Semantic Labels-Aware Transformer Model for Searching over a Large Collection of Lecture-Slides

K. V. Jobin[1] Anand Mishra[2] C. V. Jawahar[1]
[1]IIIT Hyderabad [2] IIT Jodhpur
{jobin.kv@research., jawahar@}iiit.ac.in mishra@iitj.ac.in

https://jobinkv.github.io/lecsd

Abstract

Massive Open Online Courses (MOOCs) enable easy access to many educational materials, particularly lecture slides, on the web. Searching through them based on user queries becomes an essential problem due to the availability of such vast information. To address this, we present Lecture Slide Deck Search Engine – a model that supports natural language queries and hand-drawn sketches and performs searches on a large collection of slide images on computer science topics. This search engine is trained using a novel semantic label-aware transformer model that extracts the semantic labels in the slide images and seamlessly encodes them with the visual cues from the slide images and textual cues from the natural language query. Further, to study the problem in a challenging setting, we introduce a novel dataset, namely the Lecture Slide Deck (LecSD) Dataset containing $54K$ slide images from the Data Structure, Computer Networks, and Optimization courses and provide associated manual annotation for the query in the form of natural language or hand-drawn sketch. The proposed Lecture Slide Deck Search Engine outperforms the competitive baselines and achieves nearly 4% superior Recall@1 on an absolute scale compared to the state-of-the-art approach. We firmly believe that this work will open up promising directions for improving the accessibility and usability of educational resources, enabling students and educators to find and utilize lecture materials more effectively.

1. Introduction

Online education has become increasingly popular in recent years, partly due to its convenience and flexibility. As a result, there is now a greater demand for effective presentation materials to support this mode of learning. While plenty of lecture slide presentations on various topics are available online, searching for a relevant and

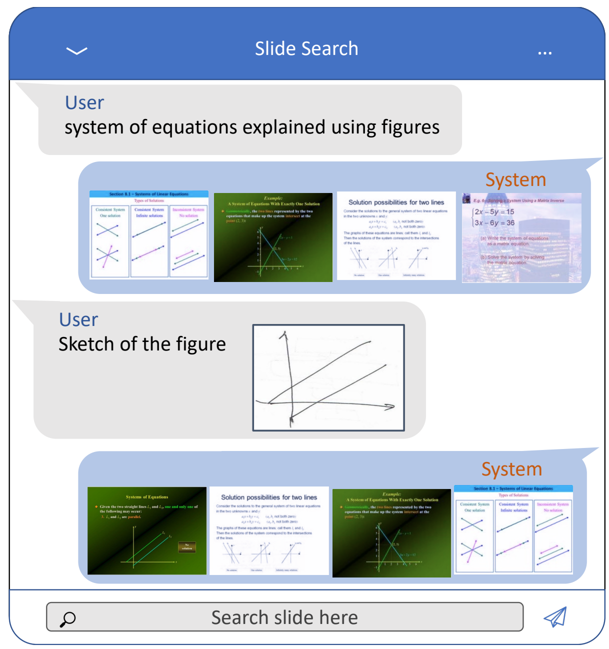

Figure 1. We present Lecture Slide Deck Search Engine – a semantic labels-aware transformer model that enables users to query and retrieve relevant lecture slides from a large collection using natural language summaries or hand-drawn sketches as queries. (Best viewed in color).

compelling slide deck for a given query can be tedious and time-consuming for educators and students. Further, in the retrieval task, using text input may not be the most suitable option in some scenarios such as, i) when the user cannot recall the correct keyword for searching, however, they have a picture in mind, ii) students with limited language proficiency struggle to express the search intent using text, and may prefer to draw the concept. Towards accelerating research on this important topic, we present Lecture Slide Deck Search Engine – a model that supports both natural language queries as well as hand-drawn sketches and performs a search on a very-large-scale collection of slide im-

Datasets	Features						Size	Avail.
	Slide Segments	Figures	Sketches	Slide Text	Transcript	Summary	# Slides	
VLEngagment [3]								✓
LectureBank [22]	✓(M)			✓(A)			51,939	✓
ALV [11]	✓(A)				✓(A)		1,498	✓
LectureVideoDB [10]	✓(M)			✓(M)			5,000	✓
GoogleI/O [4]				✓(A)	✓(A)			✓
LaRochelle [30]	✓(A)			✓(A)	✓(A)		2,350	
MLP Dataset [21]	✓(M)	✓(M)		✓(A)	✓(A)		9,031	✓
LecSD Dataset (ours)	✓(M)	✓(M)	✓(M/A)	✓(A)		✓(M/A)	54,000	✓

Table 1. Proposed LecSD Dataset as a comparison to the existing related datasets. The A and M represent that the features are extracted automatically and manually, respectively. Our dataset is larger with respect to the number of slide images and it has unique features of hand-drawn figure sketches and slide summaries as queries.

ages. This search engine's functionalities are illustrated in Figure 1.

In recent years, there has been a growing interest among researchers in developing retrieval systems for educational lecture videos and presentation slide images [13–15, 18, 24, 25]. The success of these retrieval systems can enable modern AI to design novel slides by combining search results and reusing figures and graphs from existing slide images, thereby minimizing manual effort. However, existing retrieval systems [14, 21] were developed to retrieve slides from video files and have the following limitations: (i) they rely on the transcript in the video to retrieve slides, which are not always contextually aligned with the slide image, also not always available, e.g. in lecture slide *image collection*. (ii) These retrieval systems are restricted to text queries and do not support hand-drawn sketches of diagrams as queries. (iii) Text queries often contain logical regions (semantic labels) like titles, bullet points, and figures, and figure types like line graphs, bar charts, Venn, and tree diagrams. The existing architectures are not explicitly designed to handle and learn these semantic regions and figure types. We propose Lecture Slide Deck Search Engine to overcome these limitations.

To study the problem of lecture slide image retrieval in a rigorous setting and evaluate the efficacy of our proposed model, we present a novel large-scale dataset, Lecture Slide Deck (LecSD). In Table 1, we compare our proposed dataset viz. LecSD with the existing related datasets. Our dataset is the largest with respect to the number of slides and has unique features such as the availability of figure sketches and a slide summary as queries. We selected lecture slides from *Data Structure* course to increase the dataset's complexity, especially due to their similarity, which creates challenging negatives for retrieval. The similarity among slide figures further amplifies the difficulty of the sketch-based retrieval task. Further, to assess our model's adaptability, we also included slides from *Computer Networks* and *Optimization* courses as well.

The proposed Lecture Slide Deck Search Engine is a unique architecture that handles text and sketches and combines both queries and demonstrates impressive performance on the lecture retrieval task and clearly outperforms related baselines. Compared to existing slide retrieval approaches [14, 21], our approach utilizes the semantic labels of slide images in the transformer model where the text and slide images are combined using a vision and language transformer (ViLT) [20] and the queries are encoded with PIE-Net [29].

Contributions: We make the following contributions: (i) We present the LecSD – a very large-scale dataset of lecture slide images harvested from the web. The dataset contains $54K$ slide images covering topics of *Data Structures, Computer Networks, and Optimization* and manually written natural language summaries and hand-drawn sketch queries corresponding to figures to search lecture slide images. (ii) We propose a model that leverages the semantic label of slides and encodes them into a novel semantic label-aware transformer model. The representation learned using this model is used to score against the representations for natural language summary or hand-drawn sketch query to learn the relevance of slides for a given query. (iii) We perform extensive experiments and ablation to verify the efficacy of our model. Our proposed approach significantly outperforms competitive approaches and thereby establishes a new state-of-the-art for the task.

2. Related Work

Examining images of classroom slides has been an active area of research. Early works such as Deshpande et al. [6] showcased a real-time interactive virtual classroom multimedia distance learning system; Zhu et al. [31] presented a virtualized classroom project that emphasized automatic data collection, analysis, multimodal synchronization, compression, cross-media indexing, and archiving. Recent works like Haurilet et al. [15, 25] focused on layout segmentation for classroom slide images. Building on this,

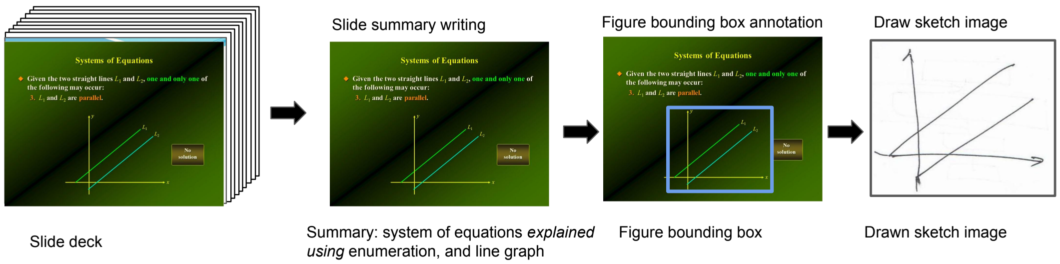

Figure 2. We present the LecSD towards developing a benchmark for retrieving educational contents, to be specific lecture slides from computer science topics. Here, we show our proposed annotation pipeline. First, annotators were asked to write a summary of the slides from the collection of slide image decks. Then, annotate the bounding boxes for the figures, and finally, draw a sketch image corresponding to the annotated figures. (Best viewed in color).

Jobin et al. [18] expanded the approach by implementing a narration system for classroom slides using the identified layout regions. In this section, we briefly present a literature survey on datasets and cross-modal retrieval in the space of lecture slides and outline the distinctions of our proposed dataset and approach.

2.1. Lecture Slide Retrieval Datasets

Several lecture slide datasets for the retrieval task have been introduced in the literature [3, 4, 10, 11, 21, 22, 30]. Some of the popular datasets are listed here: (a) The LectureBank [22] comprises $1,352$ online lecture PDF files extracted from 60 Computer Science courses covering the following five sub-domains: Machine Learning, Natural Language Processing, Deep Learning, Artificial Intelligence, and Information Retrieval. Additionally, the dataset includes more than 1K concepts automatically extracted to form an in-domain vocabulary, along with annotations for prerequisite relation pairs that involve 208 concepts. (b) The VLEngagement dataset [3] includes content-based features, e.g., stop-word frequencies and video-specific features, e.g., silence and video duration, extracted from publicly accessible 4K scientific video lectures. The tasks related to this dataset are to predict context-agnostic engagement in video lectures and rank video lectures based on their engagement levels. (c) The ALV dataset [11] contains lecture videos that were generated using automatically created transcripts from academic online videos. (d) The LectureVideoDB [10] comprises 5K frames. These frames contain annotated text characters, with a focus on detecting and recognizing text within lecture videos. The videos cover a range of 24 distinct courses in science, management, and engineering. (e) The GoogleI/O dataset [4] contains 209 presentation videos from the Google I/O conferences held between 2010 and 2012. The authors provide textual information from the spoken content and the slides in this dataset. (f) The LaRochelle dataset [30] comprises 47 French lecture recordings, totaling 65 hours of video, conducted in the author's lab. Each lecture, on average, features 50 slides. The researchers investigate cross-modal retrieval, employing a bag-of-words approach for both textual content and visual tokens. (g) The MLP Dataset [21] includes slides and spoken language, encompassing over 180 hours of video and more than 9000 slides, featuring contributions from 10 lecturers across diverse subjects such as computer science, dentistry, and biology.

Most of these datasets retrieve slides or videos based on the corresponding transcript. However, in real-world scenarios, the query may differ from the transcripts. We assess our proposed dataset – LecSD in comparison with these datasets in Table 1.

2.2. Cross-modal Image Retrieval

The baseline model for retrieving slide images given text and sketch query has been derived from the existing cross-modal image retrieval methods [5, 21, 28, 29]. Of these models, CLIP [28] stands out as a popular benchmark for aligning images with text. CLIP undergoes training to excel in a diverse range of tasks, and this acquired task learning can be effectively harnessed through natural language prompting, enabling seamless zero-shot transfer to numerous existing datasets. Further, two successful approaches in this area are: (a) PCME [5]: which characterizes each modality by modeling probabilistic distributions in a shared embedding space, employing Hedged Instance Embeddings [26]. This approach captures the uncertainty associated with mapping an image to a latent embedding space by dispersing density across plausible locations. The embeddings are treated as random variables, and the model is trained based on the variational information bottleneck principle [1]. (b) PVSE [29]: which captures one-to-many alignment for cross-modal retrieval. It learns diverse representations for the data samples by fusing global context with locally-guided features through multi-headed self-attention

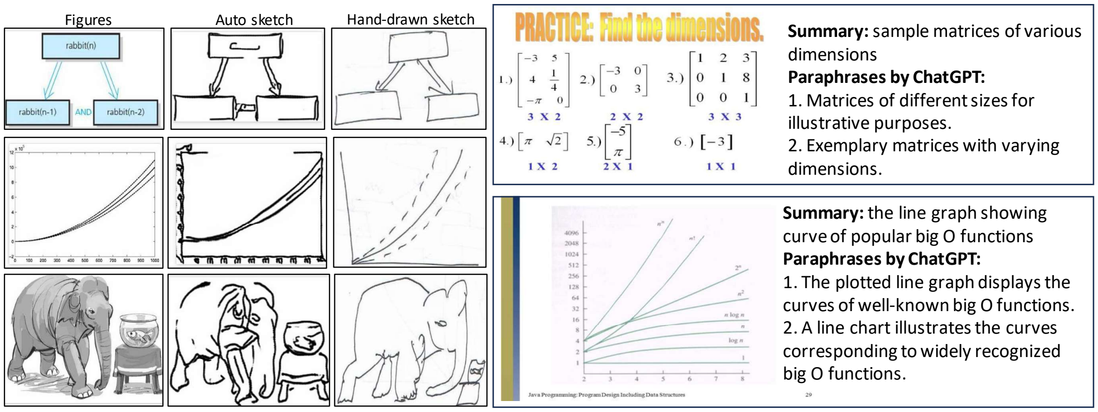

Figure 3. Sample figures in LecSD dataset. The first column shows cropped figure regions from the slide image. The second column of sketches is generated using photo-sketching [23]. The third column shows the manually drawn sketches. The last column shows the slide images and its manual summary and its two paraphrased sentences generated using chatGPT [2].

and residual learning.

In contrast to existing literature, our proposed model utilizes semantic labels within the slide image in the transformer architecture to facilitate cross-modal retrieval.

3. Lecture Slide Deck Dataset

We present a very large-scale, one-of-a-kind lecture slide dataset, namely the LecSD. This dataset's image and associated annotations can be downloaded from our project website. The slide images of the LecSD are harvested from the web[1] for a popular computer science and engineering course, namely *Data Structures*. We used the following popular sub-topics to search slide decks: *a) Arrays and Structures, b) Stacks and Queues, c) Lists, d) Trees, e) Graphs, f) Sorting, g) Hashing, h) Heap Structures, i) Search Structures, j) Algorithms, k) Stacks and Queues, l) Queues,* and *m) Binary trees*. By cleaning the collected slide image by removing duplicate slides and slides with no meaning, we obtained 1700 slide decks with around $50K$ slide images. We split the data into train, validation, and test sets of $30K$, $10K$, and $10K$ slide images, respectively. In addition, we collected and manually annotated 4000 slide images from the topic *Computer networks* and *Optimization* to demonstrate the generalizability of the proposed model.

3.1. Annotations

We obtain annotations for our dataset to enable retrieval of slide images and their components, such as figures, tables, and equations, using multimodal queries, including natural language text and drawing. We restrict our manual annotation to only the evaluation (test) and validation set since manual annotations of queries for building a large system are time-consuming, cumbersome, and not scalable. We automatically annotate the training data using the state-of-the-art slide segmentation, figure classification, and OCR modules.

3.1.1 Manual annotation

We generate organic queries for slide image retrieval with the help of five annotators. To obtain manual annotations for the test and validation set of our dataset, we provide slide images to the annotators and ask them to write a brief sentence about a given slide image. This brief sentence serves as a summary query for our dataset. Further, if figures are present in the slide, annotators were asked to draw the corresponding sketch of the figure on a paper. This paper is scanned, and the cropped sketch region is used as a sketch query for the slide. Figure 2 shows the annotation pipeline for a single slide image. We provide a modified version of the VGG image annotation tool [9] to annotators for annotating the slide image summary and figure regions. In order to overcome the annotation bias [27], we used chatGPT [2] to generate paraphrased sentences of written summaries. Sample drawn sketches and written summary are shown in Figure 3.

3.1.2 Automatic annotation

The primary goal of automatic summary annotation is to train the language model to retrieve slide images given an organic query. We conducted a study on organic queries to retrieve slide images and concluded as follows: i) we noticed that the keyword specific to a slide most frequently occurs in the title of the slide image. ii) the keywords can also occur in enumeration or paragraphs for the slides with the slide image having no titles or the title of common words such as overview, conclusion, methods, and prob-

[1] https://slideplayer.com/

lem. iii) the summary can also contain the list of logical regions such as enumeration, paragraphs, tables, equations, and various figure classes such as line graphs, bar charts, photographs, etc. iv) the logical region name need not be consistent in the summaries. As an example, the enumeration can be mentioned as bullet points. Hence, we designed the automatic slide summary as a predefined sentence structure, as $\underline{T}$ explained using $\underline{C}$, where T is the OCR text obtained from slide titles, enumeration, or paragraph regions. C lists logical regions such as enumeration, paragraphs, tables, equations, and figures. We randomly replace the logical region names with its synonyms. The slide layout segmentation model [18] is used to identify the regions. To identify the type of figures present in slide images, we use a trained model with DocFigure [17], having 28 types of figure classes.

The sketches of figures in the slide image are automatically created using photo-sketching [23]. The photo-sketching model is designed to generate contour drawings and boundary-like drawings that capture the outline of the visual scene. Hence, the model is well-suited for creating sketches of document figures. Figure 3 shows the sample sketches created by the Photo-sketching model. First, we identify the figure regions using the layout segmentation model and create sketches using the Photo-sketching model.

We manually extract text, draw the layout, and identify figure types from 100 slide images to evaluate the quality of automatic extraction of text, layout, and figure class. To extract the text in slide images, we use Google Lens OCR with a word error rate of 4.63%. The layout segmentation of slide images is performed using CSSNet [18] trained on SPaSe [25] and WiSe [15] dataset and obtaining the MIoU of 56.4%. Finally, the figure classes are identified by training the Multi-feature head model [19] using DocFig [17] dataset and obtained an accuracy of 97.85%. After annotation, the train, validation, and test have 5607, 1557, and 1487 slide images with figures, respectively.

4. The Proposed Retrieval System

We aim to retrieve the most appropriate slide image from the dataset given a query. Let $\mathcal{K} = \{(a_j, b_j)\}_{j=1}^{N}$ be the dataset consisting of a description of slide a_j and the slide image b_j. The description $a_j = (t, s)$ combines text description t and the sketch description s. Hence, the system supports both natural language and hand-drawn sketch queries. The goal is to learn an embedding space that can quantify the similarity between the slide image and description. As a result, given a description (text or sketch, or both) a_j, one could retrieve its similar slide images from $\{b_1, b_2, \ldots, b_N\}$.

We propose a novel slide image retrieval system, as shown in Figure 4. We first describe the layout, figure type, and text extraction from lecture slide images in Section 4.1). The lecture slide encoder that encodes the layouts, figures, and texts along with the slide image is described in Section 4.2. The query text and query sketch encoder are explained in Section 4.3, and finally, train the model with Multiple Instance Learning (MIL) [8] framework (Section 4.4), and provide the inference details.

4.1. Semantic Labelling of Lecture Slides

The query to retrieve a slide image need not contain all the text on the slide image. The query contains the keywords and logical regions of the slide image, such as title, enumeration, table, figures, and its various types. For example, "*system of equations explained using enumeration and a line graph*." In order to handle these queries, we propose a novel slide image retrieval system, as shown in Figure 4. Our approach utilizes a pre-trained slide image segmentation module [18] and a multi-feature head model [19] trained on DocFig [17] dataset to obtain the logical regions t_l and the type of the figure present on the slide image t_g, respectively. In addition to these modules, we also use an Optical Character Recognizer (OCR) to extract the text t_o the slide image.

4.2. Lecture Slide Image Data Encoder

In the dataset indexing, we collect the logical regions t_l, figure type t_g, and OCRed text t_o information along with the slide image b and a focus area b_f from each slide image. The focus area of a slide image is the most plausible logical region where the keyword is occurring. Most of the query keywords from our study are from the following logical regions: title, enumeration, paragraphs, and captions. In our proposed architecture, we choose the title area as a focus area. We choose the enumeration region if the title region is absent on the slide. We obtained these logical regions from the output of CSSNet [18].

Further, encoded feature vector z^x for a given slide image b is obtained using Vision and Language Transformer (ViLT) [20] as follows:

$$z^x = \text{ViLT}\left([t_l; t_g; t_o], [b; b_f]\right). \tag{1}$$

4.3. Query Feature Extraction

In our experiment, the query can be either a slide summary, sketch image, or a combination of both. The words in the text query t are encoded using pre-trained BERT model [7] and used as local features $\Psi(t) \in \mathbb{R}^{L\times 300}$ where L is a number of words in t. Then, we feed the local feature to a bi-GRU with H hidden units and take the final hidden states as global features $\phi(t) \in \mathbb{R}^H$. The query sketch image s is encoded using ResNet-152 [16]. The feature map before the final average pooling layer as local features $\Psi(s) \in \mathbb{R}^{7\times 7\times 2048}$. Further, we apply average pooling and

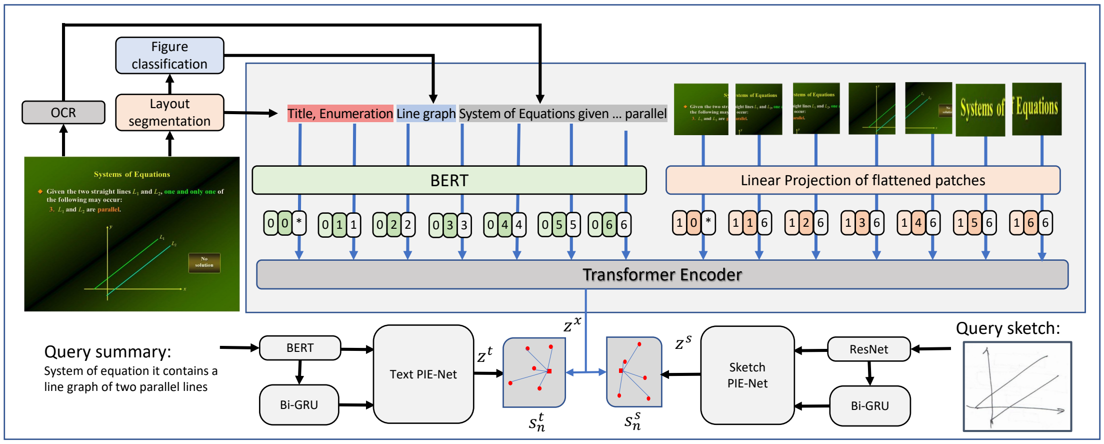

Figure 4. The proposed Lecture Slide Deck Search Engine architecture. The LecSD dataset is indexed by encoding features such as layout regions, figure classes, OCR text, slide image, and focus area using a ViL transformer. The query text and the sketch are independently encoded using PIE-Net [29]. The architecture predicts the final retrieval result based on the similarity score s_n^t and s_n^s. (Best viewed in color)

feed the output to one fully connected layer to obtain global features $\phi(s) \in \mathbb{R}^H$.

The PIE-Net proposed in [29] encodes the local and global features for text and sketch queries.

$$z^t = \text{textPIE-Net}(\Psi(t), \phi(t)), \tag{2}$$

$$z^s = \text{sketchPIE-Net}(\Psi(s), \phi(s)). \tag{3}$$

4.4. Optimization and Inference

We optimize our model to minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{mil} + \lambda_1 \mathcal{L}_{mmd} + \lambda_2 \mathcal{L}_{div}. \tag{4}$$

Where λ_1 and λ_2 are the scalar weights. $\mathcal{L}_{mil}$ is the multi instance learning (MIL) loss [8] with the learning constraint for retrieval task. The loss function $\mathcal{L}_{mil}$ only considers the minimum distance pair in the loss computation. Hence, the distribution induced by features may diverge quickly. Maximum Mean Discrepancy (MMD) [12] based loss $\mathcal{L}_{mmd}$ is introduced to regularize the discrepancy between the two distributions. The $\mathcal{L}_{div}$ is the diversity loss to ensure that the PIE-Net produces diverse representations of an instance. We follow these loss calculations described in [29].

In the training slide images of LecSD dataset, all the slide images do not contain figures. Hence, we first train the ViLT and the textPIE-Net models. Finally, the sketchPIE-Net models learn during the fine-tuning of the whole network with slide images having both summary and sketch queries.

In the inference stage, we assume that the dataset contains $\mathcal{N}$ lecture slide images. Further, the ViLT encoded vectors for i^{th} slide are represented as z_i^x. Now, given a query instance of slide summary, and sketch image, we calculate an embedding vector z^t and z^s, respectively. Then, the similarity between the query and $i = 1^{st}$ to $\mathcal{N}^{th}$ lecture slides are computed as follows:

$$s_n^t = [sim(z^t, z_1^x), \cdots, sim(z^t, z_{\mathcal{N}}^x)], \tag{5}$$

$$s_n^s = [sim(z^s, z_1^x), \cdots, sim(z^s, z_{\mathcal{N}}^x)], \tag{6}$$

$$s_n = \frac{1}{2}(s_n^t + s_n^s). \tag{7}$$

Here, s_n^t, s_n^s, and s_n are the similarity score of query text, sketch, and the combined respectively with the dataset of $\mathcal{N}$ instances. We then rank the database images with respect to these similarities.

5. Experiments

We train the proposed framework using train data obtained using automatic annotations and then assess its performance on manually annotated test data. Our dataset comprises two queries for each slide image: in the training set, there are both a generated summary and a paraphrased version of it, while in the testing and validation sets, we have the manually annotated summary and a paraphrase generated using ChatGPT [2]. During both training and testing phases, we randomly select a sentence from a query associated with a slide image. Please note that text queries are derived from either manually annotated (50% of the time) or paraphrased summaries.

Methods	*Data structure*				*Computer networks*				*Optimization*			
	@1	@5	@10	Median	@1	@5	@10	Median	@1	@5	@10	Median
Random	0.01	0.05	0.1	5000	0.05	0.25	0.5	1000	0.05	0.25	0.5	1000
PCME [5]	8.76	24.42	40.30	39	6.65	18.64	27.56	56	6.21	14.32	22.21	173
CLIP [28]	9.61	27.47	42.50	31	-	-	-	-	-	-	-	-
PVSE [29]	21.13	43.07	51.7	9	8.87	22.01	30.33	41	7.51	17.55	24.18	132
PolyViLT [21]	22.24	44.31	53.05	8	12.45	28.3	37.54	33	9.34	23.34	31.39	84
Ours	**26.45**	**48.53**	**56.82**	**6**	**16.54**	**34.2**	**42.74**	**19**	**12.53**	**28.65**	**35.44**	**60**

Table 2. The summary-based slide image retrieval performance of various slide image retrieval models that were trained on slide images from the *data structure* topics. We show evaluation results on slides related to the topics of *Data structure*, *computer networks*, and *optimization*.

Features	Summary to Slide			
	@1	@5	@10	Median
t_o, b	22.35	44.31	53.05	8
t_o, $[b; b_f]$	23.28	45.83	53.98	8
$[t_l; t_g; t_o]$, b	25.74	47.47	55.38	7
$[t_l; t_g]$, $[b; b_f]$	24.89	46.45	54.00	7
$[t_l; t_g; t_o]$, $[b; b_f]$	26.45	48.53	56.82	6

Table 3. Comparison study on the contribution of various features such as OCR text t_o, layout segmentation t_l, graphics type t_g, slide image b, and the focus area b_f on the retrieval task.

5.1. Implementation Details

We employ the Adam optimizer with a learning rate set at 2^{-4} and a weight decay of 10^{-2}. In the case of ViLT, we resize the shorter edge of input images to 384 and constrain the longer edge to be under 640 while maintaining the aspect ratio. The Patch projection of ViLT-B/32 results in $12 \times 20 = 240$ patches for an image with a resolution of 384×640. We interpolate V^{pos} of ViLT-B/32 to match the size of each image and pad the patches for batch training. The hyper-parameters λ_1 and λ_2 fall within the range of [0.01, 0.001]. We utilize the *bert-base-uncased* tokenizer and initiate the learning of textual embedding-related parameters t_{class}, T, and T^{pos} from scratch. The model undergoes training for 225K steps on four 64-bit NVIDIA GPUs, with a batch size set at 8.

5.2. Lecture Slide Image Retrieval using Natural Language Summary-based Query

To identify the contribution of various features, i.e., OCR text (t_o), layout region (t_l), graphics type (t_g), slide image (b), and the focus area (b_f) for the task, we conduct the ablation study and the result is shown in Table 3. First, we used the t_o and b features in training and obtained 22.35% recall at one. Then, we added the b_f feature that improved recall at one by 0.93%. Next, we combine $[t_l; t_g]$ with t_o and b which further improves recall at one compared to the base model by 3.39%. This indicates that the layout features $[t_l; t_g]$ is an efficient feature for slide retrieval. Further, we removed the t_o, which resulted in the recall reduction of 0.85%. Finally, we used all the features and obtained a 26.45% R@1.

We compare the proposed model on a summary-based slide image retrieval task, and the results are reported in Table 2. The proposed model outperforms the existing cross-model retrieval approaches. The PolyVilt [21], and PVSE [29] models perform reasonably in retrieving the natural images given a caption. However, these models are unable to learn logical regions from slide images. Hence, the performance was reduced, indicating that layout segmentation and document figure classification modules are essential for slide image retrieval tasks.

We further assessed the generalization ability of the proposed model. We conducted this evaluation by testing the model in a "zero-shot setting", i.e. using the model that is trained on slide images associated with the *data structure* topic to evaluate on two different sets of slide images related to *computer networks* and *optimization*. We obtain recall@1 of 16.54 and 12.53 in retrieving slide images from the computer network and optimization courses, respectively. Although there is a performance drop under this setting, our model continues to demonstrate superior performance when compared to other baselines (Refer Table 2).

Figure 5 shows the qualitative result of the proposed approach. The system retrieves the slide image having pseudo-code and block diagram in the first and second rows. We also show retrieval results when a hand-drawn sketch of a diagram is used as a query in the last row. The ground truth for this query is highlighted in the green border.

5.3. Refining Lecture Slide Image Retrieval using Hand-drawn Sketch Query

In testing, we infer the result based on text summary similarity s_n^t, sketch similarity s_n^s, and the combined similarity s_n; the result is shown in Table 4. Here, we noticed

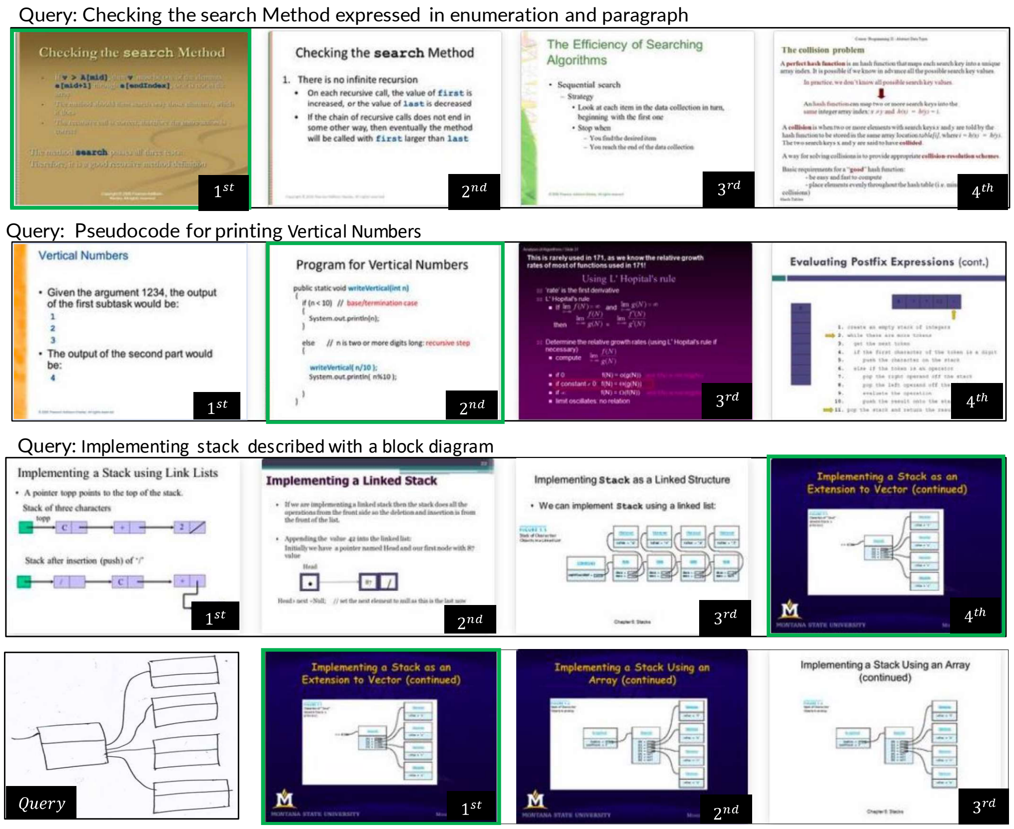

Figure 5. Qualitative result of proposed Lecture Slide Deck Search Engine. The text query and its result are shown in the top three rows. The last row shows the re-ranked result given the sketch query. The slide in the green bounding box indicates the correct image for the query. More qualitative analysis is provided in the supplementary material. (Best viewed in color).

Query type	Slide retrieval			
	@1	@5	@10	Median
Sketch	23.63	51.55	62.72	5.00
Summary	37.05	61.32	65.67	4.00
Combined	41.50	64.00	68.50	2.00

Table 4. Slide image retrieval result when only sketches, only text summary, and their combination are used as queries for retrieving the slide images from *data structure* topics.

that the combined similarity between a sketch and the summary improved the retrieval result to 41.5. The fourth row in Figure 5 shows the sketch query and the successfully re-ranked result, which matches the sketch query. However, the sketch-based retrieval fails for the figure having a smaller size compared to the slide image size.

6. Conclusion

In this paper, we introduced the LecSD - a comprehensive collection of lecture slide decks intended to serve as a benchmark for the development of educational AI systems. The dataset is unique and has rich annotations, and it is specifically created to tackle two challenging research tasks that are relevant to education: i) retrieval of lecture slide images based on brief descriptions that include logical regions and figure classes, and ii) retrieval of lecture slide images using hand-drawn sketches of the figures as queries. Our benchmarking efforts revealed that existing retrieval models fall short of accurately identifying logical regions and figure classes. On the contrary, we proposed an effective new retrieval model called Lecture Slide Deck Search Engine, which is semantic labels-aware and includes sketch-based retrieval functionality. Nonetheless, the Lecture Slide Deck Search Engine does have a few drawbacks. First, it relies on an off-the-shelf layout segmentation module which is far from being perfect. Second, when dealing with sketch queries, it encounters difficulties in retrieving small diagrams. Lastly, the model has limited success in searching slides from unseen subjects. We leave addressing these as the future scope of this paper.

Acknowledgment: Authors would like to thank NLTM and the MeitY, Govt. of India.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 3

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4, 6

[3] Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. Vlengagement: A dataset of scientific video lectures for evaluating population-based engagement. *arXiv preprint arXiv:2011.02273*, 2020. 2, 3

[4] Huizhong Chen, Matthew Cooper, Dhiraj Joshi, and Bernd Girod. Multi-modal language models for lecture video retrieval. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1081–1084, 2014. 2, 3

[5] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021. 3, 7

[6] Sachin G Deshpande and Jenq-Neng Hwang. A real-time interactive virtual classroom multimedia distance learning system. *IEEE Transactions on multimedia*, 2001. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[8] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997. 5, 6

[9] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. 4

[10] Kartik Dutta, Minesh Mathew, Praveen Krishnan, and CV Jawahar. Localizing and recognizing text in lecture videos. In *2018 16th international conference on frontiers in handwriting recognition (ICFHR)*, pages 235–240. IEEE, 2018. 2, 3

[11] Damianos Galanopoulos and Vasileios Mezaris. Temporal lecture video fragmentation using word embeddings. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part II 25*, pages 254–265. Springer, 2019. 2, 3

[12] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006. 6

[13] David Griol and Zoraida Callejas. An architecture to develop multimodal educative applications with chatbots. *International Journal of Advanced Robotic Systems*, 10(3):175, 2013. 2

[14] Anchit Gupta, CV Jawahar, Makarand Tapaswi, et al. Unsupervised audio-visual lecture segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5232–5241, 2023. 2

[15] Monica Haurilet, Alina Roitberg, Manuel Martinez, and Rainer Stiefelhagen. Wise - slide segmentation in the wild. In *ICDAR*, 2019. 2, 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[17] KV Jobin, Ajoy Mondal, and CV Jawahar. Docfigure: A dataset for scientific document figure classification. In *ICDARW*, 2019. 5

[18] KV Jobin, Ajoy Mondal, and CV Jawahar. Classroom slide narration system. In *CVIP*, 2021. 2, 3, 5

[19] KV Jobin, Ajoy Mondal, and CV Jawahar. Document image analysis using deep multi-modular features. *SN Computer Science*, 4(1):5, 2022. 5

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2, 5

[21] Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. Multimodal lecture presentations dataset: Understanding multimodality in educational slides. *arXiv preprint arXiv:2208.08080*, 2022. 2, 3, 7

[22] Irene Li, Alexander R Fabbri, Robert R Tung, and Dragomir R Radev. What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6674–6681, 2019. 2, 3

[23] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. Photo-sketching: Inferring contour drawings from images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1403–1412. IEEE, 2019. 4, 5

[24] Debabrata Mahapatra, Ragunathan Mariappan, Vaibhav Rajan, Kuldeep Yadav, and Sudeshna Roy. Videoken: Automatic video summarization and course curation to support learning. In *Companion Proceedings of the The Web Conference 2018*, pages 239–242, 2018. 2

[25] Ziad Al-Halah Monica Haurilet and Rainer Stiefelhagen. SPaSe - Multi-Label Page Segmentation for Presentation Slides. In *WACV*, 2019. 2, 5

[26] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. *arXiv preprint arXiv:1810.00319*, 2018. 3

[27] Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. Don't blame the annotator: Bias already starts in the annotation instructions. *arXiv preprint arXiv:2205.00415*, 2022. 4

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 3, 7

[29] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, 2019. 2, 3, 6, 7

[30] Nhu Van Nguyen, Mickal Coustaty, and Jean-Marc Ogier. Multi-modal and cross-modal for lecture videos retrieval. In *2014 22nd International Conference on Pattern Recognition*, pages 2667–2672. IEEE, 2014. 2, 3

[31] Zhigang Zhu, Chad McKittrick, and Weihong Li. Virtualized classroom-automated production, media integration and user-customized presentation. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004. 2