sub&gt;</td>
<td>Large</td>
<td>45.5</td>
<td>45.9</td>
<td>55.3</td>
</tr>
</tbody>
</table>
Speed and Memory Usage

● Increased Speed
  ○ TPU Efficiency + Larger Batch Training
  ○ 4x more examples per second wrt REALM

● Training Time
  ○ Reduces from 48 hours to 12 hours for same epochs

● Memory utilization
  ○ Increases ~5GB due to loading the index in memory
  ○ Fits within 12GB ~ Dragonfish
Summary!

- REALM was significantly undertrained - Works better than previously known

- Scale - plays an important role, accounts for large gains
  - Better training, optimization
  - Larger batch-size

- Dense Retrieval systems should be compared by normalizing training factors like batch size to understand the actual benefit of a method

- Reading more documents during inference is a quick easy way to boost performance!
Directions for Future Work

● Reader Bottleneck
  ○ Span Identification is problematic
  ○ Better Readers - improved reasoning
  ○ Incorporating more context - Routing Transformer, Longformer, etc

● Incorporating reranking modules
  ○ Reranking - cheap method for cross-document interaction
  ○ Optimization problems with retriever - currently doesn’t improve
  ○ Better methods to optimize pre-trained retriever and untrained reranker needed
Thank You!

Questions?

Contact -

vbalacha@cs.cmu.edu,
nikip@google.com,
avaswani@google.com
Open Domain QA

Q: How many of Warsaw’s inhabitants spoke Polish in 1933?

Document Retriever

Wikipedia

Document Reader

833,500

Sparse Retriever

Bag of Words
Tf-IDF
BM25

Dense Retriever

REALM
DPR
What's the angle of an equilateral triangle?

Answer: 60 degrees