OPAD: An Optimized Policy-based Active Learning Framework for Document Content Analysis

Sumit Shekhar  
Adobe Research  
sushekha@adobe.com

Bhanu Prakash Reddy Guda  
Adobe Research  
guda@adobe.com

Ashutosh Chaubey  
IIT Roorkee  
achaubey@cs.iitr.ac.in

Ishan Jindal  
IIT Roorkee  
ijindal@ec.iitr.ac.in

Avneet Jain  
IIT Roorkee  
ajain1@ee.iitr.ac.in

Abstract

Documents are central to many business systems, and include forms, reports, contracts, invoices or purchase orders. The information in documents is typically in natural language, but can be organized in various layouts and formats. There have been recent spur of interest in understanding document content with novel deep learning architectures. However, document understanding tasks need dense information annotations, which are costly to scale and generalize. Several active learning techniques have been proposed to reduce the overall budget of annotation while maintaining the performance of the underlying deep learning model. In this paper, we propose OPAD, a novel framework using reinforcement policy for active learning in content detection tasks for documents. The proposed framework learns the acquisition function to decide the samples to be selected while optimizing performance metrics that the tasks typically have. Furthermore, we extend to weak labelling scenarios to further reduce the cost of annotation significantly. We propose novel rewards to account for class imbalance and user feedback in the annotation interface, to improve the active learning method. We show superior performance of the proposed OPAD framework for active learning for various tasks related to document understanding like layout parsing, object detection and named entity recognition. Ablation studies for human feedback and class imbalance rewards are presented, along with a comparison of annotation times for different approaches.

1. Introduction

Documents are a key part of several business processes, which can include reports, business contracts, forms, agreements, etc. Extracting data from documents through deep networks have recently started gaining attention. These tasks include document page segmentation, entity extraction or classification. Fueled by the availability of both labeled and unlabeled data, and advances in the computation infrastructure, recently, a number of deep learning models have been proposed for modeling complex tasks [12,23,39]. The promising results from this research direction motivated development of several deep learning models which show significant performance improvements on these tasks when trained on a large amount of labelled data [35,53,55]. However, deployment of these models requires considerable effort and cost to annotate unlabeled data especially for document tasks because of requirements for dense annotations, e.g. annotating page structures with components like title, table, figures or references. Thus, there is a need to explore methods to optimize annotation budgets to accelerate the development of document analysis models.

Several approaches have been proposed in the domain of semi-supervised learning [56], unsupervised learning [52], few-shot learning [51], active learning [42] etc. to overcome the limitation of availability of labeled data. Each of these approaches have their own objectives incorporated in either modeling or data annotation or both for achieving superior performances in a limited annotated data setup. Among these, our motivation for using active learning is two-folds: (1) active learning bridges the gap in the model by querying samples in the data space, for which the model does not have enough information [42], (2) the active learning approaches seek to learn higher accuracy models within a given annotation cost, through optimizing data acquisition, which align well with our objective of optimizing annotation costs. Recent methods for pool-based active learning scenario, the query for annotations selects a subset batch of data samples for the oracle (i.e. the annotator). Pool or batch-based active learning methods are more scalable than querying single data sample per learning cycle [20]. Most of
the active learning work [2,42] formulate acquisition functions as information theoretic uncertainty estimates. While uncertainty-based methods work well for tasks like classification [17,49], where a single annotation is required per data sample, generalizations to document tasks such as page segmentation and named entity recognition, which require multiple annotations per selected data sample, have been scarcely explored. This is because methods to aggregate uncertainties over various entities present in a data sample are not well developed [7,41]. Recent techniques have been proposed to obtain a better acquisition function for active learning in these tasks [29,31]. However, these methods assume highly task-specific heuristics, and hence can not be generalized across different content detection scenarios.

In addition to active learning, in particular for dense annotation tasks in documents, weak learning can be an effective approach to reduce annotator’s efforts [36,37,50]. When there are multiple entities to be annotated in a data sample, weak learning reduces the annotation effort, either by providing faster variations of annotation techniques [37] or simply asking the annotator to verify the model predictions [36]. However, there are very few works [8,11] that combine weak learning with active learning. Furthermore, to the best of our knowledge, none of the works takes advantage of the annotator feedback (e.g. from annotator’s corrections of detected instance boundaries) during an active learning cycle.

In this work, we propose a policy-based active learning approach, taking into account the complexities of aggregating model uncertainties in the selection of samples to be labelled. We model the task of active learning as a Markov decision process (MDP) and learn an optimal acquisition function using deep Q-learning [34]. While several works rely on reinforcement learning for learning an optimal acquisition function [9,22,29,31], they assume task-specific representations of states and actions and hence are not generalizable across tasks. We further show that the proposed method can be combined with weak labelling, reducing the cost of annotation compared to strong labelling. Moreover, we incorporate class imbalance and human feedback signals into the design of MDP using suitable reward functions to further improve the performance of our approach.

To summarize, the major contributions of our work are as follows:

- We propose a policy-based task-agnostic active learning approach for complex content detection tasks, layout detection and named entity recognition in documents.
- We report that the proposed approach is generalizable, through demonstrating the performance of our active learning setup on varied detection tasks.
- We investigate the effectiveness of incorporating class balance and human feedback rewards in improving the active learning policy.
- We demonstrate the advantage of the proposed approach in reducing the costs of annotation in aforementioned complex detection tasks.

Throughout the remainder of the paper, we explain the proposed concepts, models, configurations, and discussions from the perspective of the layout and object detection, and named entity recognition tasks.

2. Related Work

Document content analysis has been studied extensively along several dimensions such as document classification (image [53,54] or text [1,38] or both [4,25]), named entity recognition in documents [32,55], content segmentation [19,35], document retrieval [10,45,48], layout analysis [5] among many others. The availability of large scale
labeled datasets of documents [21, 26, 27, 47, 58] led to the advent of several state-of-the-art deep learning models which have significantly improved these tasks in a large scale data setup. However, to the best of our knowledge, there is very limited amount of literature which uses active learning to optimize data annotation cost in a low resource setting, specifically for document analysis tasks [6, 18]. Therefore, in this section, we discuss about works that deal with general active learning policies, and active learning in a couple of related well studied domains, image classification, object detection and named entity recognition.

Active learning selects data samples with high uncertainty in the model prediction, which can provide more information to the underlying model. Different works have proposed different ways to compute model uncertainty [42]. While some methods depend on information theory for designing acquisition functions [17, 24, 49], others rely on alternative ways to approximate model uncertainty [13, 16]. Yoo et al [57] add a light-weight loss prediction module to the prediction model to predict the loss for the unlabelled samples, and use that as an uncertainty measure. Mayer et al [33] use uncertainty measure to find the optimal sample and query the data sample closest to the optimal sample.

For complex tasks such as object detection and named entity recognition, recent works [7, 41, 44] have been proposed to use uncertainty scores for the acquisition of samples. Most of these methods rely on aggregating the uncertainties of various entities within a data sample using max, sum or average functions [7, 41]. Aghdam et al [3] proposed a novel approach combining pixel-level scores to obtain an image-level score for doing active learning for the task of pedestrian detection task.

Several works have been proposed to incorporate reinforcement learning to learn an optimal acquisition function for active learning. The objective of these approaches is to model the active learning process into a Markov decision process through defining and designing suitable representations for states, actions, and rewards [15, 22, 30]. Liu et al [29] proposed an imitation learning approach for active learning in tasks related to natural language processing, relying on an algorithmic expert to find an optimal acquisition function. We differ from the work of Casanova et al [9] on using reinforced active learning approach for image segmentation, in terms of the generalize-ability of our approach on various tasks. We also report the effectiveness of using weak learning on top of policy-based active learning in consuming the budget with maximum efficiency.

3. Proposed OPAD Framework

In this section, we describe the proposed Optimized Policy-based Active Learning Framework for Document Content Analysis, OPAD. Figure 1 shows the interface for OPAD, which enables various scenarios of detection tasks for human annotators. The underlying algorithm for OPAD is a Deep Query Network (DQN)-based reinforcement learning policy, optimized for data sample selection based on the performance metrics for the task. OPAD has two stages - policy training stage and deployment stage. In the policy training stage, OPAD is trained using simulated active learning cycles to maximize performance on a validation set. While deploying, the trained policy is used to make online batch selection for annotation. The overall formulation for OPAD is described below.

3.1. Formulation

The underlying objective for policy training in OPAD is to perform an iterative selection of the samples from an unlabelled pool, $X_u$, which would maximally increase the performance of the model being trained, $\Theta$ until the annotation budget, $B$ is consumed. In each active learning cycle, the policy DQN $\Pi$ [34] selects a batch of $n_{cycle}$ samples, which are labelled, and added to the set of labelled samples $X_l$. The detection model $\Theta$ is then trained for a fixed number of epochs using the expanded set, $X_l$. The reward for the policy network for selecting the samples is the performance of the underlying model $\Theta$ computed using a metric apropos to the task (e.g. Average Precision for layout detection, and $F$-score for named entity recognition) on a separate held-out set, $X_{met}$. The training of the policy $\Pi$ is performed through episodes of active learning.

<table>
<thead>
<tr>
<th>Notations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{train}$, $X_{val}$, $X_{test}$</td>
<td>Train, Validation and Test sets of a given dataset</td>
</tr>
<tr>
<td>$X_u, X_l, X_{init}$</td>
<td>Unlabelled, labelled, and initial labelled sets</td>
</tr>
<tr>
<td>$X_{cand}$</td>
<td>Candidate unlabelled examples for an active learning cycle</td>
</tr>
<tr>
<td>$X_{met}, X_{state}$</td>
<td>Metric calculation set, State representation set</td>
</tr>
<tr>
<td>$A_t, S_t, R_t$</td>
<td>Action, State and Reward at time $t$</td>
</tr>
<tr>
<td>$\Pi, \Theta$</td>
<td>Policy deep Q network and Prediction model to be trained</td>
</tr>
<tr>
<td>$M, B$</td>
<td>Memory buffer for $Q$ learning, Total budget for active learning</td>
</tr>
<tr>
<td>$n_{cycle}$, $n_{pool}$, $n_{init}$</td>
<td>Number of samples to be acquired in one active learning cycle, Number of samples in a pool, Number of samples labelled for initial training</td>
</tr>
</tbody>
</table>

Table 1. Notations used to represent various data splits and model components.

We now describe various components of the proposed policy-based active learning approach in details.
3.2. Data Splits

Given a dataset \( D \), we split the samples (or use the existing splits of the dataset) into \( X_{train} \), \( X_{val} \), and \( X_{test} \) sets. For the two stages of \( OPAD \), the further splits are as follows.

**During policy training stage** We separate a set of samples \( X_{met} \) along with their labels from \( X_{train} \), which is used for validating the performance of underlying model \( \Theta \) and computing rewards for training the policy DQN \( \Pi \). For the RL setup of the policy DQN, we use a held-out set \( X_{state} \) which is used together with \( X_{cand} \) later to compute overall state representation. Note that, unlike \cite{9}, we do not require labels for \( X_{state} \), which further reduces the annotation budget. During this stage, we train the detection model \( \Theta \) on \( X_1 \), which is initialized with \( X_{init} \) and populated with samples from \( X_u \) as the active learning progresses. Here, \( X_{init} \) is a set with \( n_{init} \) randomly selected samples with the corresponding labels for initial training of the model \( \Theta \). Therefore, before the active learning process starts, \( X_u \) equals \( X_{train} - \{ X_{init} + X_{state} + X_{met} \} \), and \( X_1 \) equals \( X_{init} \).

**During deployment stage** We utilize the \( X_{val} \) set for training the detection model \( \Theta \). We make this differentiation from the policy training stage to ensure that sample selection by the policy happens on an unseen set. During this stage, we use the same terminology \( X_{init} \), \( X_1 \), and \( X_u \) from the previous stage. However, the \( n_{init} \) samples in \( X_{init} \) set are selected from the \( X_{val} \) set and therefore, at the start of the active learning process \( X_u \) equals \( X_{val} - \{ X_{init} \} \), and \( X_1 \) equals \( X_{init} \). We use the same set of samples for the state computation set \( X_{state} \). In this stage we do not require the \( X_{met} \) set.

Though we have ground truth annotations available for all the samples in all the three sets, to simulate the annotation setup, we mask this data from both \( \Theta \) and \( \Pi \) models and utilize the labels as and when required.

3.3. Active Learning

[hl] [1] **Input:** \( X_{train} \), budget \( B \) **Output:** Policy DQN, \( \Pi \), trained for querying the samples for annotation Randomly sample examples from \( X_{train} \) to form \( X_{state} \) and \( X_{met} \) sets. Initialize policy and target DQN Initialize memory replay buffer \( M \) convergence of DQN loss Initialize \( \Theta \) Randomly sample \( n_{init} \) from \( X_{train} - \{ X_{state} + X_{met} \} \) to form \( X_{init} \) Initialize \( X_u \) to \( X_{train} - \{ X_{state} + X_{met} + X_{init} \} \) Initialize \( X_1 \) to \( X_{init} \) Train the model \( \Theta \) on \( X_1 \) Compute the performance metric on \( X_{met} \) Consumption of budget \( B \) Sample \( n_{pool} \times n_{cycle} \) number of samples from \( X_u \) as candidates for labelling \( X_{cand} \) Compute state representation \( S_t \) using predictions of model \( \Theta \) on \( X_{state} \) and \( X_{cand} \) Select \( n_{cycle} \) samples from \( X_{cand} \) using \( c \)-greedy policy and add it to \( X_1 \) - Action \( A_t \) Retrain the model \( \Theta \) on \( X_1 \) Compute the metric on the \( X_{met} \) Compute the reward \( R_{t+1} \) as the difference in metric Re-do steps 14 and 15 - Next State \( S_{t+1} \) Add tuple \( (S_t, A_t, R_{t+1}, S_{t+1}) \) to the memory replay buffer \( M \) Optimize policy DQN, \( \Pi \) Figure 2 shows an overview of active learning (inner while loop at step 11 in Algorithm 3.3) in a single episode of policy training. In an active learning cycle, we select \( n_{pool} \times n_{cycle} \) number of samples from the set \( X_u \), which represent the candidates selected for the current active learning cycle \( X_{cand} \). The policy DQN \( \Pi \) computes \( Q \)-value for samples within each pool containing \( n_{pool} \) samples, based on candidate set \( X_{cand} \) and state representation set \( X_{state} \). The policy selection network is optimized to maximize the reward, \( R_t \):

\[
Q^*(S_t, A_t) = \max_\Pi \mathbb{E}[R_{t+1} | S_t, A_t, \Pi]
\]

(1)

The annotator then annotates the selected samples, and the labelled set \( X_1 \) is updated by adding these new samples.
We then retrain the model $\Theta$ using the updated labelled set and finally calculate the reward for the current cycle $R_t$ by measuring the performance of the model $\Theta$ on $X_{test}$.

$$R_{t+1} = \text{Performance}_{t,X_{test}} - \text{Performance}_{t-1,X_{test}}$$

(2)

where Performance is measured in terms of AP metric for layout and object detection tasks, and F-score for named entity recognition task. Algorithm 3.3 summarizes the training phase of the proposed approach.

3.4. Policy Training Stage

Policy Network Our policy network $\Pi$ is a deep query network, as shown in Figure 3. The underlying prediction model $\Theta$ computes the representations $c_t$ and $s_t$ from the sets $X_{ cand}$ and $X_{state}$ respectively (details in Section 4.3). The policy network then receives the two inputs $s_t$ and $c_t$, which we denote as the state representation $S_t$ in Figure 3. We pass the two representations through convolution layers, followed by vector product of state and candidate representations. The final Q-value is obtained by passing the combined representation through fully connected layers.

Policy Optimization The computed Q-value is used for selecting $n_{cycle}$ samples at each step. For this, a memory or experience replay buffer, $M$ is created using MDP state representation tuples, $(S_t, A_t, R_{t+1}, S_{t+1})$. Further, as a batch of $n_{cycle}$ needs to be selected, the candidate set, $X_{ cand}$, is randomly partitioned into $n_{cycle}$ mini-batches, and action set $A_t$ is set to $A_t^{n_{cycle}}$. The loss is then optimized as follows to train the policy network:

$$Loss(\Pi) = E_{t \in M}[(\gamma^i_t - Q(S_t, A_t^i); \Pi))^2]$$

(3)

The values for $\gamma^i_t$ are computed using a double DQN formulation [22] incorporating a target network, $\Pi'$ for stable training:

$$\gamma^i_t = R_{t+1} + \max_{A_{t+1}} \gamma Q(S_{t+1}, A_{t+1}^i; \Pi'); \Pi)$$

(4)

where, $\gamma$ is the discount factor for future reward, set to 0.9 in our experiments.

$\epsilon$-greedy selection To encourage exploration of diverse samples by the policy during training, an $\epsilon$-greedy strategy is followed while training the policy, which selects a random sample for the action $A_t^i$ with probability $\epsilon$-silon, instead of the sample maximizing Q-value. The $\epsilon$ value starts with 0.9 for the initial cycle, and decreases by a factor of 0.1 for subsequent cycles. For policy deployment, $\epsilon$ is set to 0. The gradient optimization is done using the temporal difference method [46].

3.5. Deployment Stage

[1] Input: $X_{val}$, $X_{test}$, $X_{state}$, budget $B$ Randomly sample $n_{init}$ from $X_{val}$ to form $X_{init}$ Initialize $X_u$ to $X_{val} - \{X_{init}\}$ Initialize $X_t$ to $X_{init}$ Initialize $\Theta$ Train the model $\Theta$ on $X_t$ Compute the performance metric on $X_{test}$ Consumption of budget $B$ Sample $n_{pool} \times n_{cycle}$ number of samples from $X_u$ as candidates for labelling $X_{ cand}$ Compute state representation $S_t$ using predictions of model $\Theta$ on $X_{state}$ and $X_{ cand}$ Select $n_{cycle}$ samples from $X_{ cand}$ using $\epsilon$-greedy policy and add it to $X_t$ - Action $A_t$ Retrain the model $\Theta$ on $X_t$ Compute the metric on $X_{test}$ and report the values in Section 4.

Algorithm 3.5 summarizes the deployment stage (or policy testing stage). We freeze the parameters of the model $\Pi$ in this stage. We use the $X_{val}$ set to iteratively select the samples and train the model $\Theta$. At the end of each active learning cycle we compute the performance of the model $\Theta$ on the held-out set $X_{test}$ and report the values in Section 4.

3.6. Weak labelling

In a usual annotation scenario (as shown in Figure 4 - top), the annotator has to mark all the entities present in a sample by drawing the bounding boxes and selecting labels for them. To reduce the annotation cost, we propose a weak labelling annotation framework (Figure 4 - bottom). Inspired from [36], the annotator is shown the document as well as the predictions with high confidence from the model $\Theta$ for that document. The annotator can then (1) add a missing box, (2) mark a box either correct or incorrect, and (3) mark a label either correct or incorrect for the associated box. The annotation interface for the weak labelling approach is shown in Figure 1.

Figure 4. Weak labelling in the case of layout detection. In the top image, the annotator has to draw and mark all the layout boxes, while in the bottom image, the annotator can verify the predictions of the model in the input image, and add new boxes. Image is best viewed in color.

The advantage of weak labelling is that it significantly reduces the annotation time. Annotation of a new entity by drawing a bounding box or selecting words takes $\sim 15$ seconds on an average in the case of detection tasks and $\sim 4$ seconds in case of named entity recognition. Verifying an
entity takes \( \sim \) 5 seconds for layout detection task and \( \sim \) 2 seconds for named entity recognition\( ^1 \).

### 3.7. Additional Rewards

We propose the following additional rewards to improve the performance of the active learning approach.

- **Class balance reward**: To reduce class imbalance in the newly acquired samples that are to be labelled, \( X_{\text{new}} \), we propose an additional class distribution entropy reward which reinforces a class-balanced selection of samples.

\[
R_{\text{cls,ent}} = \mathcal{H}(P(X_{\text{new}}))
\]

where \( \mathcal{H} \) is the Shannon entropy function \( ^{\text{[43]}} \), and \( P(X_{\text{new}}) \) is the probability distribution over various classes for the newly acquired samples \( X_{\text{new}} \).

- **Human feedback reward**: In a weak labelling scenario, where the annotator can modify the output from the prediction model, \( \Theta \), a human feedback signal could be added at each active learning cycle while training the policy. The objective is to promote the selection of those samples for which the annotator modifies the high confidence predictions of \( \Theta \) heavily because such samples would be more informative for the model \( \Theta \). Accordingly, the additional human feedback reward for detection during training time is given as,

\[
R_{\text{feedback}} = AP_{\text{after,feedback}} - AP_{\text{before,feedback}}
\]

where \( AP_{\text{after,feedback}} \) is the AP metric on the newly acquired samples, after the annotator has verified the predictions, and \( AP_{\text{before,feedback}} \) is the AP of the samples before feedback.

### 4. Experiments and Results

In this section, we provide a comprehensive experimental evaluation of the proposed policy-based active learning approach on the document understanding tasks, document layout detection and named entity recognition. Furthermore, we also evaluate our models on Pascal VOC object detection task to demonstrate the generalizability of the proposed solution across different domains.

#### 4.1. Datasets

We use the following datasets for the corresponding tasks:

- **GROTOAP2** \( ^{\text{[47]}} \) dataset is used for the complex document layout detection task. The dataset consists of 22 layout classes for scientific journals. We sampled two sets of 5000 images as training and validation sets. Among these, we hold-out 10% for reward computation set \( X_{\text{met}} \) and 256 random samples for \( X_{\text{state}} \) and use the remaining samples for the active learning setup. We use the validation set for simulating the active learning during the deployment phase and finally report the performance on a held-out subset of 2500 images. Further, we merged those classes having very few instances (e.g. *glossary, equation, etc.*) with the *body content* class, resulting into a modified dataset with 13 classes.

- **Pascal VOC-2007** \( ^{\text{[14]}} \) dataset with 20 object classes is used for the object detection task. We use the *train* set of VOC-2007 containing 2501 images during the policy training phase. Similar to layout detection task, we hold-out 10% for reward computation set \( X_{\text{met}} \) and 256 random samples for \( X_{\text{state}} \) and use remaining samples for the active learning setup. During the deployment phase, we utilize the *val* set of VOC-2007 containing 2510 images for simulating the active learning setup i.e. selecting samples using trained \( \Pi \) model and training the model \( \Theta \). We use the *test* set of VOC-2007 consisting of — samples for reporting the performance of model \( \Theta \) after each active learning cycle during the deployment stage.

We also use the following datasets for pre-training the underlying model \( \Theta \):

- **PubLayNet** \( ^{\text{[58]}} \) We use this dataset for pre-training \( \Theta \) for document layout detection. This dataset contains over 360K page samples and has typical document layout elements such as *text, list, figure, and table* as the annotations. While the *list, figure, table* and *title* classes contains the corresponding information from document, the *text* category consists of the rest of the content such as author, author affiliation; paper information; copyright information; abstract; paragraph in main text, footnote, and appendix; figure & table caption; table footnote.

- **MS-COCO** \( ^{\text{[28]}} \) This dataset consists of 91 object classes. We use this dataset to pre-train the underlying classification model \( \Theta \) (i.e. Faster-RCNN model) in the case of object detection on the VOC dataset. We pre-train the model \( \Theta \) on this dataset and remove the last layers from both the class prediction and bounding box regression branches which are class-specific.

#### 4.2. Models and configurations

We use the Faster-RCNN model \( ^{\text{[40]}} \) with RESNET-101 backbone\( ^{\text{[23]}} \) as the underlying prediction model for the
layout detection and object detection tasks. The Faster-RCNN model is pre-trained on a subset of 15000 images from PubLayNet [58] dataset for the layout detection task, and on MS-COCO [28] dataset for the object detection task to bootstrap the active learning experiments.

For active learning, we use a seed set of 512 labelled samples in case of detection tasks initially. The Faster-RCNN model is trained for 1000 iterations on the labelled set in an active learning cycle. In each of the 10 active learning cycles we select 64 samples for the detection tasks, from unlabelled dataset for labelling giving a total of 1152 and 350 labelled samples in a single episode for detection tasks and NER respectively. We run 10 episodes of these active learning cycles to train the policy network. The learning rate for training the policy DQN is set to 0.001 with a gamma value of 0.998. The learning rates of Faster-RCNN is set to 0.00025. We also apply a momentum of 0.95 to optimize the training of policy network. We set the size of memory replay buffer M to 1000 samples with first-in-first-out mechanism.

### 4.3. MDP state representation

For the layout detection and object detection tasks, we use a randomly sampled set of 256 images from the train set as the subset for representing the overall distribution of the dataset ($X_{state}$). We pass each instance from the candidate ($X_{cand}$) and state ($X_{state}$) subsets through the Faster-RCNN model, to get the top 50 confident bounding box predictions. We concatenate the class scores for these top 50 predictions to the feature map of RESNET-101 backbone to get a final representation (1256-dimension for VOC-2007, and 906-dimension for GROTOAP2) for each sample in the candidate and state subset sets. The representations thus obtained from the samples in $X_{cand}$ are stacked to form $c_t$, and similarly $s_t$ from the set $X_{state}$. Together $c_t$ and $s_t$ form the state representation $S_t$ in Figure 3.

### 4.4. Human Annotation Simulation

To simulate the role of a human annotator for weak labelling, we use the ground truths of the datasets on which we perform our experiments. In detection tasks (i.e. layout detection and object detection), we consider the predictions which have an IoU greater than 0.5 with the ground truth box as the boxes being marked as correct by the annotator. For those boxes in the ground truth which do not have any prediction with IoU greater than 0.5, we include that box into the labelled set marking as a full annotation (a strong label).

### 4.5. Results

We compare the performance of our proposed method with three baselines -

<table>
<thead>
<tr>
<th>Method</th>
<th>Strong Annotation time required(seconds)</th>
<th>Weak Annotation time required(seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROTOAP2</td>
<td>72500</td>
<td>38000</td>
</tr>
<tr>
<td>VOC2007</td>
<td>9000</td>
<td>4250</td>
</tr>
</tbody>
</table>

Table 2. Time required for one active learning cycle i.e selection of samples for various algorithms along with the model training time. Note that the model training time is constant.

<table>
<thead>
<tr>
<th>Method</th>
<th>Strong Annotation time required(seconds)</th>
<th>Weak Annotation time required(seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROTOAP2</td>
<td>66000</td>
<td>33000</td>
</tr>
<tr>
<td>VOC2007</td>
<td>7000</td>
<td>2250</td>
</tr>
</tbody>
</table>

Table 3. Annotation time required to reach an AP of 42.5 on GROTOAP2 and an AP of 45.5 on VOC-2007. These values indicate the minimum achievable best performances by all the models on the datasets.

- **Random** Data samples from the unlabelled pool are randomly chosen for annotation.

- **Entropy** [41] For the entropy-based selection, first the entropy of class prediction probability by $\Theta$ is computed over all the entities of a data sample. We present results for aggregating entropy of a single sample in two ways: 1. maximum entropy, 2. sum of entropy of all detected entities within the sample, and then the samples with the highest aggregate entropy are selected for labelling.

- **Margin** [7] Similar to entropy, a $v_{1\times 2}$ margin score is computed using the difference of prediction probability of highest and second highest class for all the instances of a sample. Then, the maximum margin score over all the instances is taken to be the aggregate margin measure for the sample. Samples with the highest aggregate margin are selected for labelling. The baseline metrics are as described in the existing prior art.

Figure 5 shows the accuracy of all the methods on the test sets of different datasets, for both strong and weak labelling settings. We can observe that the proposed policy-based
AL method significantly outperforms the baseline methods. This is because of the optimized selection policy, learned to reward the better performance of the prediction model. While the curves for VOC-2007 approach saturation, we stop the GROTOAP2 training before reaching saturation as our objective is to show the performance of the underlying model with a limited budget. Note that the proposed method uses vanilla reward in all the plots in Figure 5. Further, as shown in Table 2 and Table 3, the proposed method takes significantly less time for annotation than the baselines to reach the minimum best performance achievable by all the models, while performing only next to random algorithm for sample selection timings. The annotation times in Table 3 are based on the number of samples selected for annotation multiplied by the average human annotation times mentioned in Section 3.6.

5. Ablation Study

In this section we discuss the importance of the proposed additional rewards in improving the performance of the proposed AL approach.

5.1. Class balance reward

We conduct ablations by adding the class distribution entropy reward (Equation 5) to the vanilla reward function. The overall reward function is:

\[ \mathcal{R}_{overall} = \mathcal{R}_t + \lambda \cdot \mathcal{R}_{cls,ent} \]

where \( \lambda \) is a hyper-parameter, and \( \mathcal{R}_t \) is the vanilla reward. As seen in Table 4, we observe a significant increase in performance as compared to the vanilla reward policy.

5.2. Human feedback reward

In this experiment we report the effect of adding human feedback to the vanilla reward, i.e.

\[ \mathcal{R}_{overall} = \mathcal{R}_t + \lambda \cdot \mathcal{R}_{feedback} \]

where \( \lambda \) is a hyper-parameter. We report the results of using this overall reward in our policy in Table 5, along with the baselines and vanilla policy in a weak labelling setup. We observe that having a small weight on the feedback reward results in a jump in the performance.

---

Table 5. Performance of our method with human feedback reward for weak labelling on GROTOAP2 and VOC-2007.

<table>
<thead>
<tr>
<th>Method</th>
<th>GROTOAP2</th>
<th>VOC-2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>50.668</td>
<td>47.490</td>
</tr>
<tr>
<td>Entropy Max</td>
<td>46.229</td>
<td>46.671</td>
</tr>
<tr>
<td>Entropy Sum</td>
<td>46.634</td>
<td>47.431</td>
</tr>
<tr>
<td>Margin</td>
<td>47.428</td>
<td>47.179</td>
</tr>
<tr>
<td>OPAD</td>
<td>51.508</td>
<td>48.061</td>
</tr>
<tr>
<td>OPAD (ClsEnt ( \lambda = 0.25 ))</td>
<td>53.241</td>
<td>47.727</td>
</tr>
<tr>
<td>OPAD (ClsEnt ( \lambda = 0.50 ))</td>
<td>51.185</td>
<td>47.701</td>
</tr>
<tr>
<td>OPAD (ClsEnt ( \lambda = 0.75 ))</td>
<td>52.143</td>
<td><strong>48.566</strong></td>
</tr>
<tr>
<td>OPAD (ClsEnt ( \lambda = 1.0 ))</td>
<td>51.530</td>
<td>48.060</td>
</tr>
</tbody>
</table>

6. Conclusion and Future Works

We present a robust policy-based method for active learning task in complex content detection problems. The problem of active learning in detection is formulated using a DQN-based sampling network, optimized for task performance metrics. We extend the active learning setting to weak labelling, and propose rewards for class balance and human feedback. To the best of our knowledge, this is first-of-its-kind work optimizing active learning for detection tasks in documents. We show the efficacy of the proposed methods on a large document detection set as well as object detection. As a future direction, we would like to improve on the DQN, and further explore more recent active learning acquisition functions.
References

[2] Charu C. Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and Philip S. Yu. Active learning: A survey, pages 571–605. CRC Press, Jan. 2014.
[26] David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. Building a test col-


