CoSAEmb: Contrastive Section-aware Aspect Embeddings for Scientific Articles

Shruti Singh and **Mayank Singh**
{singh_shruti, singh.mayank}@iitgn.ac.in
LINGO, Indian Institute of Technology Gandhinagar
India

Abstract

Research papers are long documents that contain information about various aspects such as background, prior work, methodology, and results. Existing works on scientific document representation learning only leverage the title and abstract of the paper. We present CoSAEmb, a model that learns representations from the full-text of 97402 scientific papers from the S2ORC dataset. We present a novel supervised contrastive training framework for long documents using triplet loss and margin gradation. Our framework can be used to learn representations of long documents with any existing encoder-only transformer model without retraining it from scratch. CoSAEmb shows improved performance on information retrieval from the paper's full-text in comparison to models trained only on paper titles and abstracts. We also evaluate CoSAEmb on SciRepEval and CSFCube benchmarks, showing comparable performance with existing state-of-the-art models.

1 Introduction

Scientific papers contain dense detailed information in the paper text. However, existing works on scientific document representation learning are restricted to either the title and abstract of the paper or encode sentence-level information. SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2020) are the earliest transformer-based representation learning models for scientific texts, both of which use the BERT model as base architecture. Both of these models operate at the sentence level. Models such as SPECTER (Cohan et al., 2020), OAG-BERT (Liu et al., 2022), Siamese-SciBERT (Ostendorff et al., 2022a), and SciNCL (Ostendorff et al., 2022b) encode the title and the abstract of the paper. Mysore et al. (2022) learn multiple aspect-specific vectors of research papers at the sentence level. We posit that the title and the abstract embeddings or sentence vectors fail to capture the intricate details of the paper. In contrast to existing methods, our training strategy (described in Section 3) takes into consideration the full-text of the paper in a contrastive learning setup. We propose to learn section embeddings for a paper, which are clustered together in the embedding space. The relative distance of dissimilar papers and similar papers with respect to a candidate is enforced to be larger than a margin. We leverage the positive pairs in contrastive learning to minimize the distance between the section embeddings of the same paper, ensuring that intra-paper similarity is preserved. Likewise, the negative instances in the contrastive loss ensure dissimilar papers are farther away. This setup ensures that section embeddings of the same paper are closer to each other in the embedding space than embeddings of other papers, preserving the local information context. Similarly, similar papers are closer in the embedding space in comparison to dissimilar papers, preserving interdocument relations. We present a schematic of the embedding space in Figure 1.

2 Related Works

SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2020) train a BERT architecture model on the Semantic Scholar corpus and biomedical corpora, respectively. Both models take sentences as input and learn sentence representations. These are not specifically trained to represent research papers.

SPECTER (Cohan et al., 2020) is initialized from SciBERT and uses a citation-based triplet loss for fine-tuning the model. It encodes the paper's title and abstract. It leverages citations to construct triplets for contrastive loss, i.e., w.r.t. to a query paper (q), a positive paper (k_+) is one of its cited papers. Similarly, a negative paper (k_-) for q is a paper that is not cited by q (but could be cited by k_+). SPECTER is evaluated on SciDocs (Co-

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 283–292
August 16, 2024 ©2024 Association for Computational Linguistics

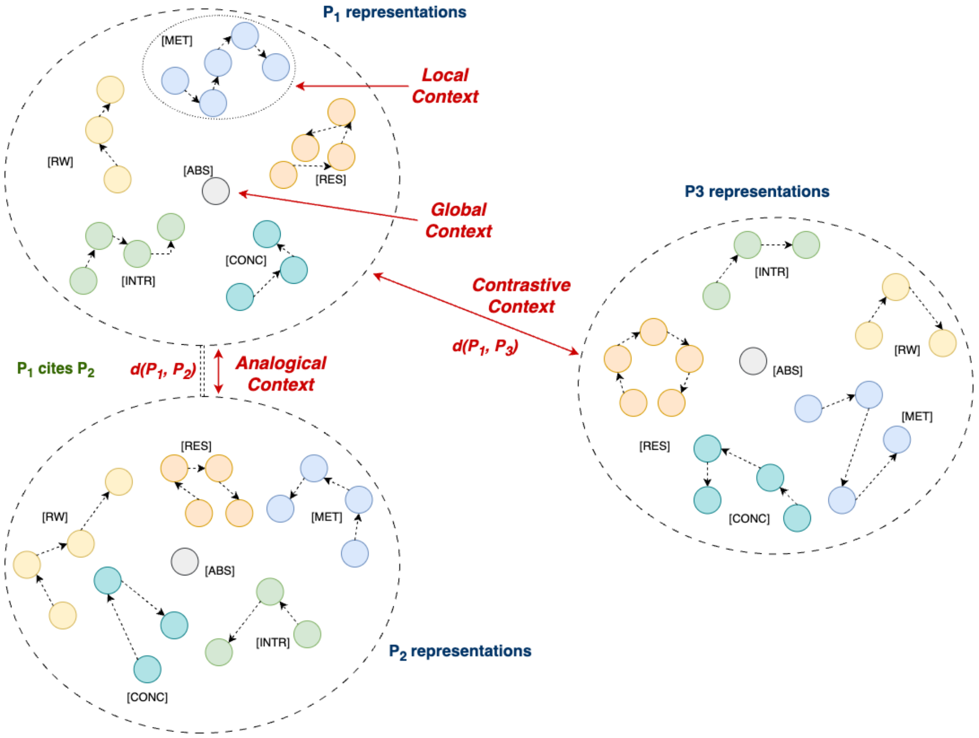

Figure 1: P_1, P_2, and P_3 represent three papers, and P_1 cites P_2. Aspect vectors are constructed from sections of research papers. Multiple vectors are learned for each section depending on model's context length. A hierarchy of distances is introduced between embeddings of different paper sections using contrastive loss with margin gradation.

han et al., 2020) benchmark consisting of seven tasks categorized into document classification, user activity prediction, citation prediction, and recommendation. Our proposed approach builds upon SPECTER by introducing a technique to utilize existing 512 context length limited transformers for large documents of any length. SciNCL (Ostendorff et al., 2022b) extends the SPECTER model with controlled nearest neighbor sampling over the citation graph that avoids collision between positive and negative samples.

OAG-BERT (Liu et al., 2022) encodes paper titles, abstracts, and heterogeneous entities such as authors, research fields, venues, and affiliations using a BERT model. They design strategies such as entity-type embeddings to denote heterogeneous entities and entity-aware 2D positional embeddings to demarcate inter and intra-entity token boundaries. The model also uses span-aware entity masking instead of BERT's random word masking to make the model learn entities.

FeRoSA (Chakraborty et al., 2016) is a faceted recommendation system for scientific articles. It recommends papers from the induced subnetwork of the candidate paper (built from cited papers and papers with highly similar content) for the facets: background (introduction), alternative approaches (related works), method, and comparison (results and conclusion). Edges in the network are assigned facets based on the section in which the paper is cited in the candidate paper. FeRoSA supports faceted recommendation, while our framework is focused on learning faceted representations for scientific articles that can be further used for downstream tasks such as faceted recommendation.

Aspect embeddings, i.e., embeddings for different facets of the paper, are learned in (Ostendorff et al., 2022a) and (Mysore et al., 2022). Embeddings for facets such as task, method, and dataset, are learned in Ostendorff et al. (2022a); however, these are again learned from the title and abstract of the paper. Ostendorff et al. (2022a) use the PWC dataset (Kardas et al., 2020), which contains task, method, and dataset labels for papers. The best-performing model Siamese-SciBERT is fine-tuned on the PWC dataset based on aspect labels. ASPIRE (Mysore et al., 2022), on the other hand, leverages co-citation sentences as a source of document similarity. However, unlike our proposed method, both these models use only the title and the abstract to learn aspect vectors. We also present a comparison of CoSAEmb with other models

Model	Input	Contrastive	Citation Information	Entity Information
SciBERT	Sentence			
SBERT	Sentence	✓		
SPECTER	Title & Abstract	✓	✓	
OAG-BERT	Title & Abstract			✓
Siamese-BERT	Title & Abstract	✓		
SciNCL	Title & Abstract	✓	✓	
ASPIRE	Sentence	✓	✓	
CoSAEmb *(ours)*	Full-text paragraphs	✓	✓	

Table 1: All existing models either encode title+abstract of the paper, or represent the full-text of the paper at sentence level.

in Table 1.

Singh et al. (2023) train multi-format scientific models for four types of tasks, specifically classification, proximity, search, and regression, in order to improve the generalization ability of models. They also introduce SciRepEval benchmark, which consists of twenty-five tasks to evaluate scientific document representations across diverse task formats. In this work, we limit our approach to learning embeddings for different facets of the paper and do not attempt to generalize them to diverse tasks.

As outlined in Tay et al. (2022), works on efficient transformers address the quadratic time complexity by using one of the techniques: (i) sparsify the attention matrix using fixed patterns, (ii) learnable patterns that learn the access pattern in a data-driven fashion (rephrase, arrange tokens based on similarity), (iii) introducing memory modules, (iv) low-rank approximation of the self-attention matrix, (v) kernels, (vi) recurrence, or a combination of these. Multiple variants of efficient transformers such as Linformer (Wang et al., 2020), Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), Reformer (Kitaev et al., 2020), Performer (Choromanski et al., 2021) have been proposed. While multiple efficient long document transformers are proposed (e.g., Tay et al. (2022) discuss thirty-two efficient transformers), only Longformer has been adapted for scientific papers and evaluation on the QASPER benchmark (Dasigi et al., 2021) shows significant scope for adoption of these models for the scientific domain. The performance and effectiveness of other efficient transformer models on scientific documents are largely unexplored. Unlike works on efficient transformers, we focus on reusing the existing models such as SciBERT and SPECTER with contrastive fine-tuning to learn full-text document representations.

3 CoSAEmb

In this section, we present CoSAEmb (pronounced ko-saam-bee), COntrastive Section-aware Aspect Embeddings. CoSAEmb is a supervised contrastive training framework for long documents using margin gradation. Scientific articles are long documents containing information on multiple aspects. We learn aspect representations for long documents instead of relying on a single document vector to capture all the information from various aspects. We briefly discuss the intuition behind our document representation framework next.

3.1 Motivation

A document presents a sequence of ideas in a continuous flow. The information in a document can be categorized into topics or aspects. For example, a product review can discuss different aspects of the product, such as its price and the quality of specific components. In scientific documents, key aspects can be defined as problem motivation, limitations of previous works, methodology, results, and conclusions drawn. This aspect-specific information is typically addressed in the Introduction, Related Works, Methods, Results, and Conclusion sections of the paper, respectively. In our work, we assume aspect-specific information is present in specific sections; and hence learn aspect vectors from the corresponding paper sections.

When reading a research paper, three types of document context contribute to its understanding:

1. ***Local Information Context:*** Each paragraph or sentence in a research paper contains information which could be understood as an answer to a highly specific fine-grained question. E.g., '*What evaluation metric was used for comparison? Why was a specific parameter*

included in the algorithm? How are the limitations of a specific previous work addressed? The paragraphs of the paper establish the local context necessary for comprehending and interpreting individual sentences.

2. ***Global Information Context:*** Global information represents the paper's main objective, primary contributions, and key findings. It provides a concise summary of the paper, combining coarse-grained information across multiple aspects. Generally, an abstract of the paper is an apt representative of global information.
3. ***Analogical or Contrastive Context:*** Human learning frequently involves drawing analogies or making comparisons to familiar concepts (Hofstadter, 1995; Schwartz and Bransford, 1998; Gentner et al., 2003). This concept positions a document with respect to others, capturing the similarities, dissimilarities, and shared properties. It emphasizes understanding inter-document relations through analogy or contrast. E.g., Two papers may be similar due to their common field of study, shared authors, or connected by a citation. This information about various inter-document relations may not be explicitly present in the document and is often acquired by humans with time.

We base our document representation framework on the above-discussed information contexts. As depicted by a schematic in Figure 1, we capture local information context for a paper P_1 by creating paragraph vectors for different aspects (sections in our case). The paragraph length is determined by the encoder model's context length. The global information is captured by the abstract vector. Finally, to capture inter-document relations, we employ citation information. If P_1 cites P_2 but does not cite P_1, then representations of P_1 is closer to P_2 in comparison to P_3. The reorientation of these document representations (paragraph or aspect vectors which constitute the local context, title and abstract vector constituting the global context) in the embedding space to capture these similarities is discussed in Section 3.3.

CoSAEmb leverages contrastive learning to represent long documents using a transformer model. CoSAEmb learns multiple representations for different aspects of the research papers specified in sections by learning section embeddings. We use a transformer model to compute representations for different sections of the paper, including the abstract. The model is fine-tuned using contrastive loss to optimize the local section and global abstract representations of the document. As we learn multiple aspect embeddings for the same paper capturing different aspects, it should be ensured that embeddings of the same paper are close to each other in the embedding space. In addition, it is also desirable that embeddings of similar papers should be closer than dissimilar papers.

3.2 Architecture

We use an encoder-only transformers model, specifically SciBERT (Beltagy et al., 2019) for encoding document sections. Each section is split into chunks of length 512, which is the maximum length supported by most models. Hence, we construct multiple embeddings for a paper from the full-text of the paper, which are reoriented using a triplet loss. Let t denote a triplet consisting of a query/anchor (q), positive (k_+), and negative key (k_-).

$$t = (q, k_+, k_-) \tag{1}$$

Triplet loss enforces an order of distances, i.e., the difference between the distance between the anchor and the positive key and the anchor and negative key is at least a certain value, called margin in triplet loss. The detailed loss function is presented in Section 3.4. We discuss the formulation of three types of triplets that reorient the embeddings to encapsulate similarity between sections and papers.

3.3 Embedding Reorientation Triplets

We construct three types of triplets to reorient the embeddings to optimize the section and abstract representations by training on local context, global context, and analogical or contrastive context.

3.3.1 Section Representations (Local context)

For each candidate paper p_c (section text represented by p_c^{intr}, p_c^{rel}, p_c^{meth}, p_c^{res}, p_c^{conc} and title+abstract represented by p_c^{abs}), we train the model with the triplet loss function. The triplet loss contains an anchor p_c, a positive (p_+), and a negative (p_-) instance with respect to the anchor.

Section representations are obtained from SciBERT. However, a section text may exceed 512 tokens, so we further break it down into chunks. The first chunk is prepended with the title, and the rest chunks prepend the last sentence from the preceding chunk. For example, let the

'Introduction' section consists of lines $\mathcal{L}$ = [l_1, l_2, ... l_n]. We split the sentences into sentence chunks such that total token size of the chunk is approximately 512. The chunk set $\mathcal{C}$ = [$\mathcal{C}_1$, $\mathcal{C}_2$, ..., $\mathcal{C}_p$]; such that

$$\begin{aligned}
&\mathcal{C}_1 = [l_1, l_2, ..., l_i] \\
&\mathcal{C}_2 = [l_{i+1}, l_{i+2}, ..., l_j] \\
&\mathcal{C}_3 = [l_{j+1}, l_{j+2}, ..., l_k] \\
&... \\
&\mathcal{C}_p = [l_{m+1}, ..., l_n]
\end{aligned}$$

Special tokens are appended to the sentence chunks to indicate the section as follows:

$$\begin{aligned}
&p_c^{intr\_\mathcal{C}_1} = \texttt{[TTL] title [INTR]}\ l_1, l_2, ..., l_i \\
&p_c^{intr\_\mathcal{C}_2} = \texttt{[INTR]}\ l_i\ \texttt{[INTR]}\ l_{i+1}, l_{i+2}, ..., l_j \\
&p_c^{intr\_\mathcal{C}_3} = \texttt{[INTR]}\ l_j\ \texttt{[INTR]}\ l_{j+1}, l_{j+2}, ..., l_k \\
&... \\
&p_c^{intr\_\mathcal{C}_p} = \texttt{[INTR]}\ l_m\ \texttt{[INTR]}\ l_{m+1}, l_{m+2}, ..., l_n
\end{aligned}$$

Chunks from the same section *sec*, are used to construct triplets of the form:

$$\begin{gathered}
(p_c^{sec\_C_{i+1}},\ p_c^{sec\_C_i},\ p_-^{abs}); i \geq 1 \\
\text{sec} \in \{\texttt{intr}, \texttt{rel}, \texttt{met}, \texttt{res}, \texttt{conc}\}
\end{gathered} \tag{2}$$

Similarly, we introduce other special tokens [INTR], [RW], [MET], [RES], and [CONC] for the introduction, related works, method, results, and conclusion sections, respectively.

3.3.2 Abstract Representations (Global context)

An abstract contains a summary of the paper and discusses multiple aspects such as method and results. So we consider an abstract to create document representations. But we also fine-tune the model using contrastive setup to influence the abstract embeddings by section embeddings. The triplets are formulated to ensure that local and global context is efficiently captured. The global context triplets for a candidate paper p_c are formulated as follows:

$$\begin{gathered}
(p_c^{sec\_C_1},\ p_c^{abs},\ p_-^{abs}); \\
\text{sec} \in \{\texttt{intr}, \texttt{rel}, \texttt{met}, \texttt{res}, \texttt{conc}\}
\end{gathered} \tag{3}$$

p_-^{abs} is the concatenated title and abstract (i.e. p_-^{abs} = [TTL] title [SEP] [ABS] abstract, where [TTL] and [ABS] are special tokens) of a paper that p_c cites. $p_c^{\text{sec\_}C_1}$ is formulated in a similar way, e.g., the 'Introduction' section p_c^{intr} = [TTL] title [SEP] [INTR] introduction, and denotes that only the first chunk is used.

As the abstract is a concise summary of the paper, we consider the abstract embedding a representative of the paper summary encoding high-level details. The paper triplets formulated in Equations (2) and (3) ensure that section embeddings representing information from various sections of the paper are similar to the abstract embedding, while preserving specific details from the section text. We posit that section information embeddings (or aspect embeddings) influence abstract embeddings and vice-versa, robustifying the embeddings to noise during the training using intra-document triplets. We also posit that the placement of abstract embedding in the vicinity of section embeddings allows orientation of the document vectors in a cluster-like structure, leading to a better-defined subspace. While initially the aspect vectors could be randomly distributed in the representation space, contrastive training reorganizes the vectors together into a cohesive document vector cluster. Section embeddings encode the specific low-level details from different sections of the paper. These embeddings capture the intra-document relations.

3.3.3 Inter-document Relation Representations (Analogical or Contrastive Context)

Multiple inter-document relations between scientific papers, such as papers belonging to the same field of study, cited papers, papers with the same authors, and co-cited papers, represent highly similar papers. In our work, similar to the SPECTER (Cohan et al., 2020), we use the citation information to capture inter-document information. We formulate inter-doc triplets based on citations as follows:

$$(p_c^{abs}, p_+^{abs}, p_-^{abs}) \tag{4}$$

With respect to the anchor paper p_c, the positive paper p_+ is cited by p_c. On the contrary, the negative paper p_- is not cited by p_c, but it could be cited by another paper that is cited by p_c. The triplets formulated using Equation (4) ensure that semantically similar papers are closer to each other in the embedding space than dissimilar papers. Positive papers p_+ can be sampled from the citations of the candidate paper p_c. Negative papers p_- can be sampled (i) randomly (easy negatives) or (ii) from the two-hop neighbor papers which weren't cited originally by p_c (hard negatives). Similar to p_c^{abs} in previous paper triplets, the format of each component is as: `[TTL]` title `[SEP]` `[ABS]` abstract.

3.4 Contrastive Learning Setup and Margin Gradation

Similar to SPECTER, we start with a SciBERT (Beltagy et al., 2019) initialized model, and train the Transformer model with the triplet loss $\mathcal{L}(t)$ defined in Equation (5). However, we use margin gradation for the triplet loss. We incrementally increase the margin value as we go from inter-document to global to local context.

$$\mathcal{L}(t) = \max(0, \\ d(q, k_+)^2 - d(q, k_-)^2 + m_t) \quad (5)$$

where, triplets *t* are sampled from the triplet set $\mathcal{T}$ defined below. $\mathcal{S}$ denotes the set of local context triplets constructed from sections, $\mathcal{G}$ denotes the set of global context triplets and $\mathcal{I}$ denotes inter-document (or cross-document) triplets. $\mathcal{P}$ denotes the set of candidate anchor papers and $\mathcal{F} = \{\texttt{intr}, \texttt{rel}, \texttt{met}, \texttt{res}, \texttt{conc}\}$.

$$\mathcal{T} = \{t_i : t_i \in \mathcal{I}\} \cup \{t_g : t_g \in \mathcal{G}\} \cup \{t_s : t_s \in \mathcal{S}\} \quad (6)$$

$$\mathcal{I} = \{t_c = (p_c^{abs}, p_+^{abs}, p_-^{abs}) : p_c \in \mathcal{P}\} \quad (7)$$

$$\mathcal{G} = \{(p_c^{abs}, p_c^{sec\_C_1}, p_-^{abs}) : p_c \in \mathcal{P}, sec \in \mathcal{F}\} \quad (8)$$

$$\mathcal{S} = \{(p_c^{sec\_C_i}, p_c^{sec\_C_{i+1}}, p_-^{abs}) : \text{ i} \geq 1, sec \in \mathcal{F}\} \quad (9)$$

The margin parameter in triplet loss for a triplet (q, k_+, k_-) enforces that the relative distance between positive pairs (q and k_+) and negative pairs (q and k_-) is greater than the margin value. A small distance in the embedding space translates to high similarity. The positive pairs in local context triplets consist of consecutive section chunks and hence are the most similar positive pair across the three triplet types. So we set the highest margin value for local context triplets. In contrast, the inter-document triplets contain abstracts from two papers as positives, so we set the margin to be the least there. To summarize, we set $m_{t_i} < m_{t_g} < m_{t_s}$, to ensure that intra-document embedding similarity is higher than inter-document similarity. This enforces an order of distances, i.e., with respect to a candidate paper q's abstract representation, a dissimilar paper (which is not cited by q) is at the maximum distance, followed by a similar paper (which is cited by q), followed by the section representation of the same paper which would be at the least distance.

Section	Frequency	Section Triplets
Introduction	81768	136417
Related Works	28663	28814
Method	97104	954025
Results	62400	115419
Conclusion	73218	21039

Table 2: Section headings identified for the 97402 papers in the dataset. Section triplets denote the number of local context triplets constructed from each section.

Triplet type	Frequency
Local Context (Eq. 2)	1255714
Global Context (Eq. 3)	343153
Inter-document (Eq. 4)	684100

Table 3: Distribution of different types of triplets in the dataset. The inter-document triplets are taken from the SPECTER training set to ensure a fair comparison. The Local context and Global context capturing triplets are created from paper full-text, only for the papers that are originally present in the SPECTER training set.

4 Experiment Details

Training Dataset CoSAEmb requires full-text of research papers and citation information to construct the section, abstract, and inter-document triplets for training the model. For a fair comparison of our methodology with SPECTER, we use the training dataset of SPECTER, which consists of 684100 triplets in the train set. Based on the citation information, these triplets are constructed from around 146000 query papers sampled from the Semantic Scholar corpus (Ammar et al., 2018). However, these triplets are constructed only from the title and abstracts of the papers (format similar to our inter-document triplets). We use the Semantic Scholar corpus to construct section triplets by extracting the full-text of the papers present in SPECTER's training set. We extract full-text of 97402 papers from the Semantic Scholar corpus.

We use keyword matching to categorize section

	background				method			
	RP	P@20	R@20	NDCG$_{\%20}$	RP	P@20	R@20	NDCG$_{\%20}$
SciBERT (2019)	17.32	23.75	38.33	45.53	9.92	8.85	29.87	30.13
SPECTER (2020)	26.48	32.19	50.77	68.07	9.21	12.58	36.36	36.66
OAG-BERT (2022)	13.85	18.12	23.63	37.32	13.62	7.47	26.81	23.73
ASPIRE (2022)	26.53	35.00	55.25	69.53	10.65	15.66	49.02	43.00
CoSAEmb *(ours)*	27.76	33.75	52.79	67.27	12.73	15.00	43.04	41.14
	result				*Aggregated*			
	RP	P@20	R@20	NDCG$_{\%20}$	RP	P@20	R@20	NDCG$_{\%20}$
SciBERT (2019)	9.92	16.53	34.33	38.51	12.32	16.17	34.06	38.00
SPECTER (2020)	17.72	25.9	59.95	56.11	17.69	23.38	48.96	53.31
OAG-BERT (2022)	9.87	13.16	30.42	34.00	12.40	12.84	26.92	31.58
ASPIRE (2022)	19.64	26.46	55.69	59.78	18.84	25.52	53.15	57.20
CoSAEmb *(ours)*	20.54	24.06	51.25	54.50	20.32	24.11	49.98	54.06

Table 4: CSFCube Benchmark: Results for the set of baselines methods. Metrics are R-Precision (RP), Precision@20 (P@20), Recall@20 (R@20), and NDCG$_{\%20}$ are computed.

headings into five classes: Introduction, Related Works, Method, Results, and Conclusion. We use the dataset of 25,482 section headings organized manually into previously mentioned five categories by Chakraborty et al. (2016) to collect seed keywords. To facilitate the categorization of new section headings in full texts, we developed a straightforward algorithm that assigns one of the following labels to each heading: Introduction, Related Works, Results, Conclusion, or Methods based on the dataset curated by Chakraborty et al. (2016). Section headings that do not match any of category headings in the dataset are automatically assigned to the Methods category. We present the distribution of different section headings in the dataset in Table 2. The statistics of total triplets in the dataset are presented in Table 3. As each section from the paper contributes to one global context triplet (as defined in Equation (3)), the total count of sections listed in Table 2 matches the number of Global Context triplets in Table 3. Likewise, the sum of section triplets presented in Table 2 equals the total number of Local Context triplets, which is 1255714.

Training Details We implement our model in PyTorch and train our model on three NVIDIA V100 GPUs. The model is trained for two epochs with a slanted triangular learning rate of 2e-5 with warmup steps equal to 0.1 fractions of the total steps. We use a batch size of 32, which fits on our three GPUs. We run ablations for margin values m_{t_i}, m_{t_g}, and m_{t_s}: (0.5, 1, 1.5), (1, 2, 3), and (2, 4, 6). The best results are obtained with m_{t_i} = 1, m_{t_g} = 2, and m_{t_s}=3.

5 Evaluation and Results

The description of datasets used for evaluating the capability of the CoSAEmb model are as follows:

CSFCube (Mysore et al., 2021) CSFCube is an annotated corpus of 50 query abstract-facet pairs from the ACL Anthology. The candidate set for each query is constructed from the S2ORC corpus using six methods, including TF-IDF and averaged word2vec. The three facets in the dataset are background, method, and result. As this dataset supports query by aspect, it is a relevant database to evaluate the model in an aspect-based retrieval scenario.

SciRepEval (Singh et al., 2023) SciRepEval is a benchmark for evaluating scientific document representations, which consists of twenty-five tasks across four formats: classification (e.g., field of study, MeSH Descriptors classification, etc.), regression (e.g., predict tweet mentions, peer-review rating, and maximum h-index of authors), proximity (e.g., same author detection, highly influential citations, etc.), and ad-hoc search (e.g., rank a candidate set for a query on TREC-COVID, CORD-19 datasets).

Full-text Paragraph Retrieval We curate a dataset of 1500 papers from arXiv (cs.CL and cs.AI categories, submitted to arXiv in 2022 so as to avoid overlap with the training dataset) and use it to evaluate if CoSAEmb model is able to generalize

	SciRepEval (SciNCL + Adapters + MTL CTRL)	CoSAEmb + Adapters + MTL CTRL
Out of Task Avg	62.5	58.2
In Task Avg	59.1	58.3
SciDocs Avg	90.6	89.0
All Avg	71.2	67.9

Table 5: As per evaluations in SciRepEval, we present the results of CoSAEmb with Adapters and MTL CTRL (control codes). While the results don't outperform SciRepEval, they are comparable.

Model	Recall@1	Recall@3
SciBERT	40.10	53.50
SPECTER	74.70	82.83
CoSAEmb (w/o margin gradation) ($m_{t_i} = m_{t_g} = m_{t_s} = 1$)	75.96	82.87
CoSAEmb (with margin gradation) ($m_{t_i} = 1, m_{t_g} = 2\ m_{t_s} = 3$)	76.71	83.44

Table 6: Full-text paragraph retrieval performance. CoSAEmb (with margin gradation, i.e. $m_{t_i} = 1$, $m_{t_g} = 2$, $m_{t_s} = 3$) shows improved performance over SciBERT and SPECTER in retrieving the paragraphs of a paper. CoSAEmb model variant trained without margin gradation (CoSAEmb ($m_{t_i} = m_{t_g} = m_{t_s} = 1$)) benefits from more data but the gain is not as significant as observed with margin gradation.

the order of distances among different paper section embeddings, on papers unseen during training. We set up a simple retrieval setup, where section embedding is used as a query and embeddings from a candidate set are retrieved based on Euclidean distance. The candidate set is composed of other section texts from the same paper as well as other papers. We evaluate if CoSAEmb embeddings retrieve other section texts from the same paper. We also evaluate a model trained with $m_{t_i} = m_{t_g} = m_{t_l}$ to evaluate if margin gradation contributes to performance gain, or if the gain is observed because of increased training data.

While the CoSAEmb models perform comparable to existing state-of-the-art models on CSFCube and SciRepEval benchmark, it shows an improved performance in retrieving the full-text paragraphs. On the CSFCube dataset, CoSAEmb performs slightly better on R-Precision metric, for most aspects. Unlike ASPIRE, CoSAEmb is not trained on any aspect-specific data and is just trained on full-text sections data only. For the CSFCube dataset, similar to other models we use the title and abstract text for query. However, for each of the specific facets, we separate the title and abstract with our special tokens in addition to the [SEP] token (for facet background, we use [INTR] and [RW]; for method facet we use token [MET], and for result facet we use [RES]). In comparison to SPECTER, which is the fairest baseline to CoSAEmb, improved performance can be observed for the 'method' aspect. This could be attributed to the full-text information available to CoSAEmb during training. On the SciRepEval dataset, we observe slightly worse performance of CoSAEmb. It could denote that it is not always required to utilize full-text information for the tasks present in SciRepEval benchmark. Singh et al. (2023) pretrain a new model similar to SPECTER to increase the domain coverage with data from 23 fields, which leads to a 15-point increase for In Task performance, which also indicates that training the model on diverse domain data would be helpful.

6 Conclusion

We present CoSAEmb, a technique to learn section-aware embeddings with any existing encoder-only transformer model without retraining it from scratch. It can be adapted to several domains that deal with long documents such as legal and research papers. We use contrastive loss to represent multiple section embeddings of papers in a compact cluster, and maximize the distance to dissimilar and minimize the distance to similar papers. Margin gradation introduces a hierarchy of distances. The representations can be used for various downstream applications, such as the generation of related works for manuscripts, predicting appropriate submission venues for manuscripts, and predicting missing citations or comparisons.

Acknowledgments

This work is supported by Prime Minister's Research Fellowship (PMRF), awarded to Shruti (PMRF-1701251) by the Government of India. We extend our heartfelt gratitude to Prof. Nipun Batra for his generous support in providing access to GPU computing facilities. Additionally, we are thankful for the use of the PARAM Ananta supercomputing facility at IIT Gandhinagar, launched under the National Supercomputing Mission by the Government of India.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tanmoy Chakraborty, Amrith Krishna, Mayank Singh, Niloy Ganguly, Pawan Goyal, and Animesh Mukherjee. 2016. Ferosa: A faceted recommendation system for scientific articles. In *Advances in Knowledge Discovery and Data Mining*, pages 528–541, Cham. Springer International Publishing.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. 2021. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. 2003. Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95:393–408.

Douglas R Hofstadter. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* Basic books.

Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. 2020. AxCell: Automatic extraction of results from machine learning papers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8580–8594, Online. Association for Computational Linguistics.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiao Liu, Da Yin, Jingnan Zheng, Xingjian Zhang, Peng Zhang, Hongxia Yang, Yuxiao Dong, and Jie Tang. 2022. OAG-BERT: Towards a unified backbone language model for academic knowledge services. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3418–3428.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceed-*

ings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Sheshera Mysore, Tim O' Gorman, Andrew McCallum, and Hamed Zamani. 2021. Csfcube - a test collection of computer science research articles for faceted query by example. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Malte Ostendorff, Till Blume, Terry Ruas, Bela Gipp, and Georg Rehm. 2022a. Specialized document embeddings for aspect-based similarity of research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, New York, NY, USA. Association for Computing Machinery.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022b. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel L. Schwartz and John D. Bransford. 1998. A time for telling. *Cognition and Instruction*, 16:475–522.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6).

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.