ICLR

SOLAR: Sparse Orthogonal Learned And Random Embeddings

Tharun Medini, Beidi Chen, Anshumali Shrivastava

tharun.medini@rice.edu

3rd May 2021

Main Contribution

- SOLAR is a high-dimensional super-sparse embedding learning method for Information Retrieval
- Significantly superior on both **query latency** and **prediction accuracy** than dense low-dimensional embedding models

DSSM (Dense Embedding Model)

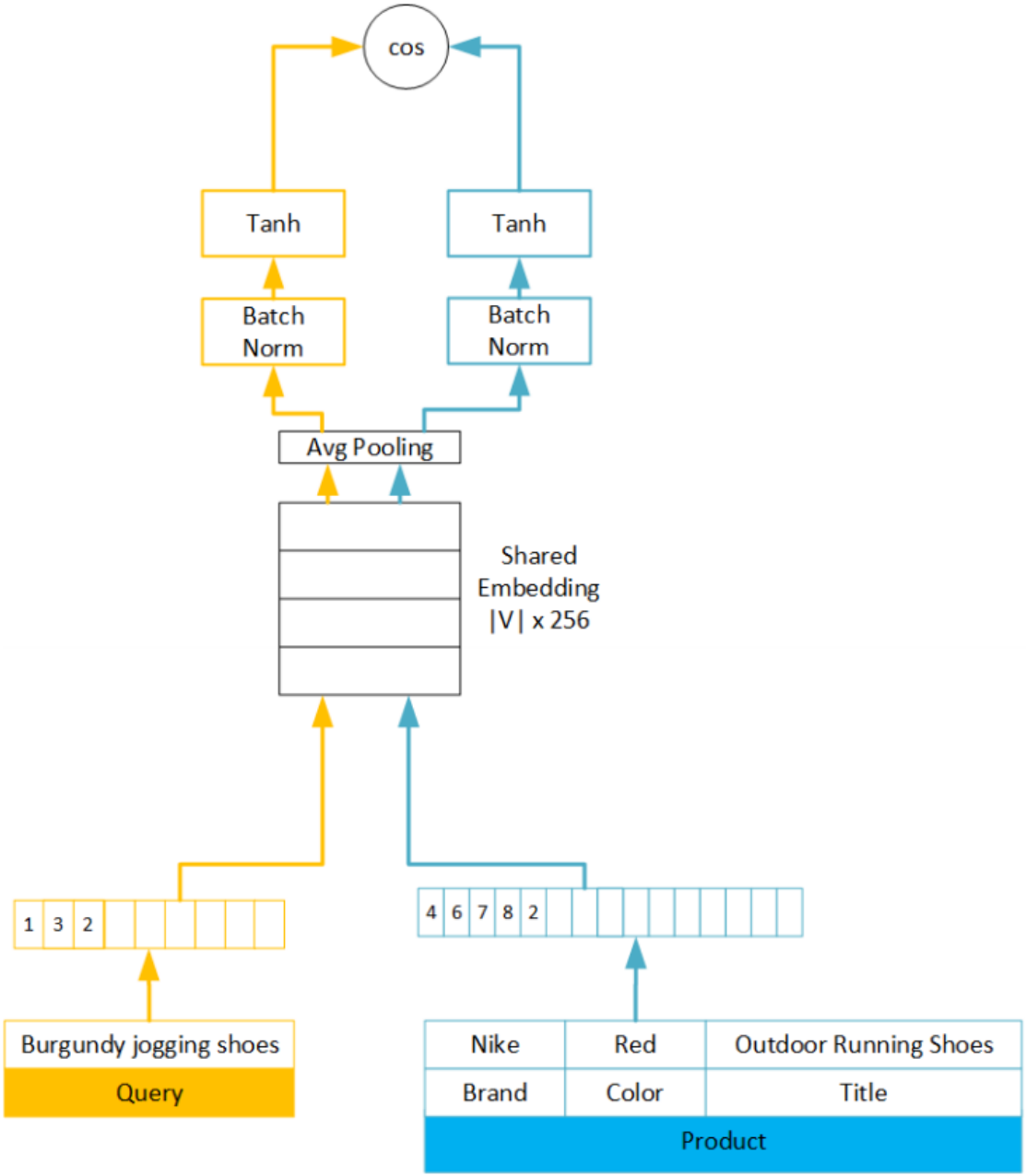


Figure credits:
Semantic Product Search,
Nigam et. al.,
KDD 2019

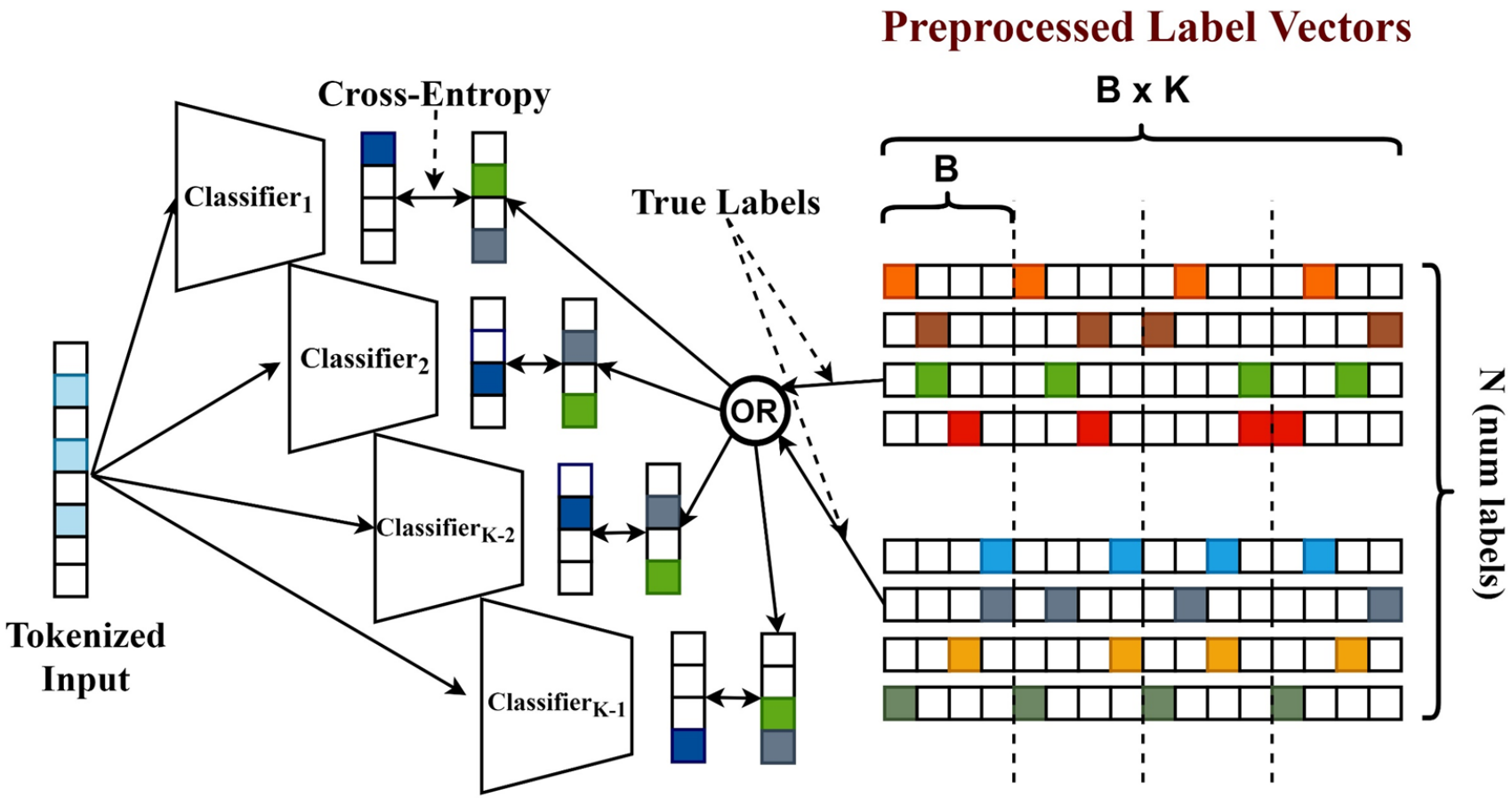
Unique Sparsity Design

- By design, all vectors have **exactly K non-zeros** and **spread** them out **uniformly** across the dimensions
- We **fix** the label vectors and only train the query vectors
- Above steps ensure persistent load-balance and orthogonality
- Expected dot-product between any two vectors is **~0**

4-fold advantage

- Matrix multiplications are replaced by Inverted-Index Lookups
- Load-Balanced Inverted Indexes
- Low embedding memory
- Trivial zero-communication distributed training

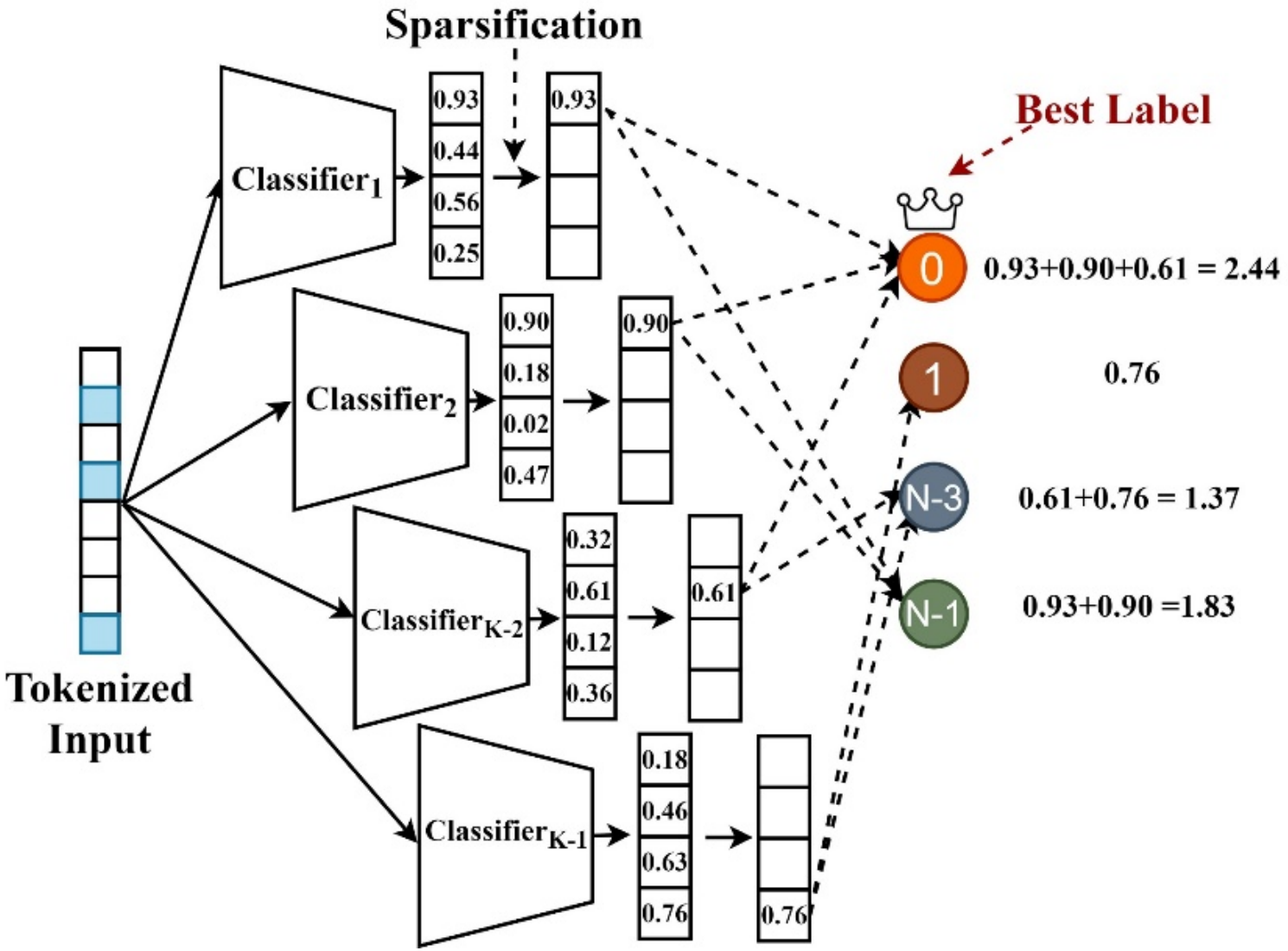
Training


Preprocessed Label Vectors
B x K
B
Cross-Entropy
True Labels
Classifier1
Classifier2
ClassifierK-2
ClassifierK-1
OR
Tokenized Input
N (num labels)

Inverted Index

- We construct one index for each of the **K** chunks.
- Each bucket will have the same number of labels by design (Load-Balanced)

Inference


Sparsification
Best Label
Classifier1
Classifier2
ClassifierK-2
ClassifierK-1
0.93
0.44
0.56
0.25
0.90
0.18
0.02
0.47
0.32
0.61
0.12
0.36
0.18
0.46
0.63
0.76
Tokenized Input
0
1
N-3
N-1
0.93+0.90+0.61 = 2.44
0.76
0.61+0.76 = 1.37
0.93+0.90 =1.83

Product-to-product Recommendation

Model	epochs	P@1	P@5	P@10	Rec@ 100	Train time (hrs)	Eval time (ms/point)
SOLAR (m=100)	10	35.24	29.71	26.98	34.19	2.65	0.96
DSSM (d=1600)	5	31.34	27.55	24.41	32.71	25.27	1.77
GLaS (d=1600)	5	32.51	28.31	25.41	33.17	37.14	1.77
SNRM (d=30K)	5	1.59	2.01	1.93	2.41	-	-
AnnexML (d=800)	10	26.31	22.22	19.37	26.13	16	3.06

Extreme Classification

Dataset	Metric	SOLAR (m=100)	SOLAR (m=50)	Annex ML	SLEEC	FyC	Parabel	PfastreXML	SLICE
	P@1	60.92	60.52	56.81	30.86	46.86	59.34	55	59.89
Wiki-500K	P@3	46.94	45.56	36.78	20.77	31.29	39.05	36.14	39.89
	P@5	45.32	45.28	27.45	15.23	25.17	29.35	27.38	30.12
	P@1	34.37	34.19	26.36	18.77	24.47	33.93	28.51	37.77
Amz-670K	P@3	32.71	32.51	22.94	16.5	20.44	30.38	26.06	33.76
	P@5	32.55	32.46	20.59	14.97	17.13	27.49	24.17	30.7
	P@1	44.89	44.61	41.79	-	-	**47.51**	43.83	-
Amz-3M	P@3	42.36	42.08	38.24	-	-	**44.68**	41.81	-
	P@5	41.03	40.69	35.98	-	-	**42.58**	40.09	-

Extreme Classification

Dataset		SOLAR (m=100)	SOLAR (m=50)	SLICE	Parabel	PfastreXML
Wiki-500K	Training time (hrs)	2.52	2.52	**2.34**	6.29	11.14
	Eval (ms/point)	1.1	**0.76**	1.37	2.94	6.36
Amz-670K	Training time (hrs)	**1.19**	**1.19**	1.92	1.84	2.85
	Eval (ms/point)	2.56	**1.58**	3.49	2.85	19.35
Amz-3M	Training time (hrs)	5.73	5.73	-	**5.39**	15.74
	Eval (ms/point)	2.09	1.87	-	**1.72**	4.05

Thank you!

- Paper

- Poster

- Please reach out to tharun.medini@rice.edu for further questions