bit.ly/dexml

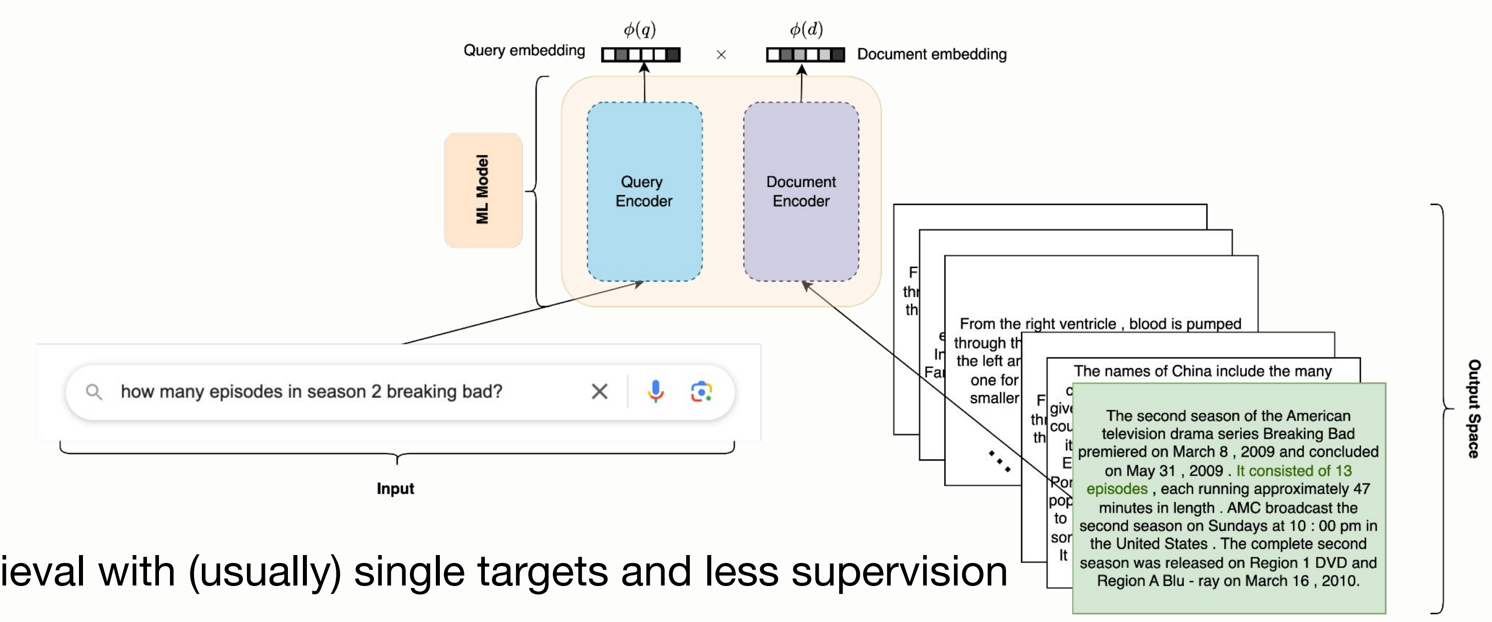
Dual-encoder for Extreme Multi-label Classification

Nilesh Gupta (nilesh@cs.utexas.edu), Devvrit Khatri, Ankit S. Rawat, Srinadh Bhojanapalli, Prateek Jain, Inderjit Dhillon

TEXAS The University of Texas at Austin · Google · ICLR

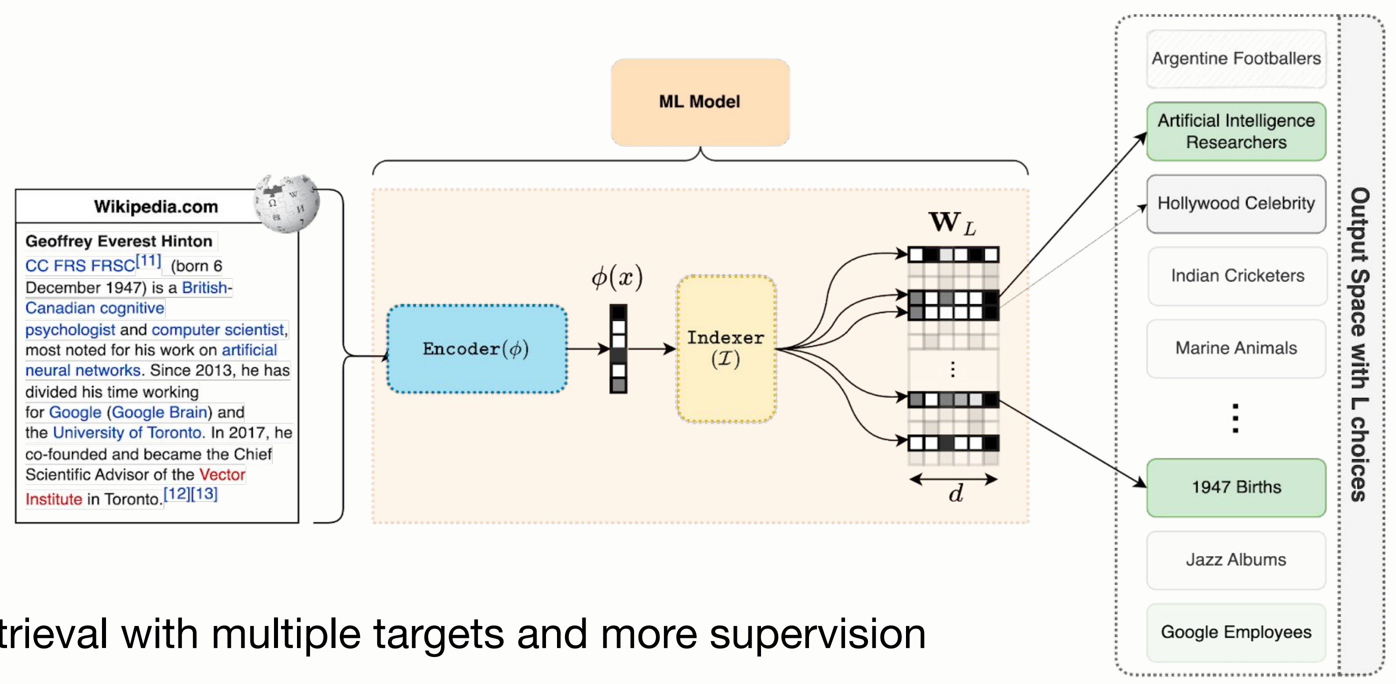
Premise

Information Retrieval



Retrieval with (usually) single targets and less supervision

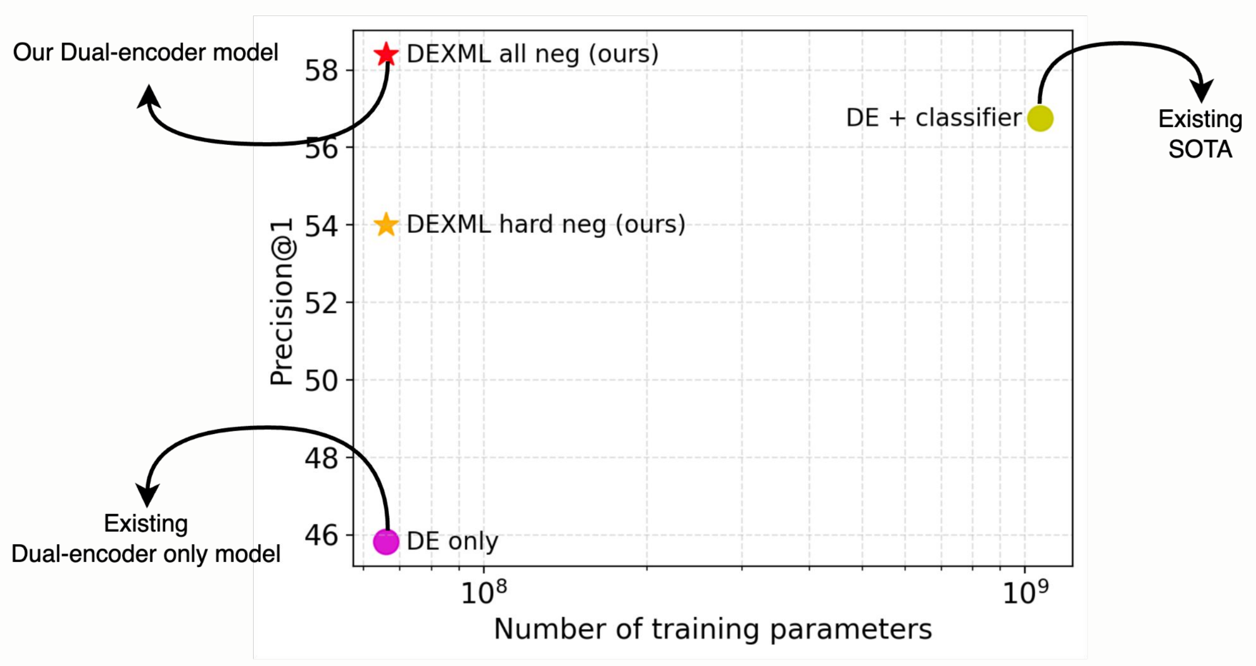
Extreme Multi-label Classification (XMC)



Retrieval with multiple targets and more supervision

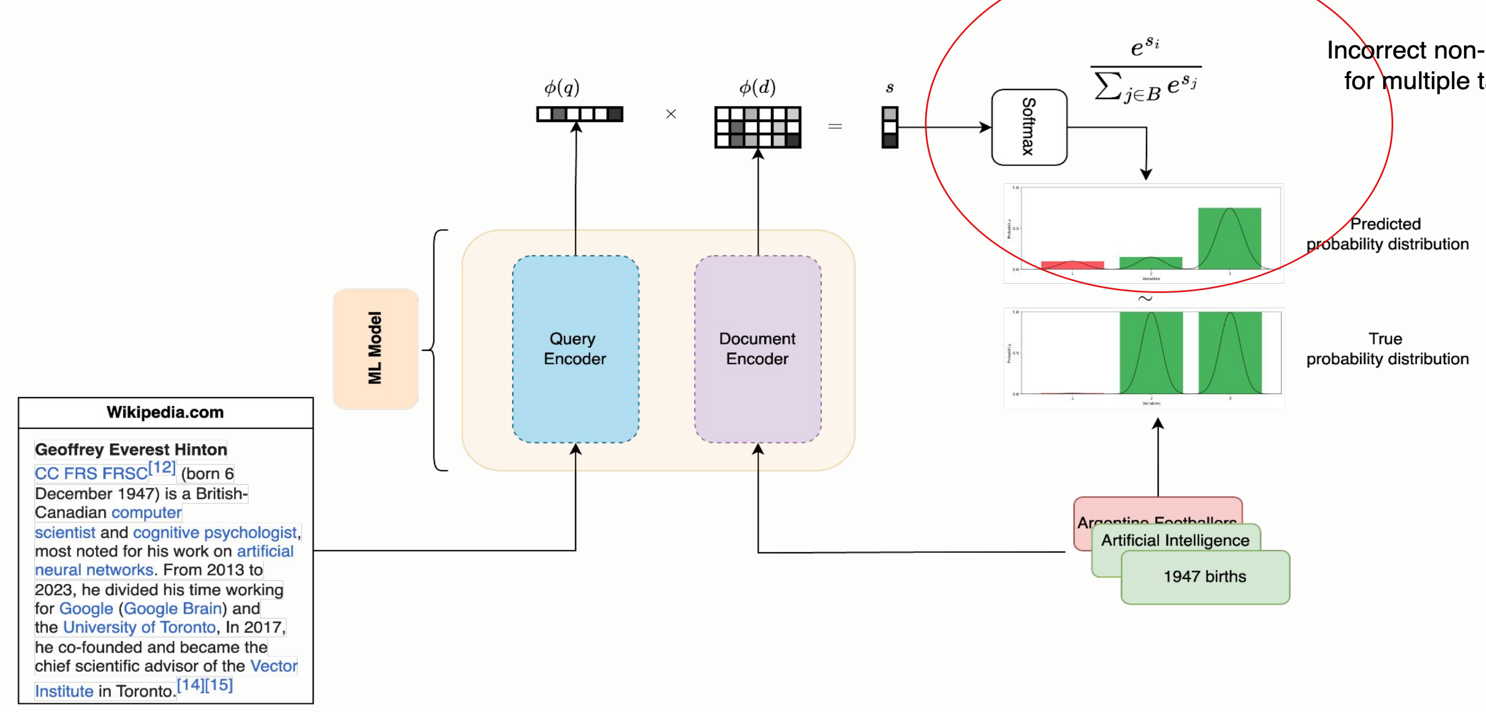
Dual-encoders (DE) for XMC?

- Model doesn't grow linearly with output space
- Better generalization on unseen items
- ~~Struggle with semantic gap~~
- ~~Underperform due to less capacity - bad memorization~~

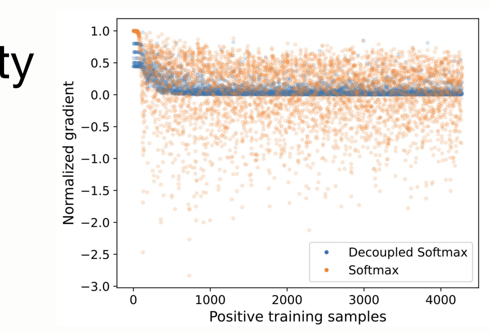


Research problem and Solution

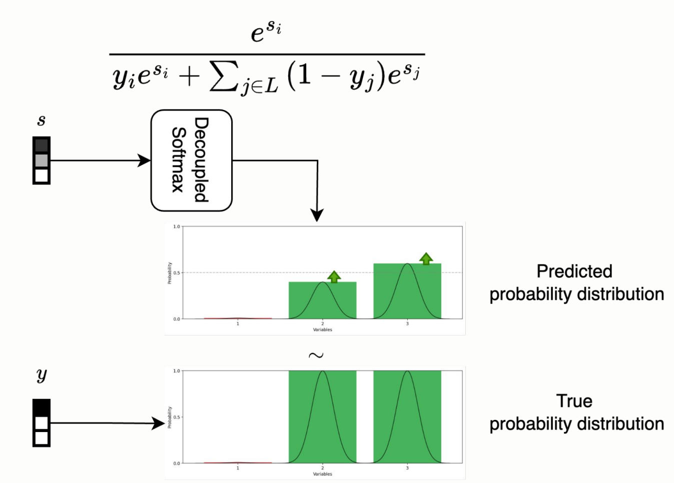
What's wrong with **InfoNCE?**



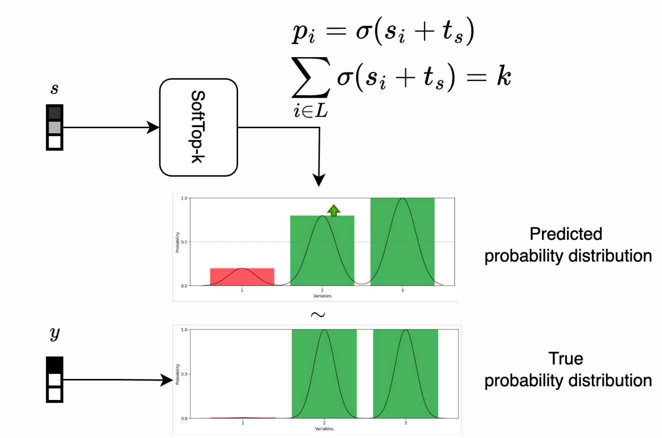
- Optimized when each positive get equal probability
- All probabilities sum upto 1
- Not all labels equally easy!



DecoupledSoftmax and **SoftTop-k**

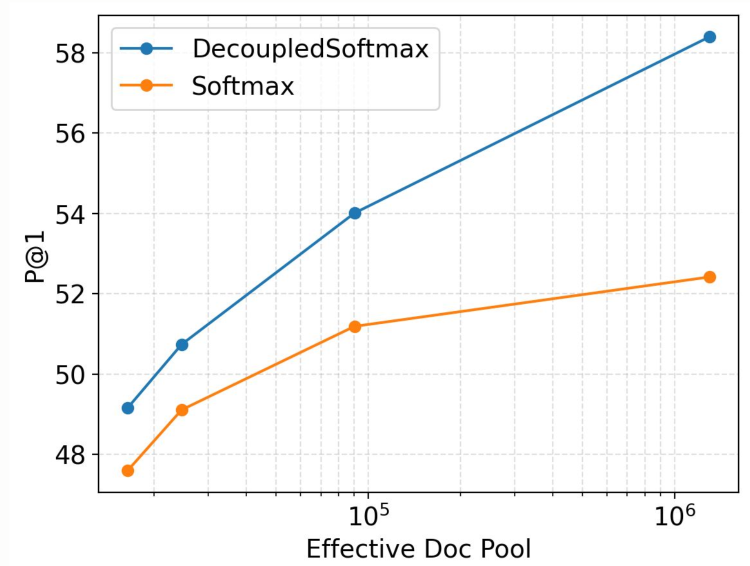


- Decouples probabilities of positives from each other
- ~~All probabilities sum upto 1~~

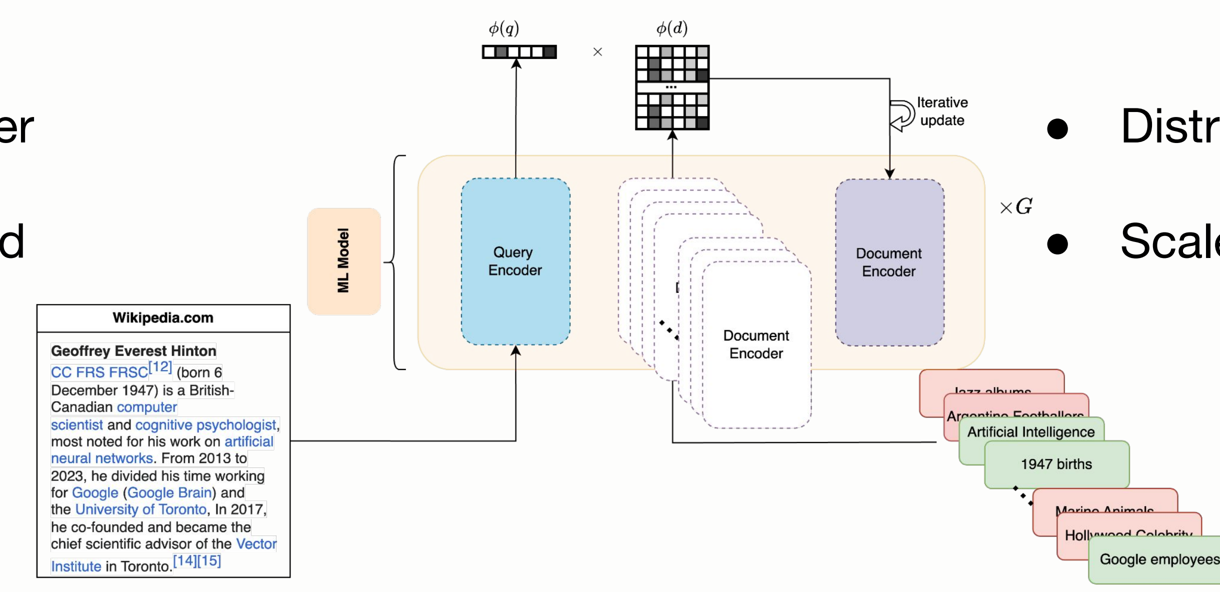


- Generalize softmax to allow probabilities to sum to given k
- Useful in top-k retrieval

Scaling Challenges



- More labels in loss computation better
- But incurs large memory footprint and computationally expensive



- Distributed Gradient caching!
- Scales to 1M labels in loss

Code
bit.ly/dexml

Results

LF-AmazonTitles-1.3M

Method	Params	P@1	P@5
XR-Transformer	3B	50.14	39.98
ELIAS	1B	49.26	39.29
NGAME	1B	56.75	44.09
DEXA	1B	56.63	43.90
DEXML (ours)	**66M**	**58.40**	**45.46**

Loss ablation (EURLex-4K)

Loss	P@1	P@5	R@100
BCE	0.1	0.07	1.84
Softmax	80.05	58.36	92.57
DecoupledSoftmax	**86.78**	60.19	91.75
SoftTop-5	83.42	**60.95**	91.30
SoftTop-100	52.34	37.41	**93.72**

Conclusions

- Showed dual-encoders are performant on XMC tasks
  - Parameter-efficient and generalizable approach for XMC
  - Universally applicable solutions for all retrieval setting
- Showed existing DE train losses not appropriate for multi-label setting
- DecoupledSoftmax and SoftTop-k, which overcomes limitations
- Applicable in multi-document retrieval augmented generation (RAG)