An Alternative to FLOPS Regularization to Effectively Productionize SPLADE-Doc

ReNeuIR’25 Workshop @ SIGIR 2025
July 18, 2025

Presented by Aldo Porco, Research Scientist, AI Engineering Group

Aldo Porco♠, Dhruv Mehra♠, Igor Malioutov♠, Karthik Radhakrishnan♠, Moniba Keymanesh♠
Daniel Preotiuc-Pietro♠, Sean MacAvaney♣, and Pengxiang Cheng♠
♠Bloomberg AI ♣University of Glasgow

TechAtBloomberg.com

Engineering

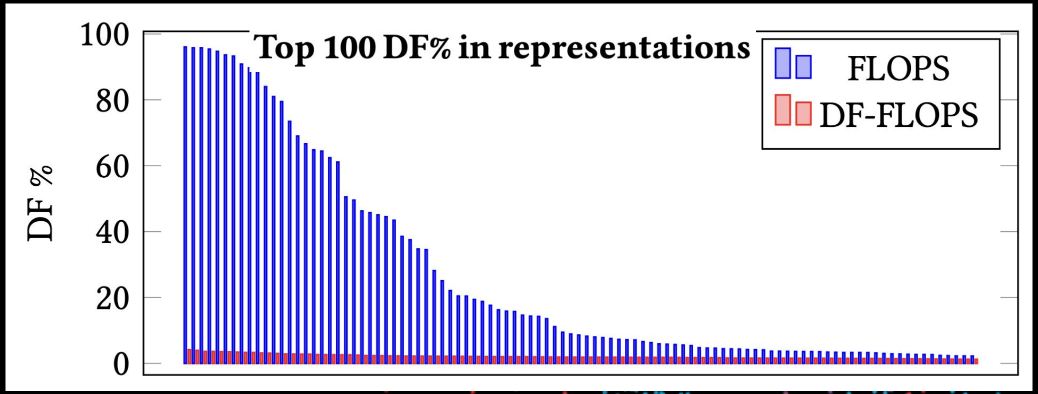
Bloomberg

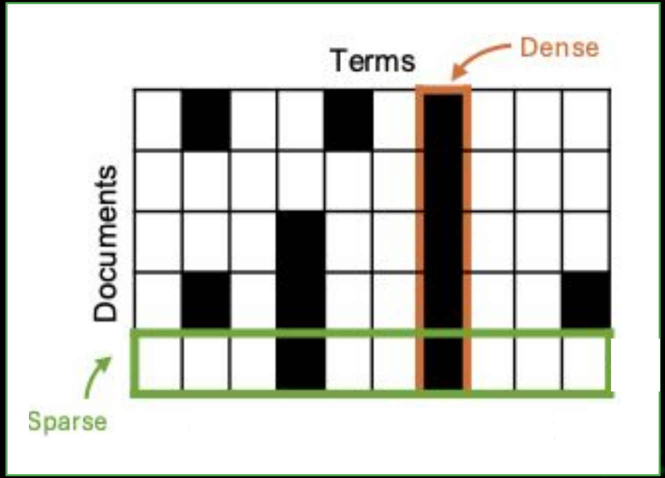
Agenda

- Problem
  - SPLADE-Doc architectures trained with FLOPS are reliant on high-frequency tokens
- Solution
  - DF-FLOPS as an alternative regularization method
- Results
- Takeaways

Problem

- Splade-Doc is an sparse retrieval architecture that achieves low latency by applying a bag-of-word encoding of the query
- FLOPS regularization keeps the outputs sparse, but reliant on high-frequency tokens
  - Large posting lists increase latency
  - BlockMaxWAND optimizations might not be possible in production systems
    — Requires complex functionality (like filtering)





Problem

Representations trained with FLOPS tend to rely on high-frequency tokens

- Having large posting lists, and thus high latency
- Production systems need functionality (like filtering) that prevents using BlockMaxWAND optimizations

FLOPS Formulation

$$\ell_{FLOPS} = \sum_{t \in V} \left(\frac{1}{N} \sum_{i=1}^{N} r_{i,t} \right)^2$$

Solution

DF-FLOPS Formulation

$$\ell_{DF-FLOPS} = \sum_{t \in V} \left(\frac{w_t}{N} \sum_{i=1}^{N} r_{i,t} \right)^2 \quad \text{where} \quad w_t = activ\left(\frac{DF_t}{|C|} \right)$$

FLOPS Formulation

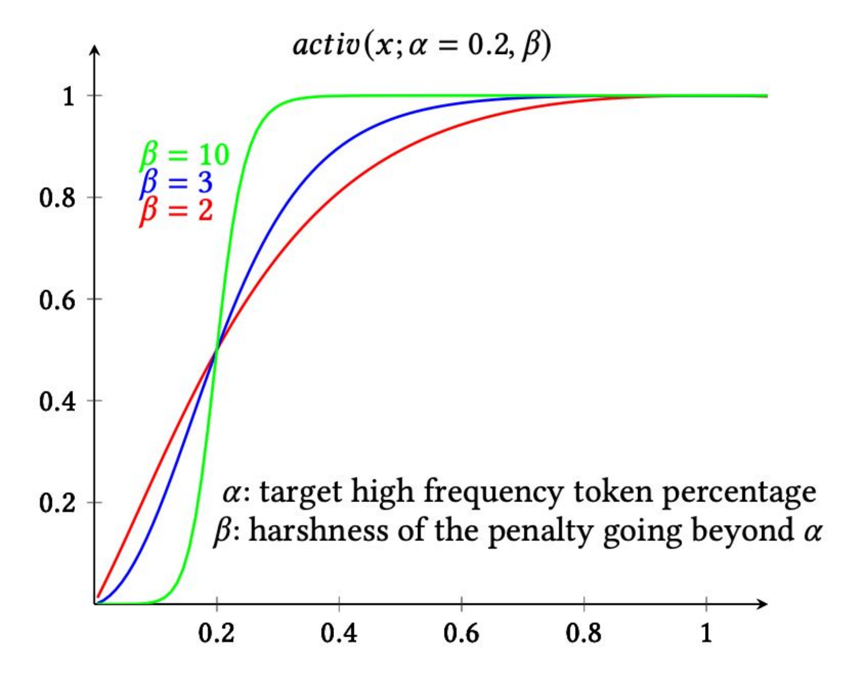
$$\ell_{FLOPS} = \sum_{t \in V} \left(\frac{1}{N} \sum_{i=1}^{N} r_{i,t} \right)^2$$

Solution

DF-FLOPS Formulation

$$\ell_{DF-FLOPS} = \sum_{t \in V} \left(\frac{w_t}{N} \sum_{i=1}^{N} r_{i,t} \right)^2 \quad \text{where} \quad w_t = activ\left(\frac{DF_t}{|C|} \right)$$

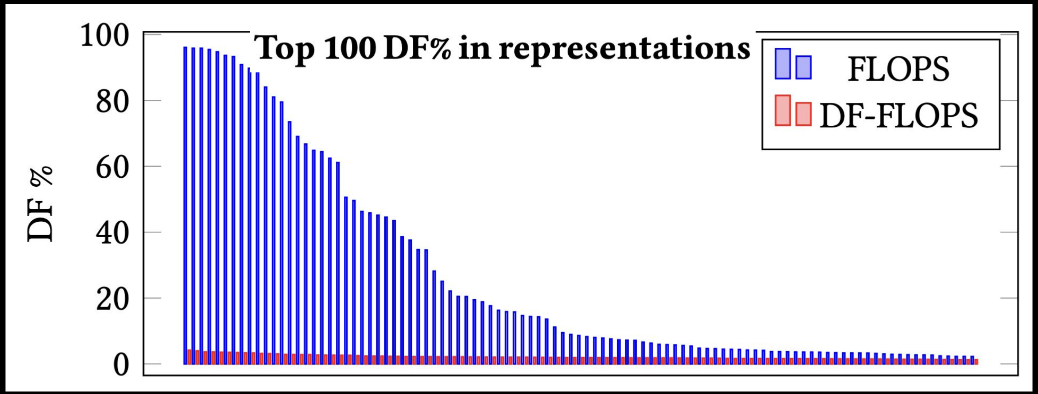
$$activ(x; \alpha, \beta) = \frac{1}{1 + (x^{log_\alpha 2} - 1)^\beta}$$

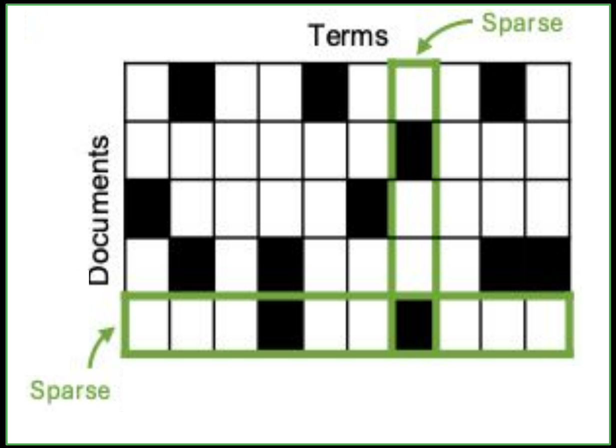


Solution

DF-FLOPS is able to:

- Reduce posting lists and latency
- Surface contextually relevant stop words





Example

***FLOPS** document representation (Top 20 weights)*

Document: Estimates of disease burden and ~~cost~~ – effectiveness ~~.~~ WHO ~~/~~ C ~~.~~ Nelson ~~.~~ Disease burden is an indicator of health ~~outcome~~ ~~.~~ disease burden can ~~be~~ expressed in ~~many ways~~ ~~,~~ ~~such as~~ the ~~number~~ of ~~cases~~ ~~(~~ ~~e .~~ ~~g .~~ ~~incidence~~ ~~or~~ ~~prevalence~~ ~~)~~ ~~,~~ ~~deaths~~ ~~or~~ ~~disability~~ – ~~adjusted~~ life ~~years lost~~ ~~(DALY s)~~ ~~associated~~ ~~with~~ a ~~given condition~~ ~~.~~ Information ~~on~~ the ~~reported~~ ~~incidence~~ of ~~vaccine~~ – ~~prevent~~ ~~able~~ diseases is ~~provided~~ to WHO ~~.~~

Expansions: ('what', 6.18), ('is', 5.5), ('of', 5.1), ('the', 4.87), ('a', 4.83), ('who', 4.81), ('does', 4.66), ('in', 4.61), ('are', 4.4), ('an', 4.39), ('for', 4.19), ('how', 4.13), ('?', 4.02), ('disease', 3.89), ('burden', 3.72), ('diseases', 3.54), ('definition', 3.51), ('health', 3.49), ('and', 3.41), ('mean', 3.4)

***DF-FLOPS** document representation (Top 20 weights)*

Document: Estimates ~~of~~ disease burden ~~and~~ cost – effectiveness ~~.~~ WHO ~~/~~ ~~C~~ ~~.~~ Nelson ~~.~~ Disease burden ~~is an~~ indicator ~~of~~ health outcome ~~.~~ disease burden ~~can be~~ expressed ~~in many ways~~ ~~,~~ ~~such as the number of~~ cases ~~(~~ ~~e .~~ ~~g .~~ ~~incidence or~~ prevalence ~~)~~ ~~,~~ deaths ~~or~~ disability – ~~adjusted life years lost~~ ~~(~~ DALY ~~s)~~ ~~associated with a given condition~~ ~~.~~ ~~Information on the reported~~ ~~incidence of~~ vaccine – ~~prevent~~ able diseases ~~is provided to~~ WHO ~~.~~

Expansions: ('burden', 3.59), ('disease', 3.44), ('estimates', 3.21), ('effectiveness', 3.12), ('diseases', 3.07), ('estimate', 2.98), ('indicator', 2.93), ('indicators', 2.82), ('nelson', 2.8), ('estimation', 2.67), ('load', 2.62), ('health', 2.55), ('vaccine', 2.53), ('illness', 2.52), ('DALY', 2.52), ('estimated', 2.51), ('who', 2.44), ('weight', 2.43), ('cost', 2.42), ('effective', 2.36)

~~Excluded~~ Stop Words Content Words

Results

ID	Model Name	MS-Marco MRR@10	R@1K	Latency Avg (ms)	Latency P99 (ms)	Matches Avg (M)	Top@1 Token DF	Avg. Emb. Length
1	BM25	18.4*	85.3*	68.9	241.3	0.952	20.6%	27.7
2	SPLADE-Doc w/ **FLOPS**	**32.2**	92.4	922.0	1945.6	8.628	95.8%	583.8
3	+ *Pruning@150*	32.0	92.1	792.1	1664.4	8.621	95.7%	147.6
4	+ $\uparrow \lambda = 0.1$	29.2	88.8	331.6	708.8	4.111	43.4%	87.6
5	+ $\uparrow \lambda = 1$	28.3	88.4	160.9	347.0	1.970	17.7%	**33.0**
6	SPLADE-Doc w/ **DF-FLOPS**	30.0	92.9	161.0	341.7	1.907	8.0%	301.6
7	+ *Pruning@150*	29.7	**93.0**	**87.8**	**187.8**	**1.078**	**5.2%**	140.3

Increasing **FLOPS** regularization improves latency, but at the cost of twice as much decrease in MRR

TechAtBloomberg.com

Bloomberg

Engineering

Results

ID	Model Name	MS-Marco MRR@10	MS-Marco R@1K	Latency Avg (ms)	Latency P99 (ms)	Matches Avg (M)	Top@1 Token DF	Avg. Emb. Length
1	BM25	18.4*	85.3*	68.9	241.3	0.952	20.6%	27.7
2	SPLADE-Doc w/ **FLOPS**	**32.2**	92.4	922.0	1945.6	8.628	95.8%	583.8
3	+ *Pruning@150*	32.0	92.1	792.1	1664.4	8.621	95.7%	147.6
4	+ $\uparrow\lambda = 0.1$	29.2	88.8	331.6	708.8	4.111	43.4%	87.6
5	+ $\uparrow\lambda = 1$	28.3	88.4	160.9	347.0	1.970	17.7%	**33.0**
6	SPLADE-Doc w/ **DF-FLOPS**	30.0	92.9	161.0	341.7	1.907	8.0%	301.6
7	+ *Pruning@150*	29.7	**93.0**	**87.8**	**187.8**	**1.078**	**5.2%**	140.3

With no decrease in Recall and ONLY a two point decrease in MRR, DF-FLOPS improves the average latency ~10x

TechAtBloomberg.com

Bloomberg

Engineering

Results

ID	Model Name	MS-Marco MRR@10	R@1K	Latency Avg (ms)	Latency P99 (ms)	Matches Avg (M)	Top@1 Token DF	Avg. Emb. Length
1	BM25	18.4*	85.3*	68.9	241.3	0.952	20.6%	27.7
2	SPLADE-Doc w/ **FLOPS**	**32.2**	92.4	922.0	1945.6	8.628	95.8%	583.8
3	+ *Pruning@150*	32.0	92.1	792.1	1664.4	8.621	95.7%	147.6
4	+ $\uparrow \lambda = 0.1$	29.2	88.8	331.6	708.8	4.111	43.4%	87.6
5	+ $\uparrow \lambda = 1$	28.3	88.4	160.9	347.0	1.970	17.7%	**33.0**
6	SPLADE-Doc w/ **DF-FLOPS**	30.0	92.9	161.0	341.7	1.907	8.0%	301.6
7	+ *Pruning@150*	29.7	**93.0**	**87.8**	**187.8**	**1.078**	**5.2%**	140.3

FLOPS trained models depend on high-frequency tokens even under heavy regularization settings

Results

ID	Model Name	MS-Marco MRR@10	MS-Marco R@1K	Latency Avg (ms)	Latency P99 (ms)	Matches Avg (M)	Top@1 Token DF	Avg. Emb. Length
1	BM25	18.4*	85.3*	68.9	241.3	0.952	20.6%	27.7
2	SPLADE-Doc w/ **FLOPS**	**32.2**	92.4	922.0	1945.6	8.628	95.8%	583.8
3	+ *Pruning@150*	32.0	92.1	792.1	1664.4	8.621	95.7%	147.6
4	+ $\uparrow \lambda = 0.1$	29.2	88.8	331.6	708.8	4.111	43.4%	87.6
5	+ $\uparrow \lambda = 1$	28.3	88.4	160.9	347.0	1.970	17.7%	**33.0**
6	SPLADE-Doc w/ **DF-FLOPS**	30.0	92.9	161.0	341.7	1.907	8.0%	301.6
7	+ *Pruning@150*	29.7	**93.0**	**87.8**	**187.8**	**1.078**	**5.2%**	140.3

DF-FLOPS is able to effectively reduce reliance on high-frequency tokens (sparser representations)

TechAtBloomberg.com

Bloomberg Engineering

Results

ID	Model Name	MS-Marco MRR@10	R@1K	Latency Avg (ms)	Latency P99 (ms)	Matches Avg (M)	Top@1 Token DF	Avg. Emb. Length
1	BM25	18.4*	85.3*	68.9	241.3	0.952	20.6%	27.7
2	SPLADE-Doc w/ **FLOPS**	**32.2**	92.4	922.0	1945.6	8.628	95.8%	583.8
3	+ *Pruning@150*	32.0	92.1	792.1	1664.4	8.621	95.7%	147.6
4	+ $\uparrow \lambda = 0.1$	29.2	88.8	331.6	708.8	4.111	43.4%	87.6
5	+ $\uparrow \lambda = 1$	28.3	88.4	160.9	347.0	1.970	17.7%	**33.0**
6	SPLADE-Doc w/ **DF-FLOPS**	30.0	92.9	161.0	341.7	1.907	8.0%	301.6
7	+ *Pruning@150*	29.7	**93.0**	**87.8**	**187.8**	**1.078**	**5.2%**	140.3

DF-FLOPS obtains comparable mean (and better P99) latency w.r.t BM25,
without further performance degradation

TechAtBloomberg.com

Bloomberg
Engineering

Results

BEIR Dataset	BM25	FLOPS Base	FLOPS $\lambda = 1$	DF-FLOPS Base
arguana	31.5*	11.16	28.83	33.25
climate-fever	**21.3***	6.89	11.96	13.44
dbpedia-entity	27.3*	31.21	30.55	**32.73**
fever	**75.3***	57.67	60.49	63.12
fiqa	23.6*	19.64	21.08	**25.56**
hotpotqa	**60.3***	42.08	48.59	55.34
nfcorpus	**32.5***	29.74	30.09	30.59
nq	32.9*	**39.82**	34.24	39.05
quora	**78.9***	7.56	12.39	48.1
scidocs	**15.8***	12.67	13.49	13.79
scifact	**66.5***	58.75	60.87	65.6
trec-covid	**65.6***	56.17	51.78	57.52
webis-touche2020	**36.7***	23.28	21.45	25.97

DF-FLOPS outperforms FLOPS in 12/13 BIER datasets; however, BM25 is better in most of them

TechAtBloomberg.com

Bloomberg Engineering

Takeaways

- SPLADE-Doc models trained with FLOPS tend to rely on high-frequency tokens, which impacts latency
- DF-FLOPS is the proposed regularization method for training SPLADE-Doc models that penalizes high-frequency tokens and promotes sparsity
- DF-FLOPS can reduce latency ~10x compared to the SPLADE-v2-Doc-max baseline

Thank You!

For more information about our work, research, and academic outreach programs, please visit:

https://www.TechAtBloomberg.com/ai/

TechAtBloomberg.com