Simple and Efficient ways to Improve REALM
Open Domain QA

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

Chen, Danqi, et al. "Reading wikipedia to answer open-domain questions." (2017).
Retrieval Augmented Language Model (REALM)

The [MASK] at the top of the pyramid

The pyramidion on top allows for less material higher up the pyramid.

[CLS] The [MASK] at the top of the pyramid [SEP] The pyramidion on top allows for less material higher up the pyramid.

Knowledge Augmented Encoder

Answer

[MASK] = pyramidion

Guu, Kelvin, et al. "Realm: Retrieval-augmented language model pre-training." (2020).
Retrieval Augmented Language Model (REALM)

The [MASK] at the top of the pyramid

Neural Knowledge Retriever

Knowledge Augmented Encoder

Answer

[MASK] = pyramidion

Unlabeled text

Textual Knowledge Corpus

Input Query

What's the angle of an equilateral triangle?

Neural Knowledge Retriever

Textual Knowledge Corpus

Knowledge Augmented Encoder

Answer

60 degrees

Guu, Kelvin, et al. "Realm: Retrieval-augmented language model pre-training." (2020).
Dense Passage Retrieval

What's the angle of an equilateral triangle?

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

Karpukhin, Vladimir, et al. "Dense passage retrieval for open-domain question answering." (2020).
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

c = 5000

\[ S_{retr}(p_i, q) = h_q^T h_{p_i} \]
REALM QA Finetuning

What's the angle of an equilateral triangle?

60 degrees

$c = 5000$

$k = 5$

$S_{rerr}(p_i, q) = h_q^\top h_{p_i}$

Neural Knowledge Retriever

Knowledge Augmented Encoder

60 degrees
REALM QA Finetuning

What’s the angle of an equilateral triangle?

60 degrees

Textual Knowledge Corpus

Neural Knowledge Retriever

Knowledge Augmented Encoder

Input Query

Supervision

Distant Sup - Does Passage Have Answer?

Gold Sup - Exact Span Match

Answer

60 degrees

$c = 5000$

$k = 5$

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

What’s the angle of an equilateral triangle?

60 degrees

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

| Metric               | NQ  | WQ  | CT  |
|----------------------|-----|-----|-----|
| Test EM (Guu et al)  | 40.4| 40.7| 42.9|
| Test EM (Ours)       | 39.4| 40.8| 39.3|
Bottlenecks in REALM

Reader Performance

|        | EM    | Upper Bound |
|--------|-------|-------------|
| NQ     |       |             |
| WQ     | 40    |             |
| CT     | 40    |             |
Bottlenecks in REALM
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

| Experiments | Test Acc | Dev Acc | R@10 |
|-------------|----------|---------|------|
| REALM       | 39.4     | 35.6    | 68.8 |
| +Scale      | 42.8     | 37.9    | 69.5 |
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

| Experiments    | Test Acc | Dev Acc | R@10 |
|----------------|----------|---------|------|
| REALM          | 39.4     | 35.6    | 68.8 |
| +Scale         | 42.8     | 37.9    | 69.5 |
| +Scale+PS      | 43.2     | 38.6    | 69.9 |
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

| Experiments                | Test Acc | Dev Acc | R@10 |
|----------------------------|----------|---------|-------|
| REALM                      | 39.4     | 35.6    | 68.8  |
| +Scale                     | 42.8     | 37.9    | 69.5  |
| +Scale+PS                  | 43.2     | 38.6    | 69.9  |
| +Scale+PS - 100 docs       | 44.8     | 38.6    | 69.9  |
Inference Scaling

Effect of no: reader documents on EM Acc

Test EM Acc

No: of documents

25 50 75 100 125 150 175
Cross-Document Passage Reranking

What's the angle of an equilateral triangle?

60 degrees
Cross-Document Passage Reranking

Document Reps

Question Rep q

Self-Attention Transformer Block

Cross-Attention Transformer Block

Cross-Document Aware Rep

Question Aware Rep

Ranking Score = d''_i q^T
## Cross-Document Passage Reranking

| Experiments                  | Test Acc | Dev Acc | R@10 |
|------------------------------|----------|---------|------|
| REALM                        | 39.4     | 35.6    | 68.8 |
| +Scale                       | 42.8     | 37.9    | 69.5 |
| +Scale+PS                    | 43.2     | 38.6    | 69.9 |
| +Scale+PS - 100 docs         | 44.8     | 38.6    | 69.9 |
| +Scale+Rerank                | 42.3     | 37.4    | 67.5 |
## Cross-Document Passage Reranking

| Model                                  | R@10 | Dev EM |
|----------------------------------------|------|--------|
| REALM                                  | 68.8 | 35.6   |
| +Scale (Fixed Ret)                     | 59.6 | 33.1   |
| +Scale +Rerank (Fixed Ret)             | 67.9 | 35.8   |
| +Scale +Rerank +PS (Fixed Ret)         | 67.5 | 37.1   |
## Cross-Document Passage Reranking

| Model                                      | R@10  | Dev EM |
|--------------------------------------------|-------|--------|
| REALM                                      | 68.8  | 35.6   |
| +Scale (Fixed Ret)                         | 59.6  | 33.1   |
| +Scale +Rerank (Fixed Ret)                 | 67.9  | 35.8   |
| +Scale +Rerank +PS (Fixed Ret)             | 67.5  | 37.1   |
| +Scale (Trained Ret)                       | 69.5  | 37.9   |
| +Scale +Rerank (Trained Ret)               | 67.5  | 37.4   |
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

| Model                                | NQ  | WQ  | CT  |
|--------------------------------------|-----|-----|-----|
| BM25+BERT (Lee et al., 2019)         | 26.5| 17.7| 21.3|
| ORQA (Lee et al., 2019)              | 33.3| 36.4| 30.1|
| REALM (Guu et al., 2019)             | 39.2| 40.2| 46.8|
| REALM\textsubscript{News} (Guu et al., 2019) | 40.4| 40.7| 42.9|
| DPR (Karpukhin et al., 2020)         | 41.5| 42.4| 49.4|
| REALM++ (10 doc)                      | 43.2| 44.5| 47.2|
| REALM++ (100 doc)                     | 44.8| 45.6| 49.7|
## REALM++ v/s Large models

| Model               | Model Size | NQ  | WQ  | CT  |
|---------------------|------------|-----|-----|-----|
| REALM               | Base       | 39.2| 40.2| 46.8|
| REALM\textsubscript{News} | Base       | 40.4| 40.7| 42.9|
| DPR                 | Base       | 41.5| 42.4| 49.4|
| REALM++ (10 doc)    | Base       | 43.2| 44.5| 47.2|
| REALM++ (100 doc)   | Base       | 44.8| 45.6| 49.7|
| RAG\textsubscript{Large} | Large     | 44.5| 45/5| 52.2|
| ReConsider\textsubscript{Large} | Large     | 45.5| 45.9| 55.3|
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?

Sparse Retriever
- Bag of Words
- Tf-IDF
- BM25

Dense Retriever
- REALM
- DPR

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies
What's the angle of an equilateral triangle?

Answer: 60 degrees