Research Paper Recommendation System

Using Document Level Embedding

Akshay Gupta | Anant Dev | Muhammed Shariq Nawaz | Gaurang Raje

Flow of our Talk

- Motivation
- Main Idea
- Existing Research
- Our Method
- Our Progress
- Our Hypothesis

Motivation

- Exponential rise in research papers published
- Accessing related documents is monotonous
- Simplify the literature review process
- Get relevant citations of related papers

Process Steps

- Language Transformer/ Sentence Transformer
- Document Embedding
- Output - The N-dimensional embeddings
- Cosine Similarity
- Final output - Top-K results

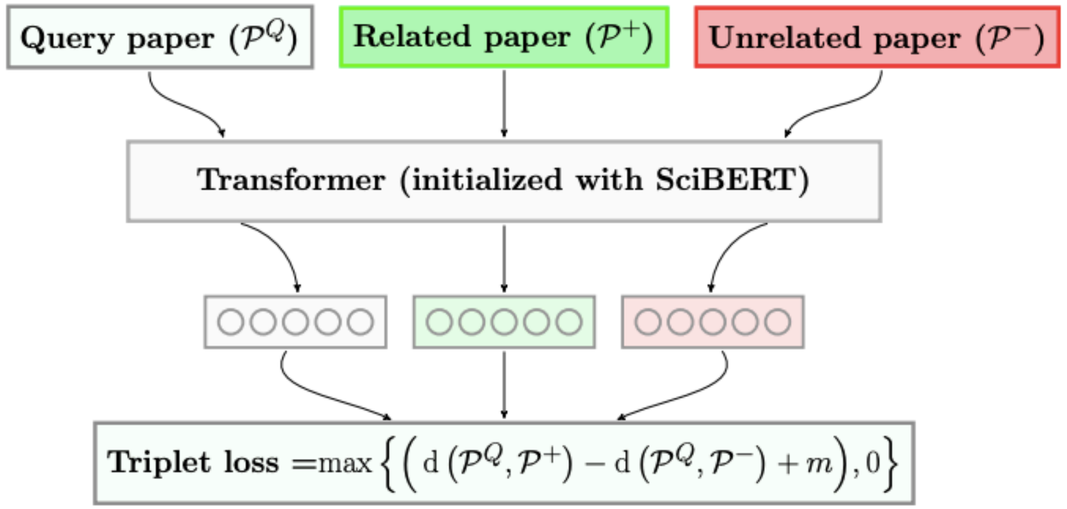


Figure 1: Overview of SPECTER.

Existing Research

Work by Haruna et al.

- Based on contextual metadata
- Heuristic algorithm to score related papers
- Normalised score returns top-N recommendations
- Evaluation metrics:

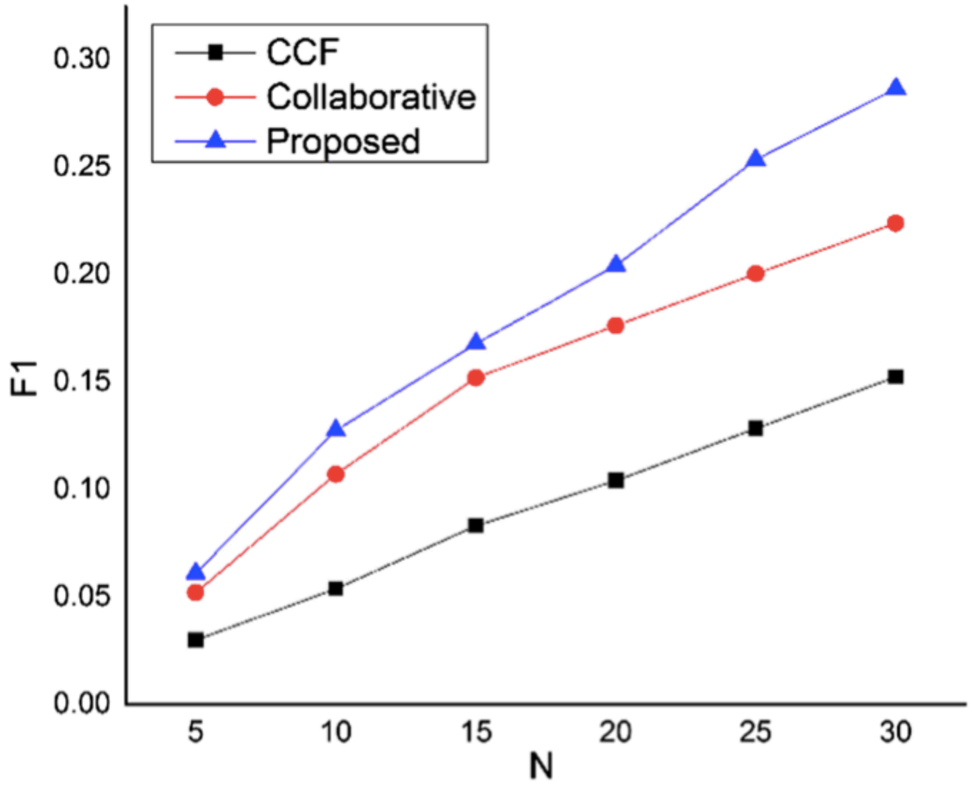
$$precision = \frac{\sum(relevant\_papers) \cap \sum(retrieved\_papers)}{\sum(retrieved\_papers)}$$

$$recall = \frac{\sum(relevant\_papers) \cap \sum(retrieved\_papers)}{\sum(retrieved\_papers)}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Work by Haruna et al.

- Results:

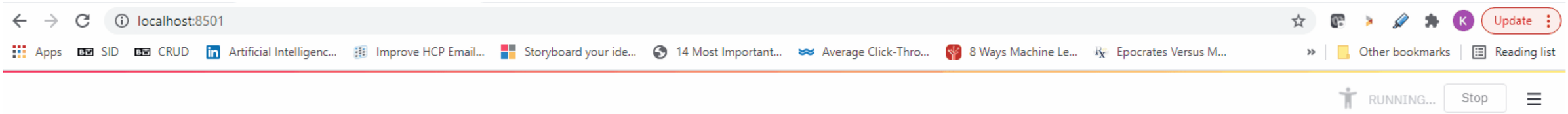


- Drawbacks:
  - Divergent Techniques papers
  - Relevance/preference score

Arxiv Neural Search Repository

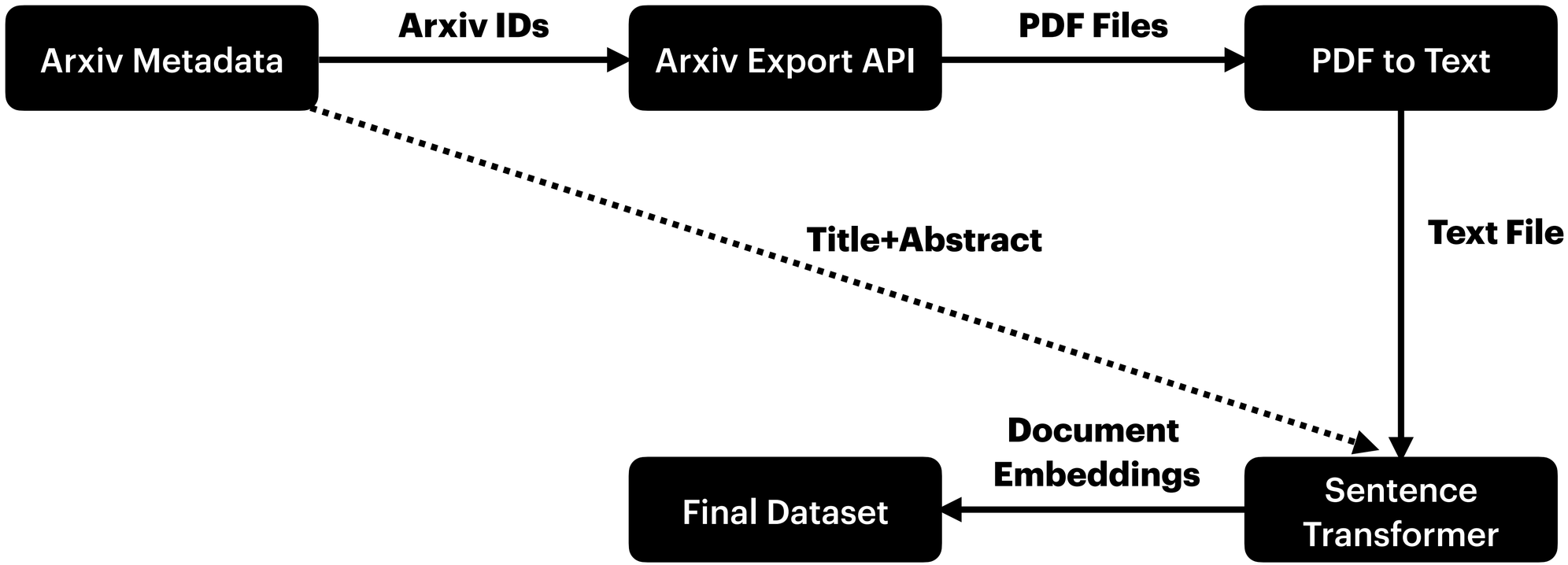
- Uses categories and abstracts for similarity search
- Sources above elements from Kaggle Arxiv Metadata dataset
- FAISS for indexing
- DistillBERT for generation of query embeddings

Arxiv Neural Search Repository

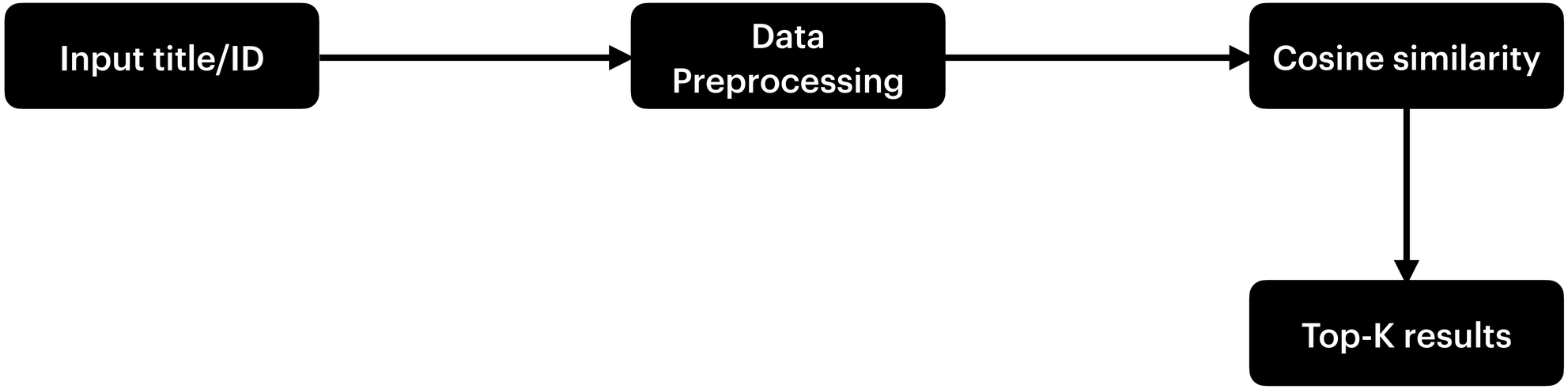


Our Method

Dataset Preprocessing

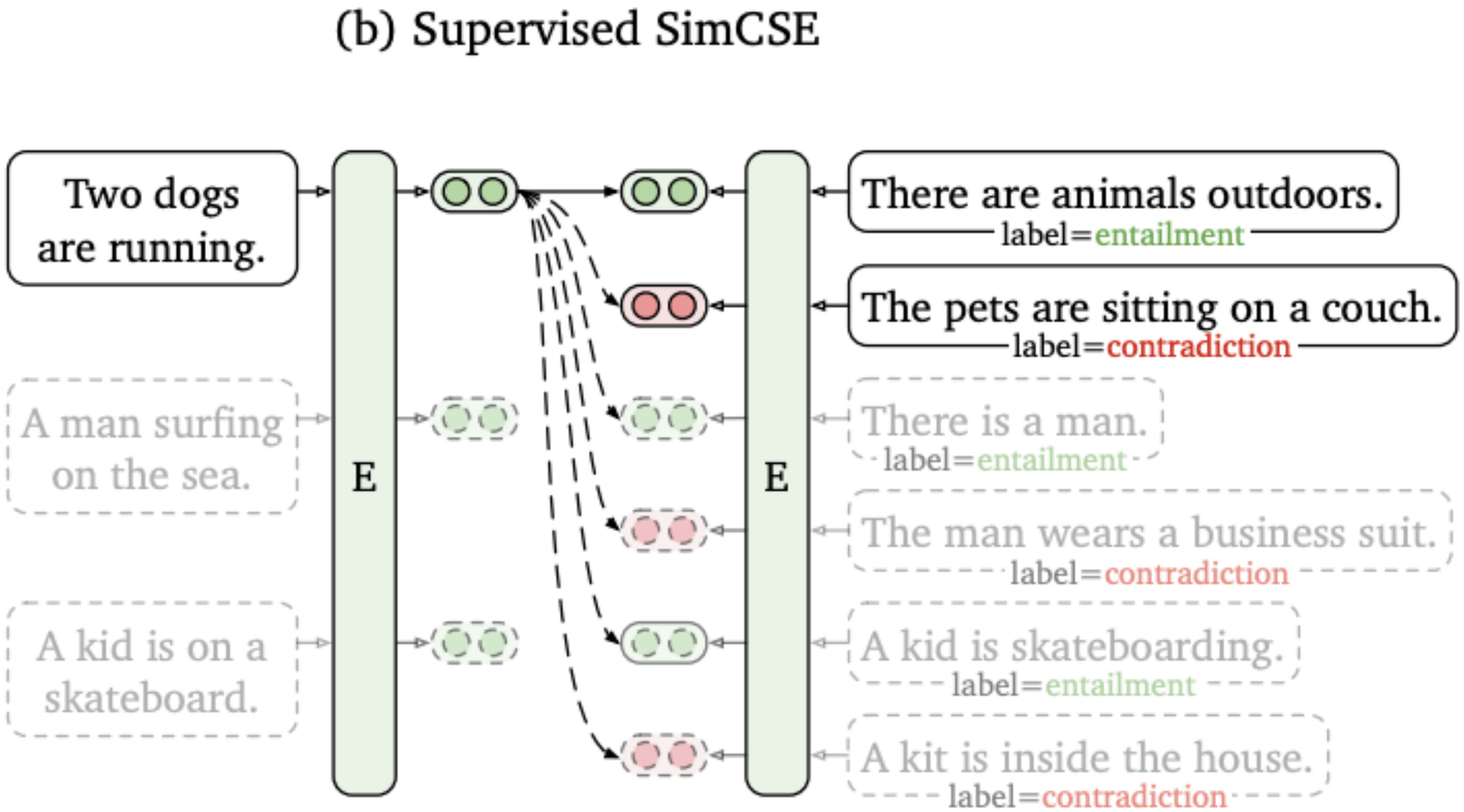


Testing on New Inputs



Pre-trained Models

- Contrastive learning:
  - SimCSE
  - DeCLUTTR
- SPECTER



Our Progress so far...

[11]:

```
# This paper was a EMNLP 2020 Honourable Mention Papers
search_papers(title='If beam search is the answer, what was the question?',
              abstract='Quite surprisingly, exact maximum a posteriori (MAP) decoding of neural language generato
```

Batches: 100% 1/1 [00:00<00:00, 20.50it/s]

	title	authors	arxiv_id	score
0	Infinite Viterbi alignments in the two state hidden Markov models	J. Lember, A. Koloydenko	0711.0928	0.819104
1	On the Vocabulary of Grammar-Based Codes and the Logical Consistency of\n Texts	{\L}ukasz D\k{e}bowski	0810.3125	0.814937
2	CLAIRLIB Documentation v1.03	Dragomir Radev, Mark Hodges, Anthony Fader, Mark Joseph, Joshua\n Gerrish, Mark Schaller, Jonathan dePeri, Bryan Gibson	0712.3298	0.806504
3	Questions & Answers for TEI Newcomers	Laurent Romary (LORIA)	0812.3563	0.803840
4	The emerging field of language dynamics	S. Wichmann	0801.1415	0.799843
5	In memoriam Maurice Gross	Eric Laporte (IGM-LabInfo)	0711.3452	0.799252
6	A constructive proof of the existence of Viterbi processes	J. Lember, A. Koloydenko	0804.2138	0.798086
7	Morphic and Automatic Words: Maximal Blocks and Diophantine\n Approximation	Yann Bugeaud, Dalia Krieger, Jeffrey Shallit	0808.2544	0.795205
8	Three Lectures on Automatic Structures	Bakhadyr Khoussainov, Mia Minnes	0809.3430	0.789193
9	UNL-French deconversion as transfer & generation from an interlingua\n with possible quality enhancement through offline human interaction	Gilles s\'erasset (IMAG, Clips - Imag, Lig), Christian Boitet (IMAG,\n Clips - Imag, Lig)	0811.0579	0.785540

Our Progress so far...

[10]:

```
is paper was the EMNLP 2020 Best Paper
ch_papers(title='Digital Voicing of Silent Speech',
          abstract='In this paper, we consider the task of digitally voicing silent speech, where silently mouthe
```

Batches: 100% 1/1 [00:00<00:00, 18.10it/s]

	title	authors	arxiv_id	score
0	EasyVoice: Integrating voice synthesis with Skype	Paulo A. Condado and Fernando G. Lobo	0706.3132	0.818029
1	Am\'elioration des Performances des Syst\`emes Automatiques de\n Reconnaissance de la Parole pour la Parole Non Native	Ghazi Bouselmi (INRIA Lorraine - LORIA), Dominique Fohr (INRIA\n Lorraine - LORIA), Irina Illina (INRIA Lorraine - LORIA), Jean-Paul Haton\n (INRIA Lorraine - LORIA)	0711.1038	0.807498
2	Adjusted Viterbi training for hidden Markov models	J. Lember, A. Koloydenko	0709.2317	0.770014
3	Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice\n Disorder Detection	Max A Little, Patrick E McSharry, Stephen J Roberts, Declan AE\n Costello and Irene M Moroz	0707.0086	0.766409
4	Probabilistic SVM/GMM Classifier for Speaker-Independent Vowel\n Recognition in Continues Speech	Mohammad Nazari, Abolghasem Sayadiyan, SeyedMajid Valiollahzadeh	0812.2411	0.761897
5	Acoustic Features and Perceptive Cues of Songs and Dialogues in Whistled\n Speech: Convergences with Sung Speech	Julien Meyer (LAB-Upc)	0711.3704	0.760309
6	Combined Acoustic and Pronunciation Modelling for Non-Native Speech\n Recognition	Ghazi Bouselmi (INRIA Lorraine - LORIA), Dominique Fohr (INRIA\n Lorraine - LORIA), Irina Illina (INRIA Lorraine - LORIA)	0711.0811	0.756685
7	Phoneme recognition in TIMIT with BLSTM-CTC	Santiago Fern\'andez, Alex Graves, Juergen Schmidhuber	0804.3269	0.753250
8	A biomechanical model of the face including muscles for the prediction\n of deformations during speech production	Julie Groleau (TIMC), Matthieu Chabanas (TIMC), Christophe Marecaux\n (TIMC), Natacha Payrard, Brice Segaud, Michel Rochette, Pascal Perrier (ICP),\n Yohan Payan (TIMC)	0803.3924	0.750757
9	Suppl\'eance perceptive par \'electro-stimulation linguale embarqu\'ee :\n perspectives pour la pr\'evention des escarres chez le bless\'e m\'edullaire	Olivier Chenu (TIMC), Nicolas Vuillerme (TIMC), Alexandre\n Moreau-Gaudry (TIMC), Anthony Fleury (TIMC), Jacques Demongeot (TIMC), Yohan\n Payan (TIMC)	0711.3786	0.746030

Our Hypothesis

Evaluation

- Use overlapping citations in target and recommended papers
- Count common citations
- Normalise scores with total number of citations
- Determine a threshold for correctness of result
- Calculate accuracy of the recommender
- Aim is to achieve better results than arxiv neural search

References

- Haruna, Khalid; Ismail, Maizatul Akmar; Qazi, Atika; Kakudi, Habeebah Adamu; Hassan, Mohammed; Muaz, Sanah Abdullahi; Chiroma, Haruna (2020).
- https://github.com/karndeb/Arxiv-Neural-Search
- DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations
- [2104.08821] SimCSE: Simple Contrastive Learning of Sentence Embeddings
- Cohan, Arman, et al. "Specter: Document-level representation learning using citation-informed transformers." arXiv preprint arXiv:2004.07180 (2020).

Thoughts?
Questions?

Thank you.