TACTFUL: A Framework for Targeted Active Learning for Document Analysis

Venkatapathy Subramanian$^{1[0000-0002-4851-628X]}$, Sagar Poudel$^{1[0009-0008-0891-2115]}$, Ganesh Ramakrishnan$^{1[0000-0003-4533-2490]}$, and Parag Chaudhuri$^{1[0000-0002-1706-5703]}$

Indian Institute of Technology Bombay, Maharashtra, India 400076
{venkatapathy,sagar.poudel,ganesh,paragc}@cse.iitb.ac.in

Abstract. Document Layout Parsing is an important step in an OCR pipeline, and several research attempts toward supervised, and semi-supervised deep learning methods are proposed for accurately identifying the complex structure of a document. These deep models require a large amount of data to get promising results. Creating such data requires considerable effort and annotation costs. To minimize both cost and effort, Active learning (AL) approaches are proposed. We propose a framework TACTFUL for Targeted Active Learning for Document Layout Analysis. Our contributions include (i) a framework that makes effective use of the AL paradigm and Submodular Mutual Information (SMI) functions to tackle object-level class imbalance, given a very small set of labeled data. (ii) an approach that decouples object detection from feature selection for subset selection that improves the targeted selection by a considerable margin against the current state-of-the-art and is computationally effective. (iii) A new dataset for legacy Sanskrit books on which we demonstrate the effectiveness of our approach, in addition to reporting improvements over state-of-the-art approaches on other benchmark datasets.

Keywords: Active Learning · Submodular Functions · Balancing Dataset · Document Layout Analysis

1 Introduction

Digitization of scanned documents such as historical books, papers, reports, contracts, *etc.*, is one of the essential tasks required in this information age. A typical digitization workflow consists of different steps such as pre-processing, page layout segmentation, object detection, Optical Character Recognition (OCR), post-processing, and storage. Though OCR is the most important step in this pipeline, the preceding steps of page layout segmentation and object detection play a crucial role. This is especially so when there is a requirement to preserve the layout of the document beyond OCR. For over two decades, the scientific community has proposed various techniques [2] for document layout analysis, yet recent deep-learning methods have attained improved performance by leaps and bounds. Several supervised, and semi-supervised deep learning methods [17, 16, 18] can accurately identify the complex structure of a document. This performance improvement, though, comes at a cost. Deep models require a significantly large amount of data to yield promising results. Creating such data requires considerable annotation effort and cost. To minimize both cost and effort, active learning (AL) approaches [19] are proposed with a constraint on the annotation budget. AL can help iteratively select a small amount of data for annotation, from a large unlabeled pool of data, on which the model can be trained at each iteration. Though such approaches work, sometimes page-level AL techniques may be biased and miss out on rare classes while selecting images. It is especially true in the case of document layout analysis where document objects (such as titles, images, equations, *etc.*) are complex, dense, and diverse. It can be somewhat challenging to select pages that lead to balance across classes in the training set and specially balanced performance across all. Most state-of-the-art AL approaches tend to decrease the models' performance on rare classes in the pursuit of overall accuracy.Summarily what we wanted is an effective AL technique that can help address the class imbalance problem while selecting page images. Towards this, we propose a framework for **T**argeted **ACT**ive Learning **F**or Doc**U**ment Layout Ana**L**ysis (TACTFUL). The proposed framework uses sub-modular mutual information (SMI) functions in its active learning strategies to select relevant data points from a large pool of unlabeled datasets. The SMI (Submodular Mutual Information) functions are useful in two complementary ways: 1) By taking advantage of the natural diminishing returns property of sub-modular functions, the framework maximizes the utility of subsets selected during each AL round. 2) By quantifying the mutual dependence between two random variables (a known rare class and an unknown object of interest in our case), we can maximize the mutual information function to get relevant objects and through them the page images for annotation. Our contribution, through the TACTFUL framework, is as follows:

1. We propose an end-to-end AL framework that can tackle object-level class imbalance while acquiring page images for labeling, given a very small set of labeled data. Within this framework, we make effective use of two complementing paradigms, *viz.*, i) AL paradigm that aims to select a subset of samples with the highest value from a large set, to construct the training samples, and ii) Submodular functions that have higher marginal utility in adding a new element to a subset than adding the same element to a superset. Within submodular functions, we use SMI functions [4, 7] that can model the selection of subsets similar to a smaller query set from rare classes thereby avoiding severe data imbalance.
2. For subset selection, we decouple the object detection and feature selection steps thereby overcoming the limitation [3] present in current object detection models. We show that the pre-trained model, without additional fine-tuning, works effectively well for representing objects. The decoupling strategy improves the targeted selection by a considerable margin against the current state-of-the-art and is computationally effective.

3. We empirically prove that our model performs well compared to the current SOTA framework having an increase of about 9.3% over the SOTA models. We also release a new dataset for document layout analysis for legacy Sanskrit books and show that our framework works for similar settings for documents in other languages and helps improve the AP by 8.6 % over the baseline.

2 Motivation

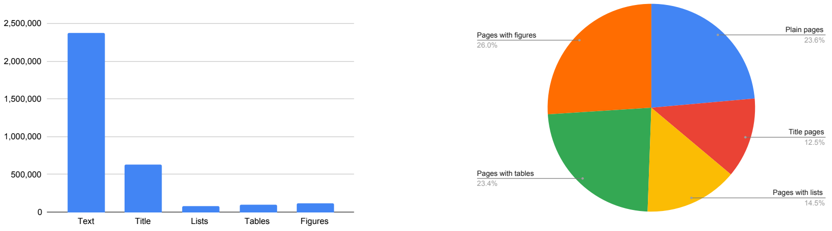
Layout detection is a continuing challenge in the field of Document Analysis. Many state-of-the-art models and datasets [23, 11, 1] have been created to address this problem. The recent availability of a large amount of annotated data has resulted in good-quality ML models for layout detection for English documents. There is still dearth of good quality ground truth data for other languages. This dearth is owing to a combination of the following reasons: (i) Manual labeling is time-consuming and expensive (ii) It is difficult, if not impossible, to replicate the alternative ways of creating ground truth data such as those created for the English documents. Large datasets such as Publaynet [23] and DocBank [11] are created in a weakly supervised manner by extracting the ground-truth layout information from an available parallel corpus such as scientific latex documents [5] and then manually post-editing or correcting the output. Such availability is rare in other languages. Hence the need for large annotated datasets for other languages remains unaddressed. One case study we consider is the digitization of ancient Sanskrit documents where there is an immediate need for high-quality document layout detection methods. In this aforementioned work, we have tens of thousands of scanned images that are old manuscripts. Thus, the only way of training a good model for layout detection is to annotate a subset of the available pages. The question that subsequently arises is *'How can we identify pages to be annotated such that the model performance improves across all classes?'*

As an attempt to answer the question, we performed a retrospective study on one of the largest datasets available for document layout analysis, *viz.*, PubLayNet [23]. The dataset was created by automatically matching the XML representations and the content of over 1 million PDF articles publicly available on PubMed Central. It contains over 360 thousand annotated document images and the deep neural networks trained on PubLayNet achieve an Average Precision (AP) of over 90%. The dataset consists of five categories of annotations and those include Title, Title, List, Table, and Figure. The statistics of the layout categories associated with this dataset are summarized in Table 1:

In Figure 1a, we depict the share of each class amongst the total objects present in the training dataset of PubLayNet. From the chart, we can notice that there is a class imbalance for the different objects. Given this imbalanced class distribution, we initially investigate how an object detection model learns. We randomly sampled a fraction of about 2000 data points from the dataset (from the official train, and test sets), while maintaining the same class imbalance as in the original set. In those 2000 pages, there were a total of 9920 Title objects,

	Training	Development	Testing
Pages			
Plain pages	87,608	1,138	1,121
Titlepages	46,480	$2,059^+$	$2,021^+$
Pages with Lists	53,793	2,984	3,207
Pages with Tables	86,950	$3,772^+$	$3,950^+$
Pages with Figures	96,656	$3,734^+$	$3,807^+$
Total	340,391	11,858	11,983
Instances			
Title	$2,376,702$	93,528	95,780
Title	633,359	19,908	20,340
Lists	81,850	4,561	5,156
Tables	103,057	4,905	5,166
Figures	116,692	4,913	5,333
Total	$3,311,660$	127,815	131,775

Table 1: Statistics of layout categories in the training, development, and testing sets of PubLayNet. PubLayNet is one or two orders of magnitude larger than any existing document layout dataset obtained from [23]



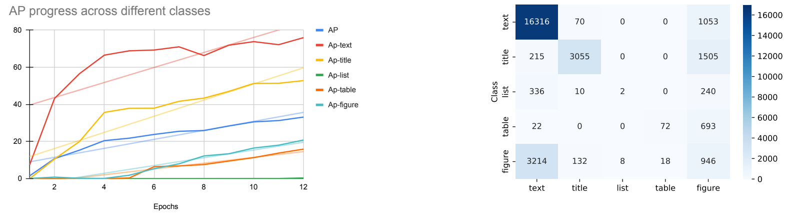
(a) Distribution of class objects

(b) Pagewise spread of different class objects

Fig. 1: Figures depicting the share of objects in the training dataset.

900 Title objects, 74 List objects, 76 Table objects, and 123 Figure objects. We subsequently trained a Faster RCNN model from scratch. The model reached an AP of 33% before plateauing. Figure 2a shows the improvement of average AP and classwise AP during the training. The AP of Title is the highest with 75% followed by Title with 50%. The worst performing classes are also the rare classes with the Figure class averaging 20%, Table 15%, and List having the worst performance with just 0.3%. It's also interesting to note that Table and Figure with counts as of List performs better. It could be that it's easier to detect Figure and Table objects due to their unique shapes compared to the rest of the page objects. The average AP is drastically reduced by the last three tail objects of the distribution. This is expected, given the skewed distribution of objects in the dataset, as the model might not have seen the rare class objects. We later selected just 500 data points but maintained the balance among the

classes. This time there were a total of 2583 Title objects, 231 Title objects, 14 List objects, 17 Table objects, 27 Figure objects. Again we trained a Faster RCNN model from scratch. Figure 2a shows the improvement of average AP and classwise AP during the training. The model reached an AP of 48% before plateauing. The difference in AP between the model trained with 2000 data points with a severely imbalanced set and the model trained on 1/4 of the former(500 data points) but without class imbalance, is large with a balanced dataset performing better. **There is then the scope for improving the AP in the same setting if there is a way to select images in such a way that the rare classes are covered and the class imbalance is avoided.**



(a) AP and classwise AP during training of 2000 data points with class imbalance

(b) Confusion Matrix for the five different classes as classified by models trained on 2000 data points

Fig. 2: Figure depiciting the AP, classwise AP and confusion matrix for different classes trained on 2000 data points

We further plotted the confusion matrix for the object detected on the initial dataset of 2000 points. The model trained on this initial dataset had good object detection accuracy, as depicted in Figure 2b. but performs poorly in the object classification task. Empirically it is clear that training a model for both classification and object detection compromises the model's ability to learn efficient features in the embedding space [3], which can be observed in our case as well. Given these observations, our goal is to mitigate the class imbalance during image acquisition for annotations, thereby improving the performance of the model for rare classes. We explain our framework in detail in the next section.

3 Our Approach

Our Active Learning (AL) paradigm[14] approach to address the problem of class imbalance discussed in Section 2 is visually depicted in Figure 3. Similar to standard AL techniques[17] we train a detection model Θ to detect n layout objects of an image X_i. An object n_j consists of the bounding box b_j and its category c_j. $Y_i = \{(b_j, c_j)\}_{j=1}^{n}$ are the object annotations for X_i. Θ is initially trained on a small labelled dataset $\mathcal{L}_0 = \{(\mathcal{X}_i, \mathcal{Y}_i\}_{i=1}^{l}$ and contains a large set of unlabelled dataset $\mathcal{U}_0 = \{(\mathcal{X}_i\}_{i=l}^{u+l}\}$. Given the unlabeled set $\mathcal{U}$, the goal is to *acquire* a subset of the images $\mathcal{A} \subseteq \mathcal{U}$ of budget $k = \|\mathcal{A}\|$, iteratively, to improve the model's performance. At each iteration t, m samples $\mathcal{M}_t = \{\mathcal{X}_i\}_{i=l}^{m} \subseteq \mathcal{U}_{t-1}$ are queried for labelling and the corresponding labeled set $\mathcal{M}_t = \{(\mathcal{X}_i\, \mathcal{Y}_{i=l}^{m})\} \subseteq \mathcal{U}_{t-1}$ is added to the existing labeled set $\mathcal{L}_t = \{\mathcal{L}_{t-1} \cup \mathcal{M}_t\}$. For the next iteration, the unlabeled set becomes $\mathcal{U}_t = \{\mathcal{U}_{t-1} \setminus \mathcal{M}\}$. Iteration is stopped when the model reaches the desired performance. As described in Section 2, we wanted to select a subset of images from the unlabelled page image distribution (eg from distribution as shown in Figure 1b) such that the model performance improves for selected rare class[es].

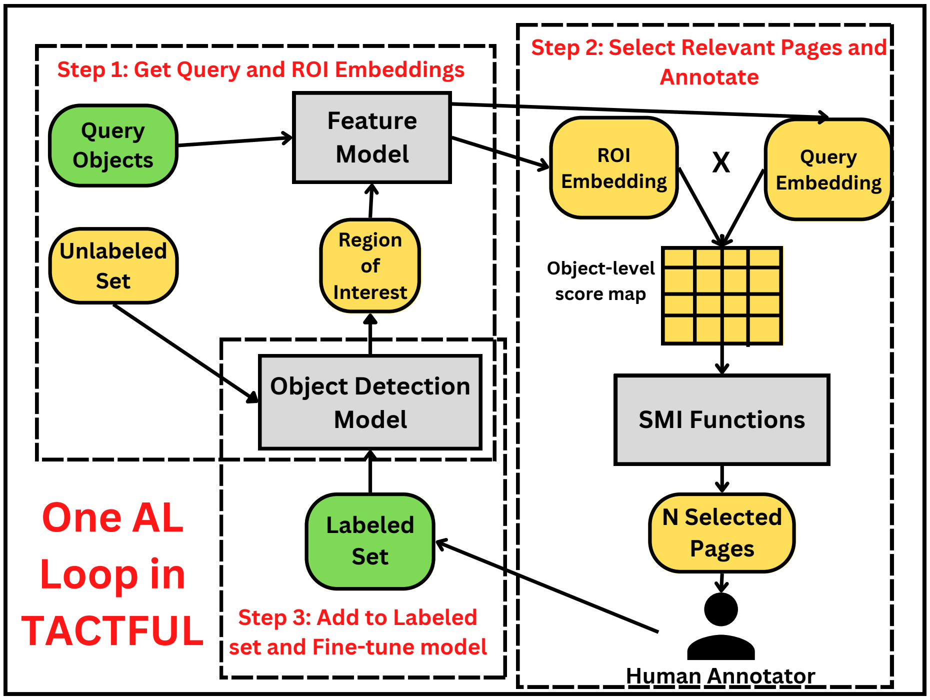


Fig. 3: Three-step approach of one AL loop of TACTFUL

To achieve this, we will use the core idea of SMI functions [21, 9, 8] as acquisition functions to acquire data points from $\mathcal{U}$. In the following section, we discuss some of the notations and preliminaries required for our work as introduced in [7, 6] and their extensions introduced in [9, 8]

3.1 Submodular Functions:

$\mathcal{V}$ denotes the *ground-set* of n data points $\mathcal{V} = \{1, 2, 3, ..., n\}$ and a set function $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$. The function f is submodular [4] if it satisfies diminishing returns, namely $f(j\|\mathcal{X}) \geq f(j\|\mathcal{Y})$ for all $\mathcal{X} \subseteq \mathcal{Y} \subseteq \mathcal{V}, j \notin \mathcal{X}, \mathcal{Y}$. Facility location, graph cut, log determinants, *etc.*, are some examples [21].

3.2 Submodular Mutual Information (SMI):

Given a set of items $\mathcal{A}, \mathcal{Q} \subseteq \mathcal{V}$, the submodular mutual information (SMI) [6, 7] is defined as $I_f(\mathcal{A};\mathcal{Q}) = f(\mathcal{A}) + f(\mathcal{Q}) - f(\mathcal{A} \cup \mathcal{Q})$. Intuitively, SMI measures the similarity between $\mathcal{Q}$ and $\mathcal{A}$ and we refer to $\mathcal{Q}$ as the query set. In our context, $\mathcal{A} \subset \mathcal{U}$ is our unlabelled set of images, and the query set $\mathcal{Q}$ is the small target set containing the rare class images and page annotations for the target set. To find an optimal subset $\mathcal{M} \subseteq \mathcal{U}$, given a query set $\mathcal{Q}$ we can define $g_{\mathcal{Q}}(\mathcal{A}) = I_f(\mathcal{A};\mathcal{Q})$, $\mathcal{A} \subseteq \mathcal{V}$ and maximize the same.

3.3 Specific SMI functions used in this work

For any two data points $i \in \mathcal{V}$ and $j \in \mathcal{Q}$, let s_{ij} denote the similarity between them.
Graph Cut MI (Gcmi): The Smi instantiation of graph-cut (Gcmi) is defined as: $I_{GC}(\mathcal{A};\mathcal{Q}) = 2\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{Q}} s_{ij}$. Since maximizing Gcmi maximizes the joint pairwise sum with the query set, it will lead to a summary similar to the query set Q. Specific instantiations of Gcmi have been used for query-focused summarization for videos [20] and documents [12, 10].
Facility Location MI (Flmi): We consider two variants of Flmi. The first variant is defined over $\mathcal{V}$(Flvmi), the Smi instantiation can be defined as: $I_{FLV}(\mathcal{A};\mathcal{Q}) = \sum_{i\in\mathcal{V}} \min(\max_{j\in\mathcal{A}} s_{ij}, \max_{j\in\mathcal{Q}} s_{ij})$. The first term in the min(.) of Flvmi models diversity, and the second term models query relevance.

For the second variant, which is defined over $\mathcal{Q}$ (Flqmi), the Smi instantiation can be defined as: $I_{FLQ}(\mathcal{A};\mathcal{Q}) = \sum_{i\in\mathcal{Q}} \max_{j\in\mathcal{A}} s_{ij} + \sum_{i\in\mathcal{A}} \max_{j\in\mathcal{Q}} s_{ij}$. Flqmi is very intuitive for query relevance as well. It measures the representation of data points that are the most relevant to the query set and vice versa.
Log Determinant MI (Logdetmi): The Smi instantiation of Logdetmi can be defined as: $I_{LogDet}(\mathcal{A};\mathcal{Q}) = \log\det(S_{\mathcal{A}}) - \log\det(S_{\mathcal{A}} - S_{\mathcal{A},\mathcal{Q}} S_{\mathcal{Q}}^{-1} S_{\mathcal{A},\mathcal{Q}}^T)$. $S_{\mathcal{A},\mathcal{Q}}$ denotes the cross-similarity matrix between the items in sets $\mathcal{A}$ and $\mathcal{Q}$.

3.4 Object-Level Feature Extraction

In [8], the similarity is calculated using the feature vector of Dimension D for T Region of Interest(ROIs) in Query images $\mathcal{Q}$ and P region proposals obtained using a Region Proposal Network(RPN) in Unlabelled images $\mathcal{U}$. Then the dot product along the feature dimension is computed to obtain pairwise similarities between T and P. One problem we encountered with this approach is that extracting the feature from the same model as the one being trained for detection, did not give a good feature representation of objects(see Section 2 and observations for Figure 2b). Through detailed analysis Chen et al., [3] also show that though the object detection networks are trained with additional annotations, the resulting embeddings are significantly worse than those from classification models for image retrieval. In line with observation, we propose decoupling the process by object-level feature selection- detecting objects with object detectors and then encoding them with pre-trained classification models.

3.5 Targeted Active learning for Document layout Analysis(Tactful)

Stitching together the concepts presented in Sections 3.1 through 3.4, we propose our approach called **T**argeted **ACT**ive Learning **F**or Doc**U**ment Layout Ana**L**ysis (**Tactful**). The algorithm is presented in 3.5. In TACTFUL a user can provide a small annotated set that can be used for initial training as well as query object selection; With a partially trained object detection model and pre-trained feature selection model, TACTFUL can be used to do targeted active learning on a new and unknown set of document pages, and can effectively avoid the curse of class imbalance and thus improving the performance of document layout analysis compared to other methods. We did extensive experiments on the proposed framework which is detailed in the following section.

```
Require: $\mathcal{L}_0$: initial Labeled set, $\mathcal{U}$ unlabeled set, $\mathcal{Q}$ under-performing query
    set, $k$: selection budget, $\Theta$ Object-Detection model, $\Phi$ feature selection model(eg
    a Pre-Trained ResNET50)
1: Train model on labeled set $\mathcal{L}_0$ to obtain the model parameters $\Theta_0$ { Obtain
    model macro AP }
2: Evaluate the model and obtain the under-performing class
3: Crop query and unlabeled set into objects set using ground truth and bounding box detection on model $\Theta$. {$\hat{\mathcal{Q}} \leftarrow n \times \mathcal{Q}$, $\hat{\mathcal{U}} \leftarrow m \times \mathcal{U}$ where n and m is
    objects in query and unlabeled set respectively }
4: Compute the embeddings $\left\{\nabla_{\Phi}\mathcal{L}\left(x_i, y_i\right), i \in \hat{\mathcal{U}}\right\}$ and $\left\{\nabla_{\theta_{\varepsilon}}\mathcal{L}\left(x_i, y_i\right), i \in \hat{\mathcal{Q}}\right\}$ {
    Obtain vectors for computing kernel in Step 5
5: Compute the similarity kernels S and define a MI function $I_f(\hat{\mathcal{U}};\hat{\mathcal{Q}}) = f(\hat{\mathcal{U}})+$
    $f(\hat{\mathcal{Q}}) - f(\hat{\mathcal{U}} \cup \hat{\mathcal{Q}})$ {Instantiate Functions }
6: $\hat{\mathcal{A}} \leftarrow \max_{\mathcal{A}\subseteq\mathcal{U},(\mathcal{A})\leq k}\left(I_f(\mathcal{A};\mathcal{T} \mid \mathcal{P}\right)$ {Obtain the subset optimally matching
    the target }
7: Retrace the subset to original data point $\{\bar{\mathcal{A}} \leftarrow \hat{\mathcal{A}}\}$ and obtain the label of
    the element in $\bar{\mathcal{A}}$ as Ł$(\bar{\mathcal{A}})$
8: Train a model on the combined labeled set $\mathcal{L}_i \in$ Ł$(\bar{\mathcal{A}})$ and repeat steps 2 to
    8 until the model reaches the desired performance
```

4 Experiments

We ran experiments for the TACTFUL framework on two datasets. First, we recreated the experiment from Section 2, where we randomly took a fraction of about 2000 data points from the Publaynet while maintaining the class imbalance among objects as in the original set. In the following experiment, we took a budget of 50 objects in each AL round in the Publaynet dataset and a total budget of 2000 pages. For the experiment, we used detectron2 [22] to fine-tune Faster RCNN [15] and train the model. The experiment was done on a system with 120GB RAM and 2 GPUs of 50GB Nvidia RTX A6000. We performed two different types of tests. The first strategy was to dynamically update the query set i.e in each active learning cycle, we found the underperforming object and update the query set $\mathcal{Q}$ with the underperforming objects. In the second test **(Static Query set)**, we selected the under-performing objects after the initial training and performed subset selection. We compare TACTFUL with TALISMAN[8] as the baseline. Here we show the effectiveness of TACTFUL's object-level selection against TALISMAN's object selection which is limited for each image. We also show the effectiveness of using a different model for feature selection. To assess the scalability and generalizability of our proposed approach, we performed a series of experiments by varying the size of the initial training set and the number of sampled pages per cycle. specifically, we utilized 6300 train data points and 10,000 unlabelled data points and allocated 200 and 2000 budgets for active learning and the total budget, respectively.

Through experiments, we show that the performance of our active learning approach is not heavily dependent on the choice of pre-trained models used for image embedding as long as the embeddings accurately capture the image content. Thus, we selected pre-trained models based on their ease of use and availability while ensuring they provide high-quality embeddings. We used a similar setting mentioned above for this experiment.

The paper introduces a new dataset called the **Sanskrit dataset**. The Sanskrit dataset is a collection of images in the Sanskrit language, which is an ancient Indian language of Hinduism and a literary and scholarly language in Buddhism, Jainism, and Sikhism. The dataset contains four types of class objects, Image, Math Table, and Text. The dataset contains 1388 training images, 88 validation images, and 82 test images. Table 2 provides the distribution of layout objects. The images in the dataset were collected from various sources, including Sanskrit textbooks, manuscripts, and art. They cover many topics, including religious texts, philosophical texts, poetry, and art.

Table 2: Statistics of layout categories in the training, development, and testing sets of Sanskrit Dataset

	Training	Validation	Testing
Pages	1356	88	82
Image	263	57	65
Math	2725	340	406
Table	63	24	31
Text	24615	1761	1309
Total	37666	2182	1811

We randomly selected 334 images from the 1388 training images to create the initial training set. The remaining 1054 training images were used as unlabelled data points. We employed an active learning strategy with a total budget of 300 images, allocating 30 for each active learning cycle. We fine-tuned a Faster R-CNN [15] model pre-trained on the COCO dataset [13] for this experiment. We conducted 10 iterations of the active learning process to evaluate our approach.

5 Results

For **Dynamic Query Set** strategy i.e new query set in each active learning cycle experiment, We observed that model performance oscillates for each round. Figure 4a shows the progression of model performance. The reason could be due to constant changes in the query set for each active learning cycle. Table 3 shows the result with respect to different strategies. The GCMI strategy performed better than other strategies. In COM strategy, the Figure object class got less AP maybe due to constant updates to the query set.

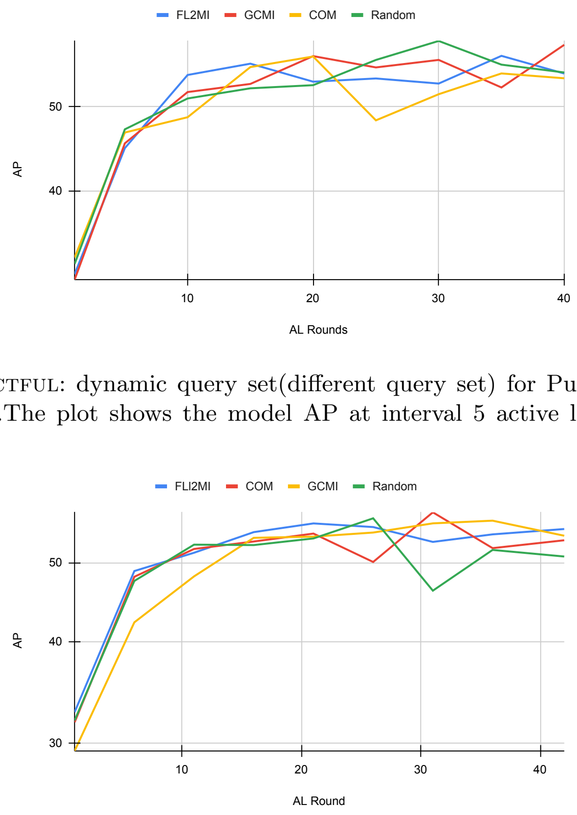
Table 3: AP with respect to different Active learning strategy with dynamic underperforming classes

	AP	AP-Title	AP-Title	AP-List	AP-Table	AP-Figure
Random	50.8703	84.5026	67.2462	38.0769	36.2	28.3259
GCMI	**54.9445**	85.4826	65.5047	50.084	36.3451	**37.3062**
FL2MI	51.1594	83.9618	65.1775	36.1761	**36.6315**	33.85
COM	48.2865	**85.7209**	**67.6509**	**44.5379**	**23.5566**	19.966

In **Static Query Set** strategy, we selected the worst performing class after initial training and the query set class is fixed for all AL rounds. Figure 4b shows that the model oscillates lesser than before, when adding a new data point to the trained labeled set. Table 4 shows the result with respect to different strategies. In this experiment, FL2MI performed better than all other strategies. In this experiment, we selected the List category as a query set. FL2MI gets more AP-List than random strategy and beats it by $\tilde{1}0\%$.

We compared Tactful with TALISMAN[8] as a baseline. In TALISMAN, the same feature model is used for selecting the image features. As can be seen from Table 5, Tactful has outperformed TALISMAN in all three strategies. This validates our proposal that decoupling detection and feature selection models have a considerable impact on SMI strategies.

Figure 5 shows the cumulative sum of rare classes augmented to the trained labeled set ($\mathcal{L}_i$) in each active learning round for dynamic query set. It can be seen that GCMI and FLMI strategies take twice as many objects from rare classes as random ones. This shows the effectiveness of our approach to tackling class imbalance during AL data acquisition.



(a) Tactful: dynamic query set(different query set) for Publaynet dataset.The plot shows the model AP at interval 5 active learning rounds.

(b) Tactful: static query set(list) for Publaynet dataset. The plot show the model AP at interval 5 active learning rounds. The model oscillates less compare to dynamic query list

Fig. 4: AP with respect to Dynamic Query and Static Query set

Table 4: AP with respect to different Active learning strategy with static rare classes

	AP	AP-Title	AP-Title	AP-List	AP-Table	AP-Figure
Random	50.8703	84.5026	**67.2462**	38.0769	36.2	28.3259
GCMI	53.949	83.90303	64.8621	39.76392	**45.3594**	35.8586
FL2MI	**55.2586**	84.482	66.0408	**48.9131**	35.348	**43.43945**
COM	53.26239	**84.68513**	63.2735	47.98171	37.80424	31.635

To evaluate our active learning approach, we conducted extensive experiments with varying initial training set sizes and different numbers of sampled pages per cycle. In the second set of experiments, we evaluated the impact of dif-

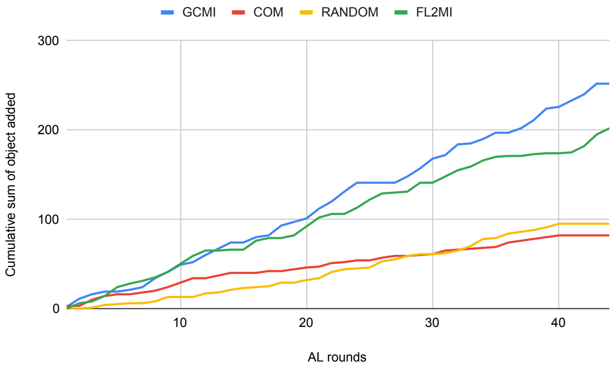


Fig. 5: Cumulative Target Object Added vs AL rounds

Table 5: AP with respect to TALISMAN[8] and TACTFUL

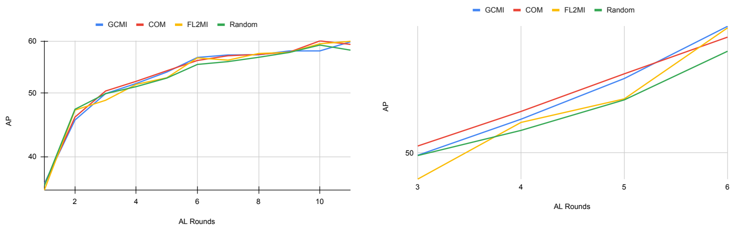
	FL2MI	GCMI	COM	Random
TADA	56.4369	55.2120	55.3783	51.64246
Talisman	51.6316	52.8814	53.0556	51.64246

ferent pre-trained models on the performance of our active learning framework. Table 6 gives the outcome of the experiment, and fl2mi gives the most significant score. A similar result can be observed with multiple pre-trained models.

Table 6: AP with respect to larger train data points and multiple pre-trained model

	FL2MI	GCMI	COM	Random
RESNET101	**79.46**	78.989	78.71	78.26
RESNET50	**79.211**	78.357	78.526	78.26
RESNET18	**79.196**	78.8507	78.5441	78.26

Finally, we experimented on **Sanskrit Dataset** that contains fewer data points as compared to large corpora like Docbank [11] and Publaynet [23]. 7 shows the result with respect to the Sanskrit dataset. All SMI strategies [4, 7, 21] performed better than the random strategy. Among all SMI strategies, GCMI gave the best result. Figure 6a shows the training plot for the Sanskrit dataset and we can see that all SMI functions give better results compared to the random function.



(a) TactfulFor the Sanskrit dataset for different AL functions with static query set

(b) Zoomed graph depicting AP between 3 to 6 AL rounds. Demonstrates SMI functions work better than random functions.

Fig. 6: AP with respect to Static Query set for the Sanskirt Dataset

Table 7: AP with respect to Sanskrit dataset using TADA

	FL2MI	GCMI	COM	Random
Sanskrit Dataset	49.4122	51.7727	49.1080	48.2420

6 Related Work

In our research, we build upon the work presented in PRISM [9], which addresses two specific subset selection domains: 1) Targeted subset selection and 2) Guided Summarization. PRISM employs a submodular function to determine the similarity between the query set and the lake setting, using distinct submodular function variations to solve the two types of subset selection problems. Our contribution enhances the approach proposed in PRISM and its extension in TALSIMAN [8] by incorporating the object embedding decoupling strategy.

Another related study, by Shekhar et al. [16], focuses on learning an optimal acquisition function through deep Q-learning and models active learning as a Markov decision process. The primary distinction between our framework and theirs is that their approach necessitates pre-training with an underlying dataset. At the same time, our method can be utilized without any pre-training requirements.

Shen et al.'s OLALA [17] attempt to leverage human annotators solely in areas with the highest ambiguity in object predictions within an image. Although they operate at the object level, images are still randomly selected. We propose that our targeted selection approach can be integrated into the OLALA framework, potentially enhancing the efficiency of both methodologies. This integration, however, remains a topic for future research. Furthermore, an additional related work, "ActiveAnno: General-Purpose Document-Level Annotation Tool with Active Learning Integration" [**?**], presents a versatile annotation tool designed for document-level tasks, integrating active learning capabilities. This tool aims to reduce the human effort required for annotation while maintaining high-quality results. It achieves this by identifying and prioritizing the most informative instances for annotation, thus optimizing the use of human expertise during the annotation process.

Our framework could potentially benefit from incorporating elements of ActiveAnno's approach to streamline the targeted selection and annotation process further. By combining our targeted selection methodology with ActiveAnno's active learning integration, we may improve both systems' overall efficiency and effectiveness. This potential integration and its implications warrant further exploration and experimentation in future research.

7 Conclusion

In this paper, we propose a **T**argeted **ACT**ive Learning **F**or Doc**U**ment Layout Ana**L**ysis (TACTFUL). That mitigates the class imbalance during image acquisition for annotations, thereby improving the model's performance for rare classes. Through different experiments, we show that our framework significantly improves the model accuracy compared to random, relative to the object level. We also decoupled the model and showed that it can perform better than TALISMAN. This approach can be used with language with fewer data points to improve accuracy.

Acknowledgements We acknowledge the support of a grant from IRCC, IIT Bombay, and MEITY, Government of India, through the National Language Translation Mission-Bhashini project.

Bibliography

[1] Abdelrahman Abdallah, Alexander Berendeyev, Islam Nuradin, and Daniyar Nurseitov. Tncr: Table net detection and classification dataset. *Neurocomputing*, 473:79–97, 2022.

[2] Galal M Binmakhashen and Sabri A Mahmoud. Document layout analysis: a comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36, 2019.

[3] Bor-Chun Chen, Zuxuan Wu, Larry S Davis, and Ser-Nam Lim. Efficient object embedding for spliced image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14965–14975, 2021.

[4] Satoru Fujishige. *Submodular functions and optimization.* Elsevier, 2005.

[5] Paul Ginsparg. Arxiv at 20. *Nature*, 476(7359):145–147, 2011.

[6] Anupam Gupta and Roie Levin. The online submodular cover problem. In *ACM-SIAM Symposium on Discrete Algorithms*, 2020.

[7] Rishabh Iyer, Ninad Khargoankar, Jeff Bilmes, and Himanshu Asnani. Submodular combinatorial information measures with applications in machine learning. *arXiv preprint arXiv:2006.15412*, 2020.

[8] Suraj Kothawade, Saikat Ghosh, Sumit Shekhar, Yu Xiang, and Rishabh Iyer. Talisman: Targeted active learning for object detection with rare classes and slices using submodular mutual information. *arXiv preprint arXiv:2112.00166*, 2021.

[9] Suraj Kothawade, Vishal Kaushal, Ganesh Ramakrishnan, Jeff Bilmes, and Rishabh Iyer. Prism: A rich class of parameterized submodular information measures for guided subset selection. *arXiv preprint arXiv:2103.00128*, 2021.

[10] Jingxuan Li, Lei Li, and Tao Li. Multi-document summarization via submodularity. *Applied Intelligence*, 37(3):420–430, 2012.

[11] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*, 2020.

[12] Hui Lin. *Submodularity in natural language processing: algorithms and applications.* PhD thesis, 2012.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[14] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[16] Sumit Shekhar, Bhanu Prakash Reddy Guda, Ashutosh Chaubey, Ishan Jindal, and Avneet Jain. Opad: An optimized policy-based active learning framework for document content analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2826–2836, 2022.

[17] Zejiang Shen, Jian Zhao, Melissa Dell, Yaoliang Yu, and Weining Li. Olala: Object-level active learning for efficient document layout annotation. *arXiv preprint arXiv:2010.01762*, 2020.

[18] Hao-Yue Sun, Ying Zhong, and Da-Han Wang. Attention-based deep learning methods for document layout analysis. In *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, pages 32–37, 2022.

[19] Alaa Tharwat and Wolfram Schenck. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics*, 11(4):820, 2023.

[20] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 582–590, 2017.

[21] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pages 1954–1963. PMLR, 2015.

[22] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[23] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019.