ACL 2023 Tutorial:

**Retrieval-based Language Models and Applications**

Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen

https://acl2023-retrieval-lm.github.io/

July 9, 2023
About presenters

Akari Asai
PhD student
@UW

Sewon Min
PhD student
@UW

Zexuan Zhong
PhD student
@Princeton

Danqi Chen
Faculty
@Princeton
Participation and Q & A

• All tutorial slides and reading list are available at:
  
  https://acl2023-retrieval-lm.github.io/

• Throughout the tutorial, we will provide Q & A on sli.do (also on the website):
  
  https://tinyurl.com/retrieval-lm-tutorial

Ask & upvote questions anytime! (+10 minutes Q & A in 1st and 2nd half)
Retrieval-based language models (LMs)

Retrieval-based LMs = Retrieval + LMs

• It is a language model
  \[ P(x_n \mid x_1, x_2, \ldots, x_{n-1}) \]

  The capital city of Ontario is ___

  (can be broadly extended to masked language models or encoder-decoder models)

• It retrieves from an external datastore (at least during inference time)

(Also referred to semiparametric and non-parametric models)
The age of large language models (LLMs)

- Transformers-based, fully parametric
- Trained on next-token prediction tasks (+ RLHF; not the focus today)
- Model size ↑, data size ↑

Image: https://lifearchitect.ai/models/
Retrieval for knowledge-intensive NLP tasks

Representative tasks: open-domain QA, fact checking, entity linking, ..

Drives a lot of research on better algorithms for dense retrieval, e.g., DPR (Karpukhin et al., 2020), ColBERT (Khattab and Zaharia, 2020), ANCE (Xiong et al., 2021), Contriever (Izacard et al., 2022), …

Why retrieval → LMs?

Image: http://ai.stanford.edu/blog/retrieval-based-NLP/
Why retrieval-based LMs?

LLMs can’t memorize all (long-tail) knowledge in their parameters

List 5 important papers authored by Geoffrey Hinton

Geoffrey Hinton is a renowned computer scientist …
Here are five important papers authored by him:

1. "Learning Internal Representations by Error Propagation" (with D. E. Rumelhart and R. J. Williams) - This paper, published in 1986, … ✓
2. "Deep Boltzmann Machines" (with R. Salakhutdinov) - Published in 2009, … ✓
3. "Deep Learning" (with Y. Bengio and A. Courville) - Published as a book in 2016,… ×
4. "Attention Is All You Need" (with V. Vaswani, N. Shazeer, et al.) - Published in 2017, this paper introduced the Transformer model,… ×

What is Kathy Saltzman’s occupation?

(Mallen et al., 2023)

GPT-3 davinci-003: 20%-30% accuracy
Why retrieval-based LMs?

LLMs’ knowledge is easily outdated and hard to update

Who is the CEO of Twitter?

As of my knowledge cutoff in September 2021, the CEO of Twitter is Jack Dorsey.

• Existing knowledge editing methods are still NOT scalable (active research!)

• The datastore can be easily updated and expanded - even without retraining!
Why retrieval-based LMs?

LLMs’ output is challenging to interpret and verify

Can trace knowledge source from retrieval results - better interpretability & control

Generating text with citations

When did the US break away from England?

The US took the first step towards gaining independence from GB when it declared independence on July 2, 1776 (although the event is now commemorated on July 4, 1776, the date when the Declaration of Independence was officially adopted by Congress) [1][2]. The Treaty of Paris was later signed on September 3, 1783, formally separating the United States from the British Empire [3].

(Nakano et al. 2021; Menick et al., 2022; Gao et al., 2023)
Why retrieval-based LMs?

LLMs’ output is challenging to interpret and verify

Create an itinerary for exploring Toronto over a span of two days.

Toronto is a vibrant and diverse city with many attractions and activities to suit different tastes and preferences. Here are some suggestions for creating an itinerary for exploring Toronto over two days:

- **Day 1**: Start your day with a visit to the **CN Tower**, the iconic landmark that offers panoramic views of the city and beyond. You can also experience the thrill of walking on the glass floor or the EdgeWalk, the world’s highest hands-free walk. Next, head to the nearby **St. Lawrence Market**, one of the world’s best food markets, where you can sample a variety of cuisines and local specialties. After lunch, take a stroll along **Queen West**, a trendy neighborhood with eclectic shops, galleries, cafes and street art. In the evening, enjoy a show at one of the many theaters or comedy clubs in the **Entertainment District**, or catch a game at the **Scotiabank Arena** if you’re a sports fan.

- **Day 2**: Explore the history and culture of Toronto at the **Royal Ontario Museum**, one of the largest and most comprehensive museums in North America, featuring exhibits on art, natural history, world cultures and more. Then, hop on a ferry to the **Toronto Islands**, a group of islands that offer a relaxing escape from the city, with beaches, parks, trails and amusement rides. You can also rent a bike or kayak to explore the islands at your own pace. For dinner, head to **Chinatown**, one of the largest and most vibrant in North America, where you can find a variety of Asian cuisines and shops.

I hope this helps you plan your trip to Toronto. Have fun! 😊

Learn more:

1. cntower.ca  
2. travel.usnews.com  
3. bing.com  
4. rom.on.ca  
5. tripadvisor.com
Why retrieval-based LMs?

LLMs are shown to easily leak private training data

Individualization on private data by storing it in the datastore

Carlini et al. 2021. “Extracting Training Data from Large Language Models”
Why retrieval-based LMs?

LLMs are *large* and expensive to train and run

Long-term goal: can we possibly reduce the training and inference costs, and scale down the size of LLMs?

e.g., RETRO (Borgeaud et al., 2021): “obtains comparable performance to GPT-3 on the Pile, despite using 25x fewer parameters”
Goals of the tutorial

• We will cover a number of key developments on retrieval-based LMs (mostly 2020–2023)
  • Problem definition
  • Architecture design
  • Training methods
  • Applications and extensions

• This tutorial is cutting-edge, and we are still far from understanding how to best develop retrieval-based LMs compared to parametric LLMs:
  • Taxonomies of existing research and key insights
  • Our perspectives on the current challenges & open problems
# Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Section</th>
<th>Presenter</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00—14:15</td>
<td>Section 1: Introduction [Slides]</td>
<td>Danqi</td>
<td></td>
</tr>
<tr>
<td>14:15—14:25</td>
<td>Section 2: Definition &amp; Preliminaries [Slides]</td>
<td>Sewon</td>
<td></td>
</tr>
<tr>
<td>14:25—15:00</td>
<td>Section 3: Retrieval-based LMs: Architecture [Slides]</td>
<td>Sewon</td>
<td></td>
</tr>
<tr>
<td>15:00—15:25</td>
<td>Section 4: Retrieval-based LMs: Training [Slides]</td>
<td>Zexuan</td>
<td></td>
</tr>
<tr>
<td>15:25—15:30</td>
<td>Q &amp; A Session I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00—16:25</td>
<td>Section 4 (Cont’d): Retrieval-based LMs: Training [Slides]</td>
<td>Zexuan</td>
<td></td>
</tr>
<tr>
<td>16:25—17:00</td>
<td>Section 5: Retrieval-based LMs: Applications [Slides]</td>
<td>Akari</td>
<td></td>
</tr>
<tr>
<td>17:00—17:10</td>
<td>Section 6: Extension: Multilingual &amp; Multimodal [Slides]</td>
<td>Akari</td>
<td></td>
</tr>
<tr>
<td>17:10—17:20</td>
<td>Section 7: Challenges &amp; Opportunities [Slides]</td>
<td>Danqi</td>
<td></td>
</tr>
<tr>
<td>17:20—17:30</td>
<td>Q &amp; A Session II</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**30min coffee break**