MISTI: Metadata-Informed Scientific Text and Image Representation through Contrastive Learning

Pawin Taechoyotin
Department of Computer Science
University of Colorado Boulder

Daniel Acuna
Department of Computer Science
University of Colorado Boulder

{pawin.taechoyotin,daniel.acuna}@colorado.edu

Abstract

In scientific publications, automatic representations of figures and their captions can be used in NLP, computer vision, and information retrieval tasks. Contrastive learning has proven effective for creating such joint representations for natural scenes, but its application to scientific imagery and descriptions remains under-explored. Recent open-access publication datasets provide an opportunity to understand the effectiveness of this technique as well as evaluate the usefulness of additional metadata, which are available only in the scientific context. Here, we introduce MISTI, a novel model that uses contrastive learning to simultaneously learn the representation of figures, captions, and metadata, such as a paper's title, sections, and curated concepts from the PubMed Open Access Subset. We evaluate our model on multiple information retrieval tasks, showing substantial improvements over baseline models. Notably, incorporating metadata doubled retrieval performance, achieving a Recall@1 of 30% on a 70K-item caption retrieval task. We qualitatively explore how metadata can be used to strategically retrieve distinctive representations of the same concept but for different sections, such as introduction and results. Additionally, we show that our model seamlessly handles out-of-domain tasks related to image segmentation. We share our dataset and methods (https://github.com/Khempawin/scientific-image-caption-pair/tree/section-attr) and outline future research directions.

1 Introduction

The famous saying "a picture is worth a thousand words" takes on a new meaning in the context of scientific publications. Scientific articles are filled with textual descriptions paired with images, diagrams, and figures that are essential for understanding the results presented. Previous work on natural images, such as CLIP (Radford et al., 2021), has shown that contrastive learning can be used to jointly represent text and images. Relative to traditionally available natural image datasets, scientific images are unique in two ways. First, they have precise textual descriptions, known as captions, which provide additional context and information. Second, they have metadata, such as a paper's title, sections, and curated concepts, that enable us to contextualize an image. New open access repositories (e.g., Wang et al. (2020); Lin et al. (2023)) allow us to take advantage of these unique characteristics. In this article, we propose and evaluate a novel model that uses contrastive learning to jointly represent text and images in scientific publications enriched with such metadata information.

Scientific images are often accompanied by captions that provide additional context and information (Li et al., 2018). For example, a caption might describe the experimental setup, the results of an experiment, or the conclusions drawn from the results. Captions are essential for understanding the research presented in scientific articles, but they are sometimes overlooked in image retrieval tasks. Recent work has shown the potential of multimodal models (Yin et al., 2023), and their application to science is yet to be explored. In addition to captions, scientific images are often associated with metadata such as the paper's title, sections, and curated concepts. This metadata provides valuable information about the context in which the image appears and can be used to enhance the representation learned by the model. For example, the section of an article potentially provides information about the type of image (e.g., experimental setup, results, discussion) while the title can provide information about the general topic of the research. Although previous work has explored this line of research (Wei et al., 2023; Eslami et al., 2023), there is still much more to be done to fully leverage the range of metadata available in science.

Understanding the content of scientific diagrams and images and producing descriptions for them

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 155–164
August 16, 2024 ©2024 Association for Computational Linguistics

are challenging tasks. For instance, a diagram might contain multiple components with a specific function or meaning. Captions provide additional context and information about the diagram but are often brief and may not fully explain its content. Conversely, writing captions for scientific diagrams is challenging, requiring concisely conveying complex information. These challenges become apparent during information retrieval, where we want to retrieve figures based on a textual description or vice versa. Building on the concepts of the CLIP model (Radford et al., 2021), we aim to build a vision text dual encoder model to facilitate scientific article generation. Similar to what happened to CLIP, the model can be used for other downstream tasks, such as image segmentation or zero-shot classification.

There are many image-caption datasets such as Flickr8k, Flickr30k (Wang et al., 2018), MSCOCO (Chen et al., 2015), RSICD (Lu et al., 2017), Medicat (Sanjay Subramanian and Hajishirzi, 2020) and other adaptations of image classification datasets namely CIFAR-10, and CIFAR-100 (Krizhevsky, 2009). These datasets are exceptional for their specific purposes, but the models do not generalize well into the scientific context. For instance, the captions and images in MSCOCO are mostly descriptions of everyday objects in everyday scenes. RSICD is a specialized dataset with captions of satellite images. Medicat is a specialized dataset with captions in the medical domain. On the other hand, captions in scientific articles are domain-specific for a particular field, with complex images representing diagrams, charts, or graphs. Apart from these datasets, there are image-caption scientific datasets with a diverse fields such as SciCap (Hsu et al., 2021), SciMMIR (Wu et al., 2024) and SCI-3000 (Darmanovic, 2022). Although these datasets are diverse but they lack the metadata associated with each image-caption pair which limits the context they belong to. Another aspect that is limited is the size of these datasets which will affect the representations learned.

To harness the richness found in images, we propose creating a new vision text dual encoder model to improve the performance of image retrieval tasks in scientific publications. We develop a dataset of scientific image captions based on open-access articles from PubMed Open Access Subset (National Library of Medicine, 2003). This dataset contains images and captions extracted from scientific articles and metadata such as a paper's title, sections, and curated concepts. We evaluate the usefulness of our model and dataset on various tasks. In sum, the contributions of our work are as follows:

- Develop a dataset of image captions derived from open-access scientific articles on PubMed.
- Implement a caption-enhancement technique to improve learned representations.
- Make the trained model and dataset publicly available to the scientific community.
- Present a variety of new tasks to assess the efficacy of contrastive learning within the scientific domain.

2 Related Work

2.1 Contrastive Representation Learning

Contrastive Learning is when we learn to tell apart objects by comparing the similarities between two or more similar objects and the differences between those objects and dissimilar ones. The goal is to make similar samples close in embedding space, and dissimilar samples far apart (Zimmermann et al., 2021). Recently, there has been a surge in the popularity of using such techniques to achieve better model performances (Yuan et al., 2022; Ciga et al., 2022; Wang et al., 2021). The performance improvements include the learning speed of the models, relevance in information retrieval or recommendation tasks, and reduced effort in dataset creation (Jaiswal et al., 2020). Even when we do not have access to supervised data, we can use contrastive learning effectively in a self-supervised setting (Falcon and Cho, 2020). In this paradigm, contrastive learning has thrived the most, especially in multi-modal tasks (Radford et al., 2021).

2.2 Dataset Enhancement in Deep Representation Learning

There are three key generalization factors for self-supervised contrastive learning (Huang et al., 2021): alignment of positive samples, divergence of class centers, and concentration of augmented data. The first key factor is related to the model's architecture, where the difference between the positive samples from the same class is minimized. The second key factor is related to the information behind the dataset: different classes should be significantly different; otherwise, we might encounter a "feature collapse," where the representations learned cannot distinguish between distinct classes. The third key factor is related to the density of positive samples, which adds more signal for the model to learn from. The work of (Wei et al., 2023) applied these principles to learn a joint representation of a text and an image. In their work, the captions were too generic and thus were enhanced with dictionary definitions. This approach used the exact architecture as proposed in (Radford et al., 2021). With such augmentation, there was a significant improvement in both zero-shot classification and zero-shot retrieval.

2.3 Scientific Article Metadata

Recent datasets have made available open access publications, including their full text and figures, such as the PubMed Open Access Subset (National Library of Medicine, 2003). Datasets of scientific publications can give us access to metadata in an easy-to-use format. For example, OpenAlex (Priem et al., 2022) contains indexes and metadata for a large number of articles spanning almost all human knowledge. These metadata include the publishing journals of each article, the fields associated with that journal, the concepts associated with each article, the authors associated with each article, and many more.

2.4 Scientific Vocabulary Embeddings

To process the text, we need to transform it into embeddings. The original CLIP model is based on RoBERTa. As demonstrated in MIREAD (Razdaibiedina and Brechalov, 2023), SciBERT (Beltagy et al., 2019), and SPECTER2 (Cohan et al., 2020), the vocabulary used to process text plays a great deal of importance in the performance of text encoding in the scientific domain. Therefore, other text encoders should be used for specific domains. These representations can be used as a starting point for the text encoding aspect of text-image contrastive learning methods.

3 Dataset and Methodology

This work focuses on enhancing the representation learning of scientific texts and images through metadata-informed embeddings. Our approach leverages scientific publications' unique vocabulary and metadata, such as concepts, keywords, and field-specific terminologies. To this end, we introduce a novel method that directly incorporates metadata into the training process of the contrastive learning models, and a specialized tokenizer to handle scientific metadata. This approach aims to improve the performance of embedding models on tasks such as zero-shot classification, information retrieval, and recommendation systems in the scientific domain.

This section will describe the model architecture of contrastive learning with a dual vision-text encoder, the information retrieval system, the process of building the dataset, the training parameters, and the evaluation metrics for information retrieval. The process includes extracting image-caption pairs from articles and metadata from OpenAlex.

3.1 A Scientific Image-Caption dataset

To build a new scientific image-caption dataset, we have analyzed over 4 million scientific articles from PubMed Open Access Subset (National Library of Medicine, 2003). A sample of a scientific article in XML format and the location of image-captions can be seen in figure 1A. This resulted in around 12 million image-caption pairs. Note that apart from the 12 million image-caption pairs, we have filtered out approximately 8 million image-caption pairs that were not found under a section in the articles. Among 12 million image-caption pairs, 9.78 million were used as the training dataset, 1.1 million as the validation set, and 1.1 million as the test set. One extra process was adding metadata from OpenAlex to each image caption based on the document's DOI. This led to an image-caption dataset that included the title of the related article, the related concept keywords, and the authors.

3.2 Model Architecture

The contrastive learning model architecture primarily consists of two components (Figure 1B). The first component, known as the text encoder, handles the tokenization and encoding of text into a 512-dimensional vector. A key subcomponent of the text encoder is the language embedding or vocabulary. To address this, we experiment with RoBERTa, MIREAD, SciBERT, and SPECTER2. RoBERTa is the only language embedding that was not trained on scientific text. The second component, the image encoder, is responsible for encoding images into a 512-dimensional vector. The model used for this component is the pre-trained Vision Transformer from CLIP (Radford et al., 2021), which takes 32 by 32 patches of the image and outputs a 512-dimensional vector. Training is performed by evaluating each batch separately. For each batch, all texts are encoded with the text encoder, and all images are encoded with the image encoder. We then maximize the cosine similarity between the paired encoded texts and images while minimizing the difference between non-paired images and texts. This process is illustrated in Figure 1B, where the green cells indicate the values to be maximized, and the white cells indicate the values to be minimized.

3.3 Information retrieval pipeline

The process of retrieving captions involves three steps. The first step is to encode all the reference captions, which we refer to as the dictionary. The second step is to encode the query image. The third step involves comparing the encoded query image with all the encoded captions. The outcome is a ranking of captions based on the similarity between the encoding of the query image and the captions. In information retrieval tasks, we typically define a parameter k, which specifies the number of samples to be returned. In this instance, we will use k equal to 1 to assess the model recall, formally known as Recall@1. We use Recall@1 due to each image having only 1 caption associated with it. Therefore, there is only 1 caption that is perfectly relevant for each image. Increasing k for Recall@5 or Recall@10 will only increase the evaluation score since there is a higher probability where the original caption will be in the top n rankings. On the other hand, there are other metrics that are commonly used such as ROUGE-n (Chin-Yew, 2004), BERTScore and BLEU. The reason we are not using such scores is that we are evaluating the representations not the output captions. Due to this, ROUGE-n which compares the matching n-gram between the original caption and the retrieved caption does not give us significant information regarding the representations learned. For BLEU, the score is used to evaluate the quality of the text compared to multiple reference captions. Since we only have 1 reference caption, it is essentially comparing only the retrieved caption to the single reference caption. This might result in the illusion that the retrieved caption is good in cases where both captions are similar in words but mean different things. Therefore, BLEU is not suitable in this case. For BERTScore, such metric is inherently used due to the ranking of captions based on the similarity between encodings.

3.4 Model training

The model was trained on a machine with 64 vCPUs and a NVIDIA A100. The vision encoder was a pre-trained Vision Transformer from OpenAI that takes in 32 by 32 patches of the image. As mentioned in Section 2.4, we systematically varied the language embedding model and the type of caption augmentation. This resulted in 16 variations, which comprise four language embedding models (MIREAD, ROBERTA, SCIBERT and SPECTER2) and four types of caption augmentations (no augmentation, augmenting with section, augmenting with section and metadata, and augmentation with operationalized metadata). Operationalized metadata means introducing special tokens representing these sections (see below). Since one training session spans 6 days, we created a subset of the full training data. This subset had a size of 700k image-caption pairs. These 16 models were trained on a batch size of 64 over 3 epochs with a learning rate of 0.00005 and a weight decay of 0.1.

Once the performance was measured, we selected the best validation variation to train on the rest of the training dataset. We increased the batch size to 192 and compared it to the equivalent model with textual metadata to demonstrate the effects of tokenizing the metadata.

3.5 Augmenting captions with additional metadata

We used three variations of augmentation on each caption: augmenting with sections, metadata, and metadata with separator tokens. The metadata comprises the article title, related concepts, and the location of the image caption within the article (i.e., section). The sections of the article are "Introduction," "Methods," "Results and Discussion," "Conclusion," and "Other." "Other" is used when the section title does not fit any known pattern.

Augmenting with section Enhancing the caption with the section is done by concatenating the section at the beginning of the caption. For example, a caption such as "western blot of protein nurturing" that appears in the "result and discussion" section will have an enhanced caption such as "result and discussion western blot of protein nurturing." The template is "<section> <caption>".

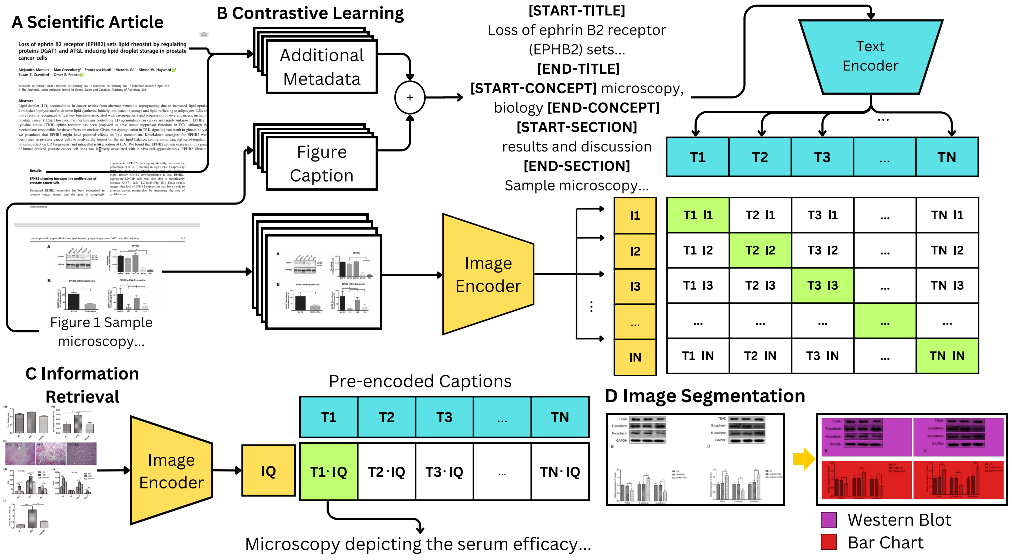

Figure 1: **A)** A sample scientific article and the XML representation depicting how images appear in scientific articles. **B)** A dual vision-text encoder architecture for contrastive learning between images and captions. It also illustrates how additional metadata are operationalized to enhance the captions during training. **C)** An illustration of text retrieval where each caption is pre-encoded with the text encoder. The query image is encoded and compared to each encoded caption. During retrieval, the image is encoded and compared to each encoded caption. **D)** A sample task such as image segmentation is performed by the learned representation. In this image, segments colored purple are classified as "Western Blot," and segments colored red are classified as "Bar Chart"

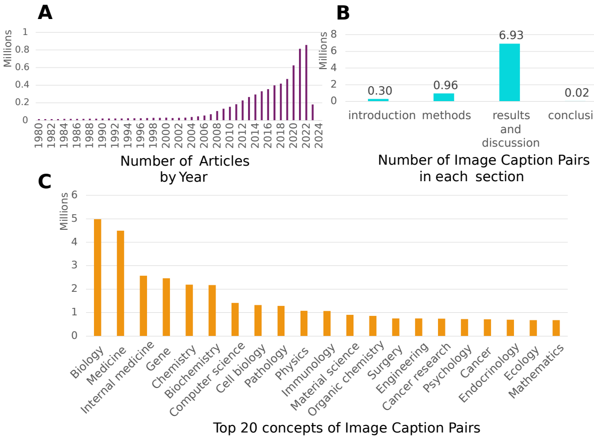

Figure 2: Statistics of scientific articles and training dataset. A) The extracted data was from 2.54 million articles. Most of the articles were published in the year 2016 to 2023. B) Distribution of the sections where each article's image captions are located. Most of the image-caption pairs were found in the result and discussion sections. C) Distribution of concepts related to each image-caption pair. Note that each image-caption pair might have more than one concept related to them.

Augmenting with section and metadata Enhancing the caption with metadata is similar to enhancing with only the section but includes more metadata about the caption which are the article title and the concepts of the article. For example, a caption as "structural integrity of ..." which appears in the "Methods" section in an article "Novel materials for..." with the related concepts as "Material Science, Tensile Strength" will have the enhanced caption as "Novel materials for... Material Science, Tensile Strength Methods structural integrity of ...". In this sense, the template we use is "<title> <concepts> <section> <caption>"

Augmenting with operationalized metadata To help the model distinguish text for each type of metadata and the caption, we enhanced the caption further by adding a special separating token for the title, concept, and section. The template we use is "[START-TITLE] <title> [END-TITLE] [START-CONCEPT] <concepts> [END-CONCEPT] [START-SECTION] <section> [END-SECTION] <caption>". We will call this type of augmentation as MISTI.

3.6 Relevance of Information Retrieval Sample

Information retrieval was used to evaluate the models. The query is an image, and the expected result is a caption. One key factor of information retrieval is defining the method of measuring relevance between the query and the retrieved sample. To address this, we used 3 metrics. The first metric is called *identity* metric, where the retrieved sample is only considered relevant if it matches the original caption paired with the query image. The second metric is called *section* metric, where the retrieved sample is only considered relevant if the section associated with the retrieved sample matches the section associated with the query image. The third metric is called *concept* metric, where we used the Jaccard Similarity index (Fletcher et al., 2018) to compute the overlap between the set of concepts associated with the retrieved caption and the set of concepts associated with the query image.

4 Results

4.1 Caption Retrieval

The performance of all models is shown in Table 1, where we illustrate the identity, section, and concept metric. We have included an average score across all three metrics to compare these models effectively. The baseline models had a Recall@1 of 0 on the identity metric. After fine-tuning the model on our scientific dataset, we see an increase in performance across all 3 metrics. The identity metric increased from 0 to around 0.06 with the original captions and captions augmented with sections. For the augmentation with Textual Meta and MIST, the score increased from 0 to around 0.05. Regarding the section metric, the scores increased to around 0.7 across all adaptations, and the concepts increased from 0.04 to 0.15. Overall, the RoBERTa adaptation saw the least improvement for the identity, section, and concept metric. Finally, the SPECTER2 MISTI + large batch performed the best with an average score of 0.4940. To summarize, all the baseline models performed sub-optimally. Fine-tuning the models on the scientific dataset improved the performance of the models. This indicates that the models have the potential to adapt to the dataset (Information in the dataset is not present in the pre-trained model). Lastly, all caption augmentations improved the performance of the model. However, each type of caption augmentations have different magnitudes of improvements on the models.

4.2 Qualitative Analysis of Information Retrieval

The results shown in Figure 3A and Figure 3B are for accurate and non-accurate retrieval, respectively. All the images associated with the retrieved caption are line charts that resemble the query image (Figure 3A). The images associated with the rank 0 and rank 1 caption resemble geographical plots, which align with the query image (Figure 3B). Conversely, images in rank 2 and 3 differ greatly from the query image, where rank 2 is a line chart and rank 3 is a stacked bar chart. To summarize, all the images for accurate retrieval are similar in nature but for non-accurate retrieval some retrieved results do not resemble the query image.

4.3 Retrieving images with different article sections

Figure 3C shows the images in different sections retrieval results. In such a task, we used the caption as "bar chart" and varied the section as "introduction," "method", "results and discussion", "conclusion", and "other". The results are the top 4 most similar images starting from rank 0, 1, 2, and 3. The "Other" section is used for section names that do not fit a standard keyword criterion. The results show that all section types, excluding "results and discussion," have matching rank 0 results but differ in the subsequent ranks. All retrieved images are bar charts, indicating that the retrieval results are qualitatively relevant to the queried caption.

4.4 Classification on Different parts of an image

Figure 4 shows a simple classification of image segments between the caption "bar charts" and "western blot". This task was done by encoding each textual category as a reference encodings. The image segments are then encoded and compared to each encoded textual category. The textual category that is the most similar to each encoded image segment is considered the output classification for those image segments. The result shows that image segments that resemble a "bar chart" are classified as a bar chart, and the image segments that resemble a "western blot" are classified as a western blot. Some segments are misclassified due to the segment being similar to both "bar chart" and "western blot." Additionally, some segments lack apparent

Model	Base + Adaptation	Recall@1			
		Identity	Section	Concepts	Average
Baseline	MIReAD	0.0000	0.5744	0.0465	0.2070
	(Radford et al., 2021) RoBERTa	0.0000	0.6175	0.0380	0.2185
	SciBERT	0.0000	0.2842	0.0423	0.1088
	SPECTER2	0.0001	0.4971	0.0428	0.1800
Text Model Finetuned Caption	MIReAD	0.0602	0.7157	0.1497	0.3085
	RoBERTa	0.0490	0.7091	0.1374	0.2985
	SciBERT	0.0595	0.7156	0.1484	0.3078
	SPECTER2	0.0595	0.7150	0.1485	0.3077
Caption + Section	MIReAD	0.0601	0.7220	0.1495	0.3105
	RoBERTa	0.0506	0.7197	0.1384	0.3029
	SciBERT	0.0609	0.7193	0.1499	0.3100
	SPECTER2	0.0603	0.7187	0.1501	0.3097
Caption + Section + Textual Meta	MIReAD	0.0549	0.7253	0.1497	0.3100
	RoBERTa	0.0366	0.7112	0.1333	0.2937
	SciBERT	0.0539	0.7193	0.1489	0.3074
	SPECTER2	0.0565	0.7231	0.1505	0.3100
	SPECTER2 + large batch	0.1717	0.7780	0.2826	0.4108
MISTI (Text Model Finetuned + Tokenized Meta)	MIReAD	0.0500	0.7163	0.1448	0.3037
	RoBERTa	0.0347	0.7128	0.1326	0.2934
	SciBERT	0.0518	0.7208	0.1462	0.3063
	SPECTER2	0.0527	0.7217	0.1485	0.3076
	SPECTER2 + large batch	**0.3000**	**0.7932**	**0.3889**	**0.4940**

Table 1: Table summarizes Recall@1 for caption retrieval across various models using a 70k-caption dictionary, evaluated on identity, section, and concepts metrics, alongside an overall average. Models are categorized into baseline (not finetuned), text model finetuned (subdivided into no augmentation, section augmentation, and section plus metadata augmentation), and text model finetuned with tokenized meta. The top performer, SPECTER2 large batch on text model finetuned with tokenized meta, is highlighted in bold.

features, such as the 4 segments at the bottom right corner.

5 Discussion

5.1 The effects of caption augmentation

Our analyses point out that the sub-optimal performance of the baseline model was due to the embedding of each sample being extremely close to each other. This closeness does not allow the model to distinguish between samples. This suggests that the representations of scientific contexts are even more densely packed, creating such bad performance. This trend of results applies to other tasks such as text to image retrieval where text to image and image to text retrieval performance are identical when varying the dictionary size. This aligns with how contrastive learning works in general (e.g., see Radford et al. (2021); Zimmermann et al. (2021); Yuan et al. (2022); Ciga et al. (2022); Wang et al. (2021)). Such a hypothesis is confirmed with the inclusion of the fine-tuned models, as performance increases due to the dispersion of samples. Moreover, augmenting the captions with Textual Metadata also increases the performance when compared to the baseline models. The degree of improvement is slightly lower than that of the model without the additional metadata on the identity metric, but the average performance is higher. In other words, the improvement on the section and concepts metric combined was higher than the delta of the identity metric. We understand that this is due to the additional metadata acting as a signal for the model to group samples with similar sections and concepts closer in the embedding space. This phenomenon correlates with adding dictionary definitions for vague captions as observed in (Wei et al., 2023). The metadata leads the model to create regions for each section/concept in the embedding space. Consequently, this leads to improvements in the performance of the section and concept metric. We tokenized the metadata with a special separator token to improve distinguishable abilities, increasing the model's performance. We see a substantial difference between 0.1717 for textual and 0.3 for tokenized metadata. The results of these new tokens led to the model capable of producing a substantial differentiation of samples, ultimately resulting in better representations, similar to what is described in (Huang et al., 2021). Thus, our caption augmentation task demonstrates how important it is to develop specialized scientific representations to tell samples apart.

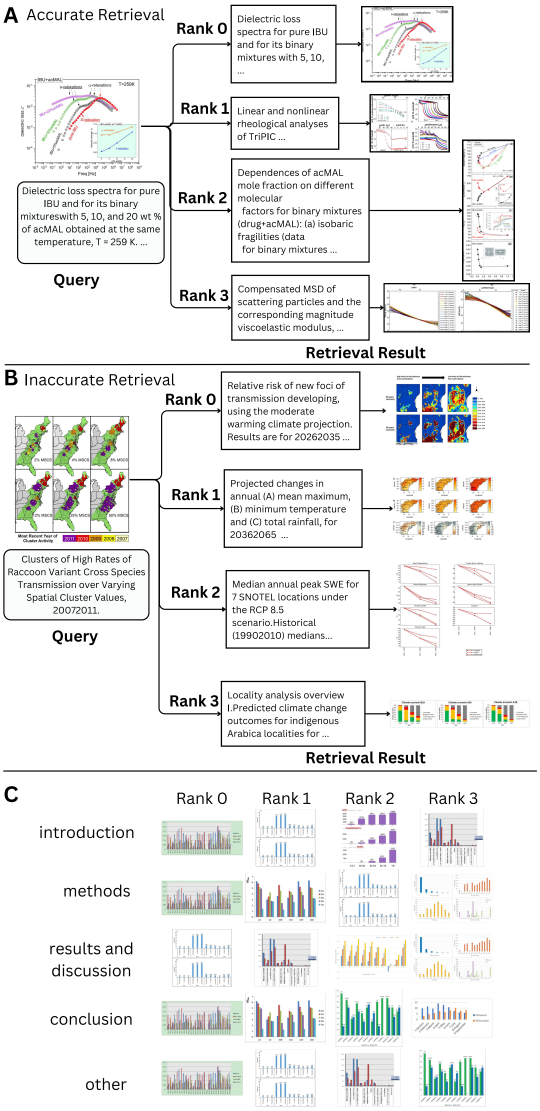
A Accurate Retrieval
Rank 0
Dielectric loss spectra for pure IBU and for its binary mixtures with 5, 10, ...
Rank 1
Linear and nonlinear rheological analyses of TriPIC ...
Rank 2
Dependences of acMAL mole fraction on different molecular factors for binary mixtures (drug+acMAL): (a) isobaric fragilities (data for binary mixtures ...
Rank 3
Compensated MSD of scattering particles and the corresponding magnitude viscoelastic modulus, ...
Dielectric loss spectra for pure IBU and for its binary mixtureswith 5, 10, and 20 wt % of acMAL obtained at the same temperature, T = 259 K. ...
Query
Retrieval Result
B Inaccurate Retrieval
Rank 0
Relative risk of new foci of transmission developing, using the moderate warming climate projection. Results are for 20262035 ...
Rank 1
Projected changes in annual (A) mean maximum, (B) minimum temperature and (C) total rainfall, for 20362065 ...
Rank 2
Median annual peak SWE for 7 SNOTEL locations under the RCP 8.5 scenario.Historical (19902010) medians...
Rank 3
Locality analysis overview I.Predicted climate change outcomes for indigenous Arabica localities for ...
Clusters of High Rates of Raccoon Variant Cross Species Transmission over Varying Spatial Cluster Values, 20072011.
Query
Retrieval Result
C
Rank 0
Rank 1
Rank 2
Rank 3
introduction
methods
results and discussion
conclusion
other

Figure 3: **Caption retrieval with corresponding images** A) Shows the top 4 captions that are retrieved when the top 1 caption retrieved is originally paired to the query image in the dataset. It also shows the original corresponding image pair for each retrieved caption. B) Shows the top 4 captions that are retrieved when the top 1 caption retrieved is not the originally paired to the query image in the dataset. It also shows the original corresponding image pair for each retrieved caption. C) Shows the retrieved image with different sections.

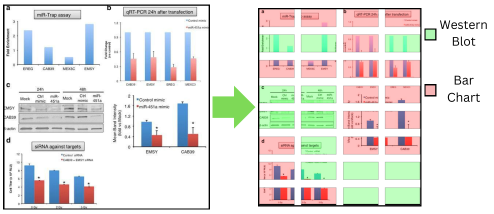

Figure 4: **Image Segment classification on "bar chart" and "western blot"**. The image on the left is the input image. The images on the right are segments of the image overlayed with colors to illustrate whether that segment is a "bar chart" or a "western blot". Green is for western blot and Red is for bar chart.

5.2 Qualitative Analysis of Information Retrieval

One hypothesis regarding the low performance on caption retrieval is that the model was retrieving results similar to the query but not the original pair (Wei et al., 2023). Based on the result in 4.2, accurate retrievals have similar images even though the images were never compared during training. This aligns with the findings of learning the image representation via text supervision (Radford et al., 2021). Simultaneously, we see the same behavior with non-accurate retrievals, but only the top 2 ranks are relevant. This implies that while quantitatively, we have a low recall, the retrieved information is still qualitatively relevant.

5.3 Retrieving Images with different article sections

According to the results in 4.3, the sections do affect the embedding, but the caption dominates in the end. This is interpreted from the result in rank 0 being the same but the subsequent ranks being different (Li et al., 2018). In other words, the results demonstrate how sections are essentially a style and can be applied to an image to adjust the visual details while conveying the original meaning of the caption. Similar results were studied for applying styles on images for fashion generative models (Baldrati et al., 2022; Sain et al., 2021).

5.4 Classification on image croppings

Classification of image segments to classes is one of the long-standing tasks for classification as seen in ResNet50 (Targ et al., 2016), VGG (Kaur and Gandhi, 2019) and Fast-R-CNN (Girshick, 2015).

These works performed well but were limited to fixed classes. Essentially, the final output layer of these models is a representation of the images compared to the predetermined classes. Therefore, it is understandable that our model can perform the classification tasks given the classes can be described in text form. Interestingly, our model performs on par with such dedicated models. However, prediction results can sometimes be incorrect when the similarity scores between the image and each possible class are extremely close to each other. This task showcases the potential of using our model on the downstream tasks in a zero-shot fashion.

5.5 Summary

In conclusion, our results show the important role of metadata augmentation in enhancing model performance in the scientific context. Our systematic incorporation of fine-tuning and specialized tokenization techniques shows that metadata improves contrastive learning substantially by helping the model distinguish between samples. The findings underline the potential of targeted data augmentation strategies in advancing the capabilities of NLP models, suggesting promising avenues for future research.

6 Conclusion

The goal of this project was to create representations of images and captions in scientific images based on the simultaneous learning of their relationship using contrastive learning. This representation can be applied to multiple tasks, such as improving the understanding of figures and textual descriptions of scientific concepts and helping generate better images and captions. Our zero-shot segmentation example demonstrates the numerous downstream tasks where our model can be applied. The substantial performance differences between the generic image-text models and our metadata-informed model show that more work is needed in the scientific domain. For future work, additional metadata such as authors, publication venues, and citations might be explored further to augment the data.

References

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Conditioned image retrieval for fashion using contrastive learning and clip-based features. In *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*, MMAsia '21, New York, NY, USA. Association for Computing Machinery.

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.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server.

Lin Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.

Ozan Ciga, Tony Xu, and Anne Louise Martel. 2022. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

Filip Darmanovic. 2022. *SCI-3000: A novel dataset for the task of figure, table, and caption extraction from scientific PDFs*. Ph.D. thesis, Technische Universität Wien.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163.

William Falcon and Kyunghyun Cho. 2020. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104*.

Sam Fletcher, Md Zahidul Islam, et al. 2018. Comparing sets of patterns with the jaccard index. *Australasian Journal of Information Systems*, 22.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. 2021. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Taranjit Kaur and Tapan Kumar Gandhi. 2019. Automated brain image classification based on vgg-16 and transfer learning. In *2019 international conference on information technology (ICIT)*, pages 94–98. IEEE.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. In *Learning Multiple Layers of Features from Tiny Images*.

Pengyuan Li, Xiangying Jiang, and Hagit Shatkay. 2018. Extracting figures and captions from scientific publications. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1595–1598.

Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. 2023. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2017. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195.

National Library of Medicine. 2003. Pmc open access subset. `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`. May 15, 2024.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Anastasia Razdaibiedina and Alexander Brechalov. 2023. Miread: Simple method for learning high-quality representations from scientific documents.

Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. 2021. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8504–8513.

Sachin Mehta Ben Bogin Madeleine van Zuylen Sravanthi Parasa Sameer Singh Matt Gardner Sanjay Subramanian, Lucy Lu Wang and Hannaneh Hajishirzi. 2020. MedICaT: A Dataset of Medical Images, Captions, and Textual References. In *Findings of EMNLP*.

Sasha Targ, Diogo Almeida, and Kevin Lyman. 2016. Resnet in resnet: Generalizing residual architectures.

Jian Wang, Zhiguo Cao, Yang Xiao, and Xinyuan Qi. 2018. Supervised guiding long-short term memory for image caption generation based on object classes. In *MIPPR 2017: Pattern Recognition and Computer Vision*, volume 10609, pages 148–155. SPIE.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. 2021. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033.

Yixuan Wei, Yue Cao, Zheng Zhang, Houwen Peng, Zhuliang Yao, Zhenda Xie, Han Hu, and Baining Guo. 2023. iclip: Bridging image classification and contrastive language-image pre-training for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2776–2786.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. Scimmir: Benchmarking scientific multi-modal information retrieval.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Liangzhe Yuan, Rui Qian, Yin Cui, Boqing Gong, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, and Ting Liu. 2022. Contextualized spatio-temporal contrastive learning with self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13977–13986.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. 2021. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12979–12990. PMLR.