Investigating the Effects of Sparse Attention on Cross-Encoders

ECIR 2024, 24–28 March, Glasgow, Scotland

Ferdinand Schlatt

Maik Fröbe

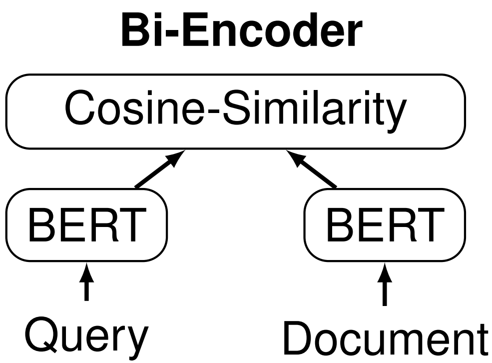
Matthias Hagen

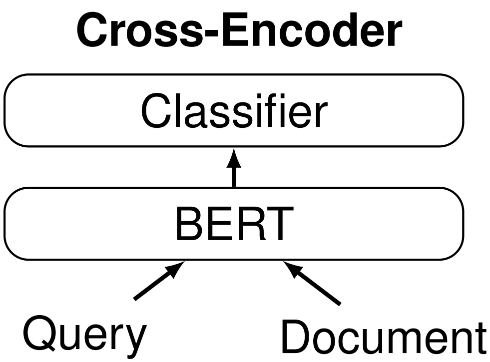
FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

webis.de

Investigating the Effects of Sparse Attention on Cross-Encoders

Motivation

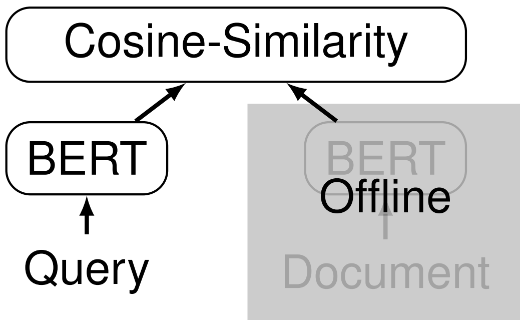




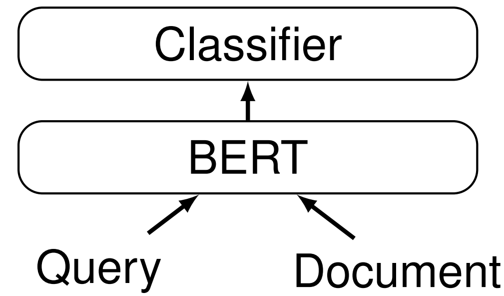
Investigating the Effects of Sparse Attention on Cross-Encoders

Motivation

Bi-Encoder



Cross-Encoder

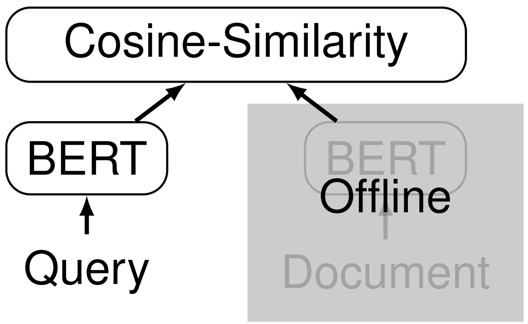


- ❑ Query and document processed separately
- \+ Documents can be processed and indexed offline
- \- No interaction between queries and documents

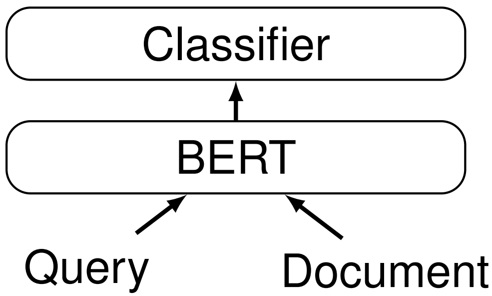
Investigating the Effects of Sparse Attention on Cross-Encoders

Motivation

Bi-Encoder



Cross-Encoder



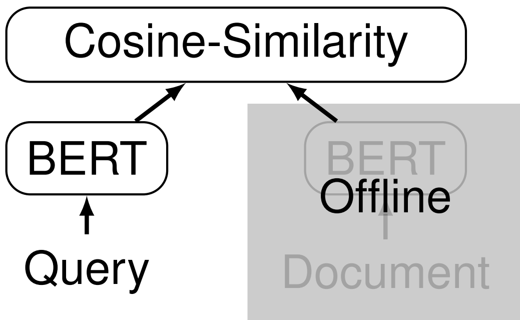
- ❑ Query and document processed separately
- \+ Documents can be processed and indexed offline
- \- No interaction between queries and documents

- ❑ Query and document processed jointly
- \- Document must be processed at query time
- \+ Interaction between queries and documents

Investigating the Effects of Sparse Attention on Cross-Encoders

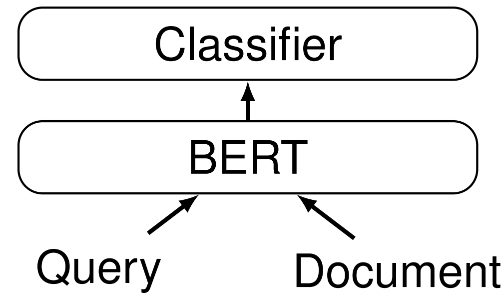
Motivation

Bi-Encoder



- ❑ Query and document processed separately
- \+ Documents can be processed and indexed offline
- \- No interaction between queries and documents

Cross-Encoder



- ❑ Query and document processed jointly
- \- Document must be processed at query time
- \+ Interaction between queries and documents

Cross-encoders are effective but slow and expensive to run. [Scells et al., SIGIR'22]

Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

Query:	python
Document:	Python is a great programming language to learn.

Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

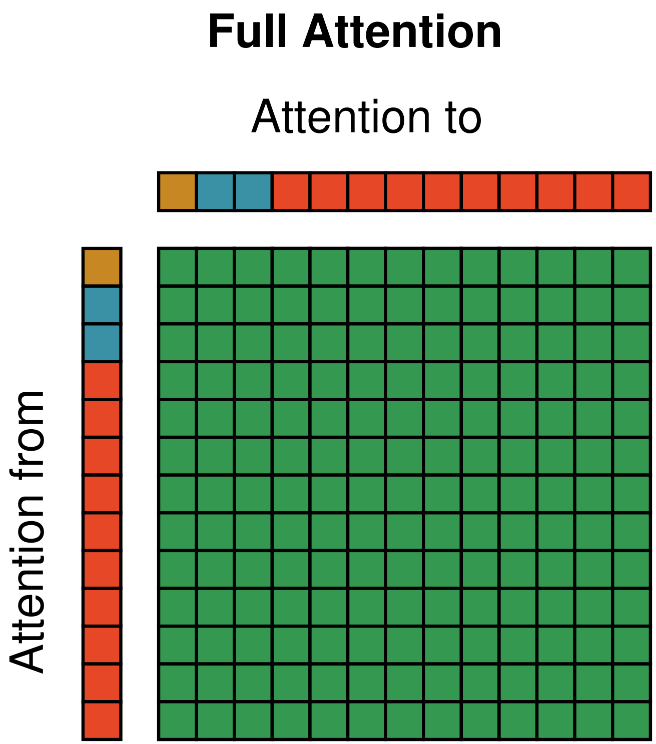
[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]



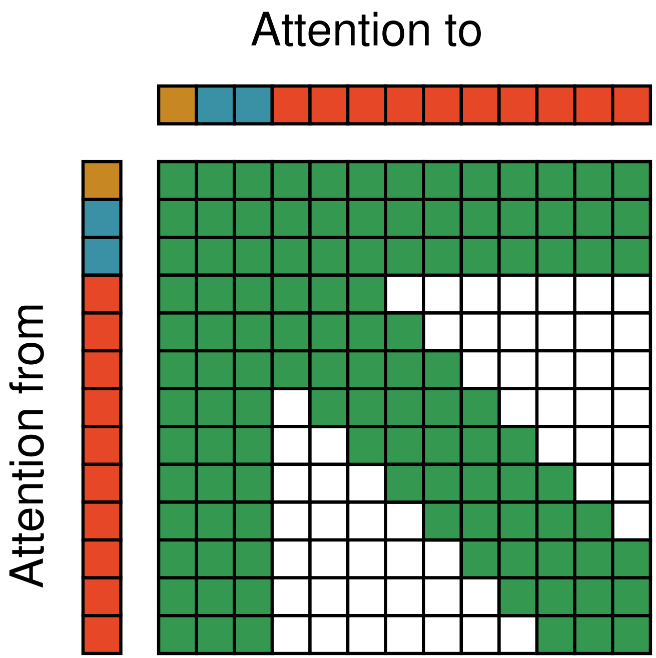
Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Longformer [Beltagy et al., arXiv'20]



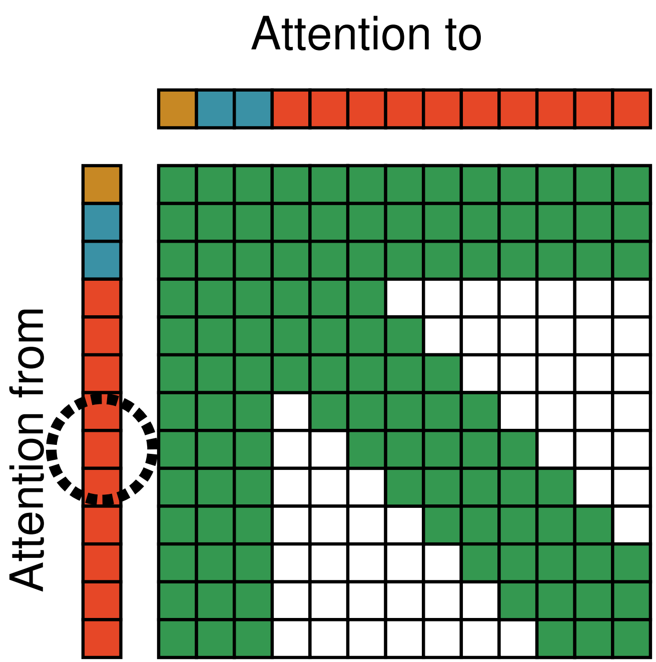
Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Longformer [Beltagy et al., arXiv'20]



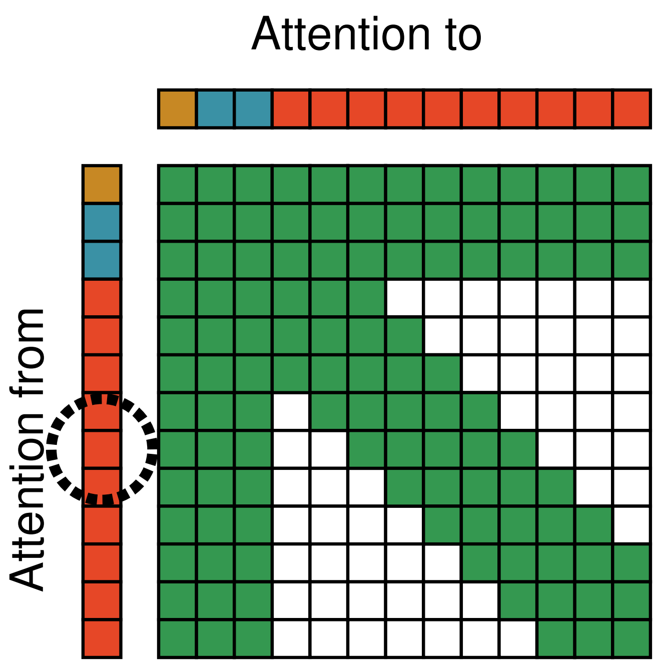
Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Longformer [Beltagy et al., arXiv'20]



- ❑ Document tokens' attention restricted to context window of length w

➜ Semantic "gist" suffices to determine the relevance of a document token

- ❑ Previous work used $w = 64$ to save memory and re-rank longer documents

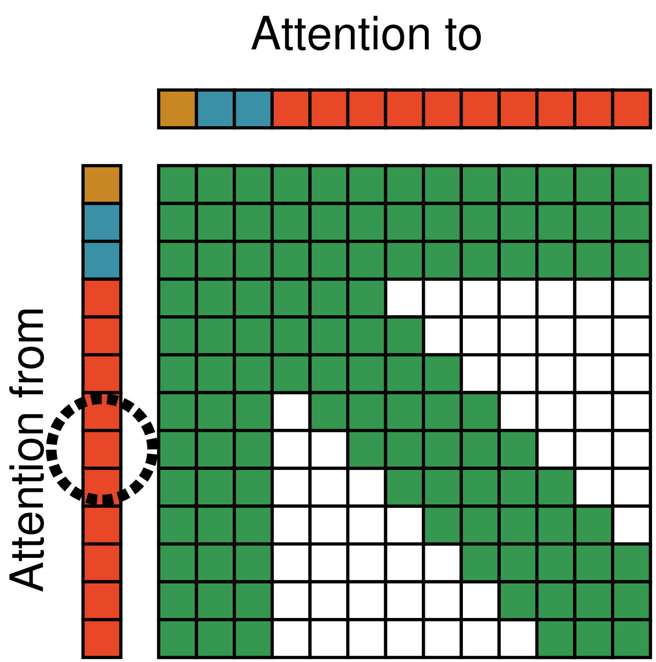
Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Longformer [Beltagy et al., arXiv'20]



- ❑ Document tokens' attention restricted to context window of length w

➜ Semantic "gist" suffices to determine the relevance of a document token

- ❑ Previous work used $w = 64$ to save memory and re-rank longer documents

Hypothesis: Very small window sizes are as effective as full attention.

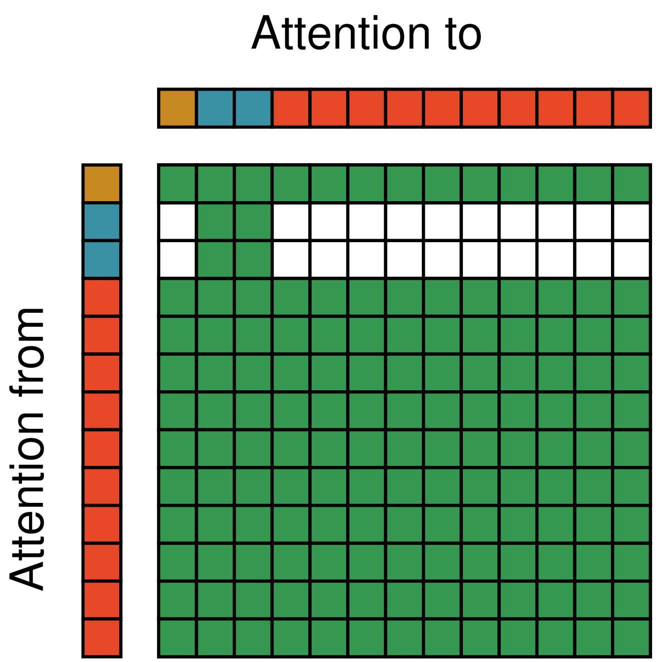
Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Query Independent Attention



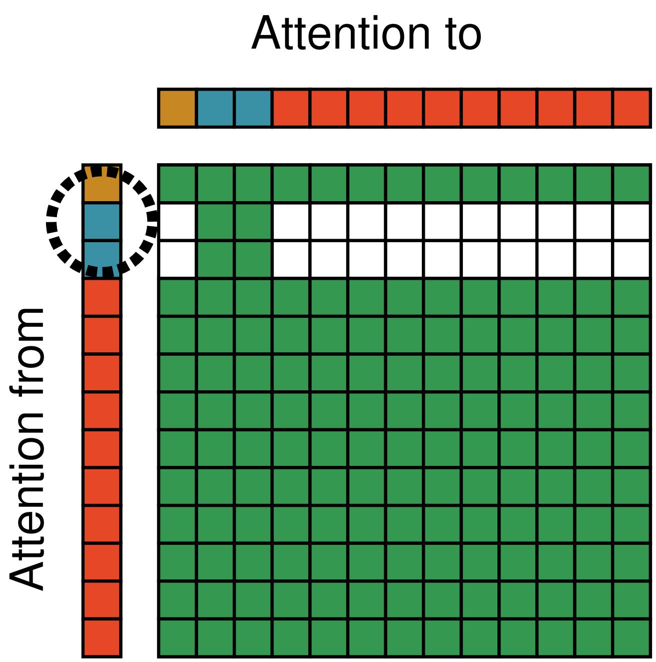
Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Query Independent Attention



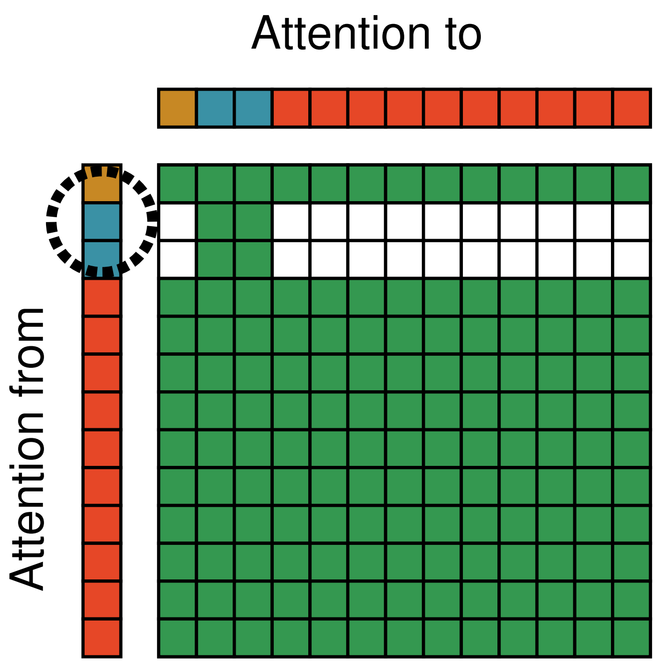
Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Query Independent Attention



- ❑ A document is relevant to a query and not vice versa

➜ The query–document relevance relationship is asymmetric

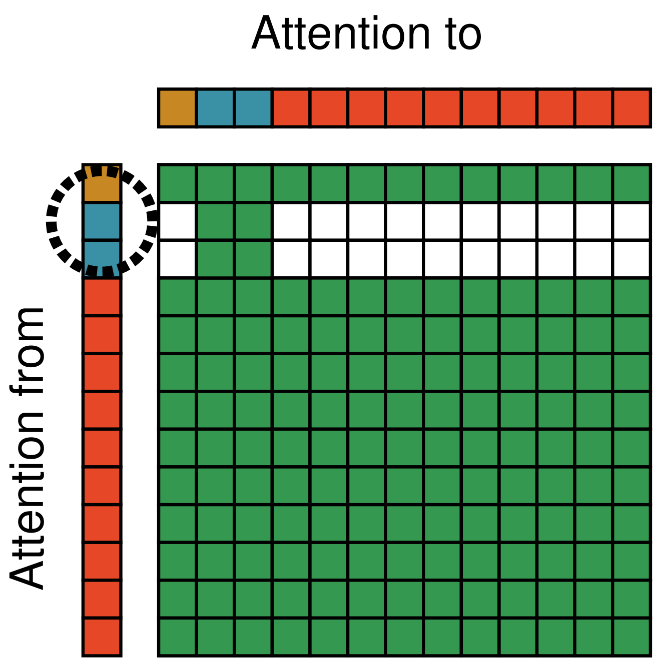
Investigating the Effects of Sparse Attention on Cross-Encoders

Making cross-encoders more efficient

One paradigm that improves cross-encoder efficiency is reducing the number of tokens that interact with each other. [Sekulic et al., TREC'20; Jiang et al., EMNLP'20]

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Query Independent Attention



- A document is relevant to a query and not vice versa

➜ The query–document relevance relationship is asymmetric

Hypothesis: Deactivating attention from query tokens to other tokens is as effective as full attention.

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder architecture

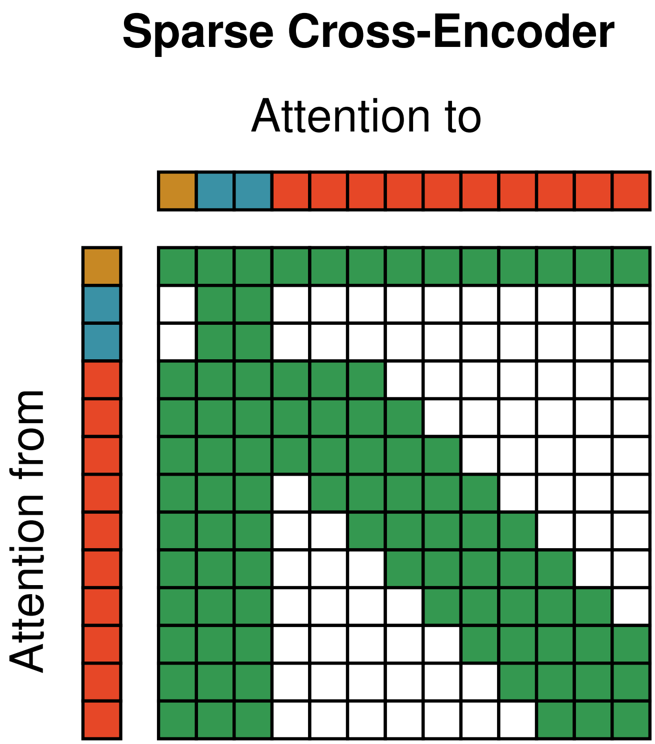
Our sparse cross-encoder architecture combines windowed self-attention and asymmetric cross-attention between sub-sequences.

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder architecture

Our sparse cross-encoder architecture combines windowed self-attention and asymmetric cross-attention between sub-sequences.

[CLS] python [SEP] Python is a great programming language to learn . [SEP]



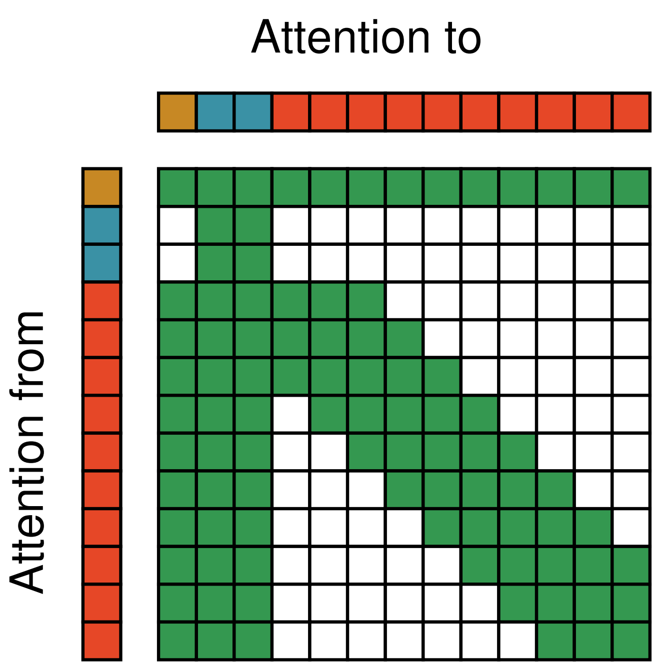
Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder architecture

Our sparse cross-encoder architecture combines windowed self-attention and asymmetric cross-attention between sub-sequences.

[CLS] python [SEP] Python is a great programming language to learn . [SEP]

Sparse Cross-Encoder



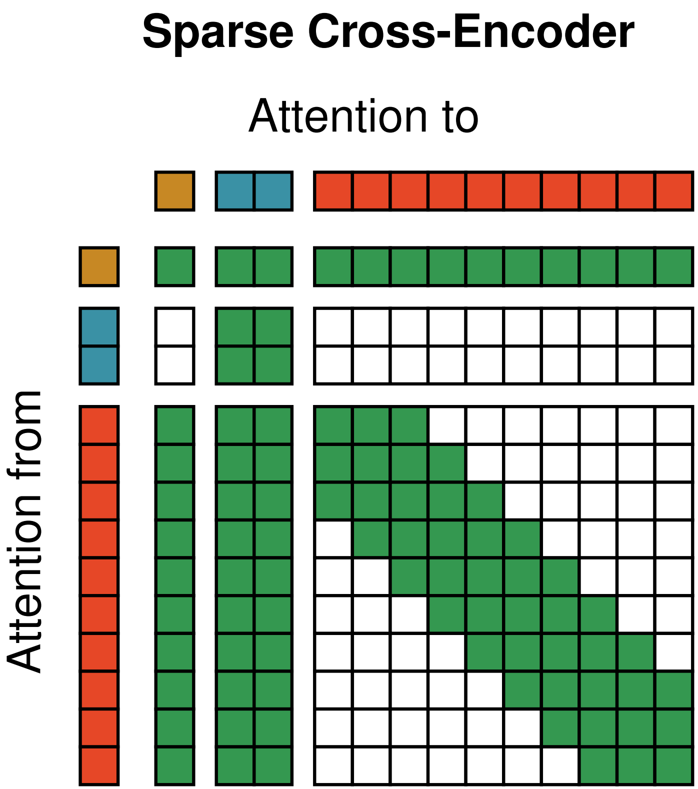
- Asymmetric attention not supported by standard transformer architectures

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder architecture

Our sparse cross-encoder architecture combines windowed self-attention and asymmetric cross-attention between sub-sequences.

[CLS] python [SEP] Python is a great programming language to learn . [SEP]



- ❑ Asymmetric attention not supported by standard transformer architectures

➜ Custom architecture with cross-attention between sub-sequences

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder effectiveness

nDCG@10 on TREC Deep Learning 2019–2022 passage and document

Task	Full Attention / Longformer						Sparse Cross-Encoder					
$w =$	∞	64	16	4	1	0	∞	64	16	4	1	0
Passage	0.62	0.62[†]										
Document	0.58	0.58										

[†] denotes significant equivalence within ± 0.02 (paired TOST) with underlined score per row. MaxP results are grayed out.

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder effectiveness

nDCG@10 on TREC Deep Learning 2019–2022 passage and document

Task	**Full Attention / Longformer**						**Sparse Cross-Encoder**					
$w =$	∞	64	16	4	1	0	∞	64	16	4	1	0
Passage	0.62	0.62†					0.62†					
Document	0.58	0.58					0.57					

† denotes significant equivalence within ± 0.02 (paired TOST) with underlined score per row. MaxP results are grayed out.

1. Asymmetric query attention does not impact effectiveness . . .

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder effectiveness

nDCG@10 on TREC Deep Learning 2019–2022 passage and document

Task	Full Attention / Longformer						Sparse Cross-Encoder					
$w =$	∞	64	16	4	1	0	∞	64	16	4	1	0
Passage	<u>0.62</u>	0.62†					0.62†	0.62†				
Document	0.58	<u>0.58</u>					0.57	0.59				

† denotes significant equivalence within ± 0.02 (paired TOST) with underlined score per row. MaxP results are grayed out.

1. Asymmetric query attention does not impact effectiveness even combined with windowed self-attention

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder effectiveness

nDCG@10 on TREC Deep Learning 2019–2022 passage and document

Task	Full Attention / Longformer						Sparse Cross-Encoder					
$w =$	∞	64	16	4	1	0	∞	64	16	4	1	0
Passage	0.62	0.62†	0.62†				0.62†	0.62†	0.61			
Document	0.58	0.58	0.59†				0.57	0.59	0.59			

† denotes significant equivalence within ± 0.02 (paired TOST) with underlined score per row. MaxP results are grayed out.

1. Asymmetric query attention does not impact effectiveness even combined with windowed self-attention

2. Window size of $w = 16$ is on par with full attention

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder effectiveness

nDCG@10 on TREC Deep Learning 2019–2022 passage and document

Task	Full Attention / Longformer						Sparse Cross-Encoder					
$w =$	∞	64	16	4	1	0	∞	64	16	4	1	0
Passage	<u>0.62</u>	0.62†	0.62†	0.62†			0.62†	0.62†	0.61	0.61†		
Document	0.58	<u>0.58</u>	0.59†	0.59			0.57	0.59	0.59	0.58		

† denotes significant equivalence within ± 0.02 (paired TOST) with underlined score per row. MaxP results are grayed out.

1. Asymmetric query attention does not impact effectiveness even combined with windowed self-attention

2. Window size of $\cancel{w = 16}$ $w = 4$ is on par with full attention

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder effectiveness

nDCG@10 on TREC Deep Learning 2019–2022 passage and document

Task	Full Attention / Longformer						Sparse Cross-Encoder					
$w =$	∞	64	16	4	1	0	∞	64	16	4	1	0
Passage	0.62	0.62†	0.62†	0.62†	0.61		0.62†	0.62†	0.61	0.61†	0.60	
Document	0.58	0.58	0.59†	0.59	0.58†		0.57	0.59	0.59	0.58	0.59	

† denotes significant equivalence within ± 0.02 (paired TOST) with underlined score per row. MaxP results are grayed out.

1. Asymmetric query attention does not impact effectiveness even combined with windowed self-attention
2. Window size of $\cancel{w = 16}$ $w = 4$ is on par with full attention
3. Window size of $w = 1$ still competitive

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder effectiveness

nDCG@10 on TREC Deep Learning 2019–2022 passage and document

Task	Full Attention / Longformer						Sparse Cross-Encoder					
$w =$	∞	64	16	4	1	0	∞	64	16	4	1	0
Passage	<u>0.62</u>	0.62†	0.62†	0.62†	0.61	0.57	0.62†	0.62†	0.61	0.61†	0.60	0.56
Document	0.58	<u>0.58</u>	0.59†	0.59	0.58†	0.56	0.57	0.59	0.59	0.58	0.59	0.56

† denotes significant equivalence within ± 0.02 (paired TOST) with underlined score per row. MaxP results are grayed out.

1. Asymmetric query attention does not impact effectiveness even combined with windowed self-attention
2. Window size of $\cancel{w = 16}$ $w = 4$ is on par with full attention
3. Window size of $w = 1$ still competitive
4. Window size of $w = 0$ slightly less effective

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder effectiveness

nDCG@10 on TREC Deep Learning 2019–2022 passage and document

Task	Full Attention / Longformer						Sparse Cross-Encoder					
$w =$	∞	64	16	4	1	0	∞	64	16	4	1	0
Passage	$\underline{0.62}$	0.62†	0.62†	0.62†	0.61	0.57	0.62†	0.62†	0.61	0.61†	0.60	0.56
Document	0.58	$\underline{0.58}$	0.59†	0.59	0.58†	0.56	0.57	0.59	0.59	0.58	0.59	0.56

† denotes significant equivalence within ± 0.02 (paired TOST) with underlined score per row. MaxP results are grayed out.

1. Asymmetric query attention does not impact effectiveness even combined with windowed self-attention
2. Window size of $\cancel{w = 16}$ $w = 4$ is on par with full attention
3. Window size of $w = 1$ still competitive
4. Window size of $w = 0$ slightly less effective

➜ Also translates to out-of-domain effectiveness on TIREx [Fröbe et al. SIGIR'23]

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder efficiency

Latency and memory consumption on synthetic query document pairs

Unit	Full Attention	Longformer	Sparse CE	Sparse CE
$w =$	∞	64	64	4
Query length 10, Passage length 164				
µs	368	980 (+166%)		
MB	9	15 (+67%)		
Query length 10, Document length 4086				
ms	49 (+250%)	14		
MB	1608 (+905%)	160		

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder efficiency

Latency and memory consumption on synthetic query document pairs

Unit	Full Attention	Longformer	Sparse CE	Sparse CE
$w =$	∞	64	64	4
Query length 10, Passage length 164				
µs	368	980 $(+166\%)$	527 $(+43\%)$	
MB	9	15 $(+67\%)$	9 $(+0\%)$	
Query length 10, Document length 4086				
ms	49 $(+250\%)$	14	12 (-14%)	
MB	1608 $(+905\%)$	160	111 (-31%)	

1. Sparse cross-encoder with $w = 64$ is more efficient than the Longformer

Investigating the Effects of Sparse Attention on Cross-Encoders

Sparse cross-encoder efficiency

Latency and memory consumption on synthetic query document pairs

Unit	**Full Attention**	**Longformer**	**Sparse CE**	**Sparse CE**
$w =$	∞	64	64	4
Query length 10, Passage length 164				
µs	368	980 $(+166\%)$	527 $(+43\%)$	364 (-1%)
MB	9	15 $(+67\%)$	9 $(+0\%)$	7 (-22%)
Query length 10, Document length 4086				
ms	49 $(+250\%)$	14	12 (-14%)	8 (-43%)
MB	1608 $(+905\%)$	160	111 (-31%)	66 (-59%)

1. Sparse cross-encoder with $w = 64$ is more efficient than the Longformer

2. Window size $w = 4$ is more efficient than full attention on passages

Investigating the Effects of Sparse Attention on Cross-Encoders

Conclusion

We introduced a sparse cross-encoder architecture that combines windowed self-attention and asymmetric cross-attention between sub-sequences.

- ❑ Attention from query tokens to other tokens can be deactivated without losing effectiveness.
- ❑ Very small window sizes are still effective for re-ranking with cross-encoders.
- ❑ Our sparse cross-encoder reduces memory consumption and runtime.

Conclusion

We introduced a sparse cross-encoder architecture that combines windowed self-attention and asymmetric cross-attention between sub-sequences.

- Attention from query tokens to other tokens can be deactivated without losing effectiveness.
- Very small window sizes are still effective for re-ranking with cross-encoders.
- Our sparse cross-encoder reduces memory consumption and runtime.

Thank you!

Code and models @
https://github.com/webis-de/ECIR-24

Investigating the Effects of Sparse Attention on Cross-Encoders

Full TREC DL Table

Task		Full Att. / Longformer						Sparse Cross-Encoder						QDS
	$w =$	∞	64	16	4	1	0	∞	64	16	4	1	0	64
Passage	2019	.724	.719†	.725†	.719	.714	.694	.722	.717	.724	**.728**	.715	.696	.720†
Passage	2020	.674	.681†	.680	**.684**	.676	.632	.666	.672	.661	.665	.649	.605	.682
Passage	2021	**.656**	.653	.650	.645	.629	.602	**.656**	.650	.639	.647	.625	.593	**.656**†
Passage	2022	**.496**	.494†	.487	.486	.481	.441	.490	.492†	.479	.484	.471	.427	.495†
Passage	Avg.	.619	.619†	.616†	.615†	.607	.572	.615†	.615†	.607	.612†	.596	.560	**.620**†
Document	2019	.658	.683	.678	.667	.689	.663	.638	.672	.685	.669	.692	.646	**.697**
Document	2020	.622	.640	.639	**.661**	.655	.644	.636	.638	.650	.642	.657	.638	.639
Document	2021	.678	.671	.681	**.683**	**.683**	.629	.677	.681	.681	.670	.679	.644	.676
Document	2022	.424	.425	.431	.425	.409	.389	.421	**.446**	.443	.417	.424	.405	.428
Document	Avg.	.575	.582	.586†	.587	.584†	.556	.573	.590	**.594**	.577	.589	.561	.587†

Investigating the Effects of Sparse Attention on Cross-Encoders

TIREx Table

Corpus	Doc. Len.	First Stage	monoT5			monoBERT		Sparse CE	
			Base	Large	3b	Base	Large	512	4096
Antique	49.9	.510	.505	.527	.537	.507	.484	**.540**	.174
Args.me	435.5	**.405**	.305	.338	.392	.314	.371	.313	.180
CW09	1132.6	.178	.186	.182	.201	.192	.134	.198	**.212**
CW12	5641.7	**.364**	.260	.266	.279	.263	.251	.312	.338
CORD-19	3647.7	.586	.688	.636	.603	**.690**	.625	.673	.642
Cranfield	234.8	.008	.006	.007	.007	.006	.006	**.009**	.003
Disks4+5	749.3	.429	.516	.548	**.555**	.514	.494	.487	.293
GOV	2700.5	.266	.320	.327	**.351**	.318	.292	.316	.292
GOV2	2410.3	.467	.486	.513	**.514**	.489	.474	.503	.460
MED.	309.1	**.366**	.264	.318	.350	.267	.298	.237	.180
NFCorpus	364.6	.268	.295	.296	**.308**	.295	.288	.284	.151
Vaswani	51.3	.447	.306	.414	.458	.321	**.476**	.436	.163
WaPo	713.0	.364	.451	**.492**	.476	.449	.438	.434	.296
Average	–	.358	.353	.374	**.387**	.356	.356	.365	.260

Investigating the Effects of Sparse Attention on Cross-Encoders

Efficiency Graphs