Contrastive Out-of-Distribution with Multi-modal Information for Document Classification

Shuxian (Trinity) Fan
Department of Statistics
University of Washington
fansx@uw.edu

Hao Li
Department of Applied Mathematics
University of Washington
haoli25@uw.edu

Warren Paris-Moe
Department of Applied Mathematics
University of Washington
parismoe@uw.edu

Yilun (Ellen) Xing
Industrial & Systems Engineering Department
University of Washington
yilunx19@uw.edu

Abstract

Pretrained transformers have demonstrated remarkable performance in document classification when dealing with training and test data from the same distribution. However, in real-world applications, the model often faces out-of-distribution (OOD) instances that can lead to significant performance degradation. In this project, we tackle this issue in the context of document classification, where a reliable model should be capable of identifying OOD instances while maintaining good performance on the primary classification task. Additionally, we conduct extensive experiments to validate the hypothesis that incorporating multi-modal information improves OOD generalization for document classification. We investigate OOD detection performance using (i) text-layout information based on the LayoutLM transformer and (ii) image-layout-text information based on LayoutLM and a ResNet-based feature extractor. Our results demonstrate that, as a proof of concept, incorporating image information enhances OOD detection performance.

1 Introduction

Most natural language classifiers rely on the closed-world assumption, which means that they assume that both the training and testing data come from the same distribution. This assumption is particularly crucial in NLP tasks such as Named Entity Recognition (NER), where the model must identify entities from a predetermined set of categories. One real-world example is automating claims processing, which involves using machine learning to extract relevant client information from scanned health insurance claim forms. However, these forms may contain irrelevant pages, such as legal pages or additional descriptions, which can lead to inaccurate information being extracted. Therefore, accurately classifying the scanned images and filtering out the irrelevant pages is critical. To accomplish this, an OOD detection task is essential.

Despite the importance of OOD detection in document understanding tasks, only a few attempts have been made in this area, as documented in studies such as [6, 12, 5, 20, 15]. Furthermore, there is an increasing interest in NLP models that can effectively handle multi-modal data, such as LayoutLM [19], which utilizes both text and layout information for document image understanding and information extraction, and LayoutLMv2 [18], which models the interaction between text, layout, and image in a single multi-modal framework.

In this project, we aim to validate the hypothesis that, incorporating multi-modal information can improve the OOD generalization for document classification based on pre-trained Transformers. We

Preprint. Under review.

focus on the OOD scenario with semantic shift, where OOD data are those with labels that do not belong to the training label set. Mathematically, let $\mathcal{Y}$ and $\mathcal{X}$ denote the training label set and the training feature space and define the training distribution $\mathbb{P}(\mathcal{X}, \mathcal{Y})$. In the context of this project, we consider an instance $(\boldsymbol{x}, y)$ to be OOD if $y \notin \mathcal{Y}$ and in-distribution (ID) if $y \in \mathcal{Y}$ and $\boldsymbol{x} \in \mathcal{X}$.

Our goal, given a main task of document image classification built on pre-trained transformers, is to develop a separate OOD detector that can distinguish between ID instances and OOD instances based on a predefined metric while maintaining the performance of the main classification task on the ID data. We adopt the setting where only ID data are available during the training stage for practical purposes. To this end, we consider an unsupervised OOD detection task with contrastive representation learning, which consists of a contrastive loss and an OOD scoring function. The contrastive loss aims to increase the inter-class discrepancies, and the OOD scoring function maps the representations to OOD scores, which indicates the likelihood of an instance being OOD. Studies have shown that contrastive representation learning can help the model learn discriminative features for ID/OOD distinctions [20]. To incorporate multi-modal information, we use a ResNet-based model to extract image features, along with LayoutLM to extract text and layout information. We examine different combinations of contrastive losses and OOD scores with and without image information.

The report is structured as follows. Section 2 provides a brief review of related work in the literature. Section 3 describes the data used in this project. Section 4 details the proposed method and experimental settings. The method is then applied in Section 5 on the motivating datasets, and the results are summarized as proof of concept of the proposed hypothesis. Section 6 provides our concluding remarks.

2 Related Work

Transformers In this project,we implement our method using Huggingface's Transformers [17]. Specifically, we adopt LayoutLM, which was first introduced by Yiheng Xu et al. (2020) [19] to jointly model the interactions between text and layout information of scanned documents instead of only considering text-level manipulation [2]. This model can automatically classify, extract, and structure information from business documents. While the state-of-the-art NLP models consider text-layout information, we aimed to take additional image information into account. It is noteworthy that the authors have introduced an improved version of LayoutLM, LayoutLM2 [18], with similar ideas to learn the interaction between visual and textual information. In LayoutLMv2, the authors have introduced a spatial-aware self-attention mechanism, which incorporates a 2-D relative position representation for token pairs and offers a more comprehensive view for contextual spatial modeling compared to the absolute 2-D position embeddings used in LayoutLM. Nevertheless, we found that the model structure with token-wise image information presented in LayoutLMv2 was too complicated for our document classification task. Instead, we have implemented a simpler model structure with multi-modal information, similar to that in [7]. Specifically, we only use the document-wise ResNet-based image features along with the LayoutLM representations in the fine-tuning stage. This approach has been found to be effective in our experiments.

Out-of-Distribution Detection In the realm of natural language classifiers, the closed-world assumption is typically employed, where the training and test data are sampled from the same distribution. However, in real-world scenarios, document images may come from unknown categories, causing the model to mistakenly classify them as one of the existing categories. Therefore, it is crucial to introduce out-of-distribution (OOD) detection to pre-trained models to detect such exceptions. For example, an unsupervised OOD detection framework was introduced by Zhou et al. [20], which proposes a contrastive loss to fine-tune transformers and improve the compactness of representations, thereby better differentiating OOD instances from in-distribution (ID) instances. Two types of contrastive losses were considered: supervised contrastive loss and margin-based contrastive loss. However, the authors only considered text data as input, ignoring the layout information contained in image data, which plays a critical role in differentiating document types. To address this issue, we propose to incorporate multi-modal information in the OOD detection task, motivated by the methods proposed in [19, 18]. Additionally, recent research has improved contrastive learning techniques [15].

3 Datasets

We introduce the following two datasets used to validate the hypothesis of OOD detection performance with multi-modal information.

RVL-CDIP [3](Ryerson Vision Lab Complex Document Information Processing) consists of $400,000$ grayscale document images across 16 categories, with 25000 images per class. Following the original split, 320000 images are used for training, 40000 images for validation, and 40000 images for testing. This dataset is publicly available[1]. Figure 1 gives selected scanned document image examples.

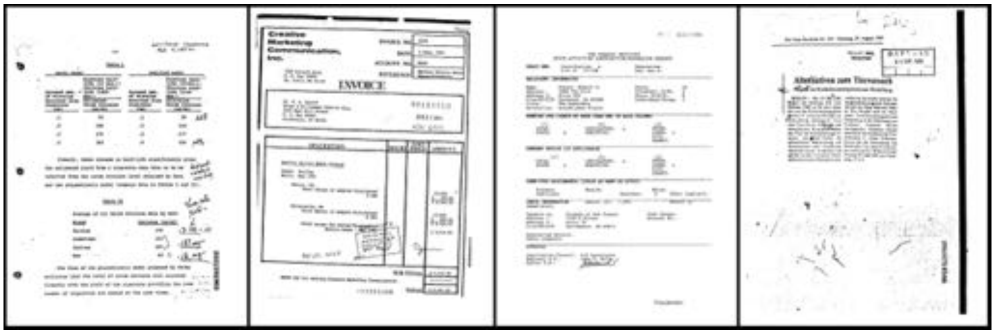
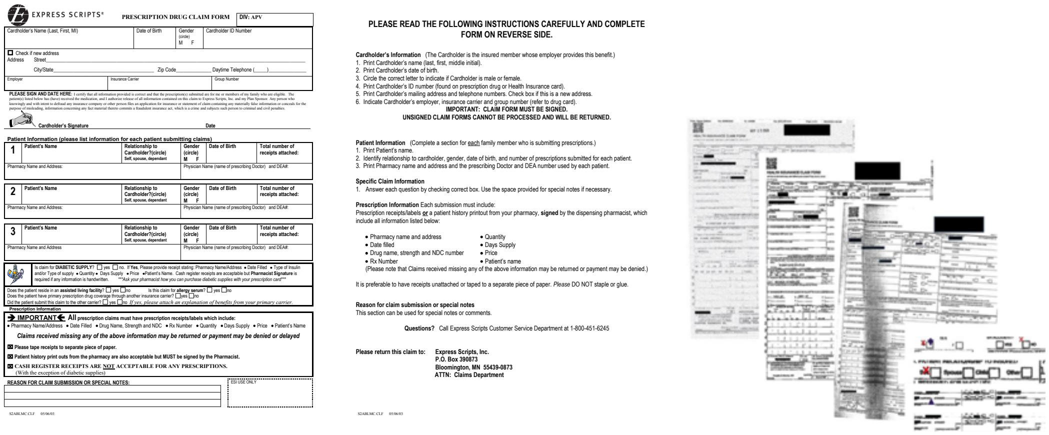


Figure 1: Scanned documentation example: `scientific report, invoice, budget, news article` from left to right.

Health Insurance Claim Forms consist of $\approx$ 3000 scanned medical claim documents belonging to four categories, collected within American Family Insurance company for developing automated claim processing methods. Unfortunately, due to company privacy policies, this dataset is not publicly available. However, we provide similar examples in Figure 2, taken from released public information. The ID instances are those with the standard claim tables required for filing insurance claims. The OOD instances are irrelevant pages that are not used for making claims. From processing the available claim forms, we obtain $\approx$ 3000 ID standard medical forms. They are split into $\approx$ 1800 samples for training, $\approx$ 600 samples for validating and $\approx$ 600 samples for testing.



EXPRESS SCRIPTS® PRESCRIPTION DRUG CLAIM FORM DIV: APV

Cardholder's Name (Last, First, MI)	Date of Birth	Gender (circle) M F	Cardholder ID Number

□ Check if new address
Address Street__________
City/State__________ Zip Code______ Daytime Telephone (____)______

Employer	Insurance Carrier	Group Number

PLEASE SIGN AND DATE HERE: I certify that all information provided is correct and that the prescription(s) submitted are for me or members of my family who are eligible. The patient(s) listed below has (have) received the medication, and I authorize release of all information contained on this claim to Express Scripts, Inc. and my Plan Sponsor. Any person who knowingly and with intent to defraud any insurance company or other person files an application for insurance or statement of claim containing any materially false information or conceals for the purpose of misleading, information concerning any fact material thereto commits a fraudulent insurance act, which is a crime and subjects such person to criminal and civil penalties.

Cardholder's Signature **Date**

Patient Information (please list information for each patient submitting claims)

1	**Patient's Name**	**Relationship to Cardholder?(circle)** Self, spouse, dependant	**Gender (circle)** M F	**Date of Birth**	**Total number of receipts attached:**
Pharmacy Name and Address:			Physician Name (name of prescribing Doctor) and DEA#:		
2	**Patient's Name**	**Relationship to Cardholder?(circle)** Self, spouse, dependant	**Gender (circle)** M F	**Date of Birth**	**Total number of receipts attached:**
Pharmacy Name and Address:			Physician Name (name of prescribing Doctor) and DEA#:		
3	**Patient's Name**	**Relationship to Cardholder?(circle)** Self, spouse, dependant	**Gender (circle)** M F	**Date of Birth**	**Total number of receipts attached:**
Pharmacy Name and Address			Physician Name (name of prescribing Doctor) and DEA#:		

Is claim for **DIABETIC SUPPLY**? □ yes □ no. If **Yes**, Please provide receipt stating: Pharmacy Name/Address • Date Filled • Type of Insulin and/or Type of supply • Quantity • Days Supply • Price • Patient's Name. Cash register receipts are acceptable but **Pharmacist Signature** is required if any information is handwritten. ****Ask your pharmacist how you can purchase diabetic supplies with your prescription card****

Does the patient reside in an **assisted living facility**? □ yes □no Is this claim for **allergy serum**? □ yes □no
Does the patient have primary prescription drug coverage through another insurance carrier? □yes □no
Did the patient submit this claim to the other carrier? □ yes □no *If yes, please attach an explanation of benefits from your primary carrier.*

Prescription Information

➔ IMPORTANT ➔ All prescription claims must have prescription receipts/labels which include:
• Pharmacy Name/Address • Date Filled • Drug Name, Strength and NDC • Rx Number • Quantity • Days Supply • Price • Patient's Name

Claims received missing any of the above information may be returned or payment may be denied or delayed

☒ **Please tape receipts to separate piece of paper.**
☒ **Patient history print outs from the pharmacy are also acceptable but MUST be signed by the Pharmacist.**
☒ **CASH REGISTER RECEIPTS ARE NOT ACCEPTABLE FOR ANY PRESCRIPTIONS.**
(With the exception of diabetic supplies)

REASON FOR CLAIM SUBMISSION OR SPECIAL NOTES: ESI USE ONLY

S2ABLMC.CLF 05/06/03

PLEASE READ THE FOLLOWING INSTRUCTIONS CAREFULLY AND COMPLETE FORM ON REVERSE SIDE.

Cardholder's Information (The Cardholder is the insured member whose employer provides this benefit.)
1. Print Cardholder's name (last, first, middle initial).
2. Print Cardholder's date of birth.
3. Circle the correct letter to indicate if Cardholder is male or female.
4. Print Cardholder's ID number (found on prescription drug or Health Insurance card).
5. Print Cardholder's mailing address and telephone numbers. Check box if this is a new address.
6. Indicate Cardholder's employer, insurance carrier and group number (refer to drug card).

IMPORTANT: CLAIM FORM MUST BE SIGNED.
UNSIGNED CLAIM FORMS CANNOT BE PROCESSED AND WILL BE RETURNED.

Patient Information (Complete a section for each family member who is submitting prescriptions.)
1. Print Patient's name.
2. Identify relationship to cardholder, gender, date of birth, and number of prescriptions submitted for each patient.
3. Print Pharmacy name and address and the prescribing Doctor and DEA number used by each patient.

Specific Claim Information
1. Answer each question by checking correct box. Use the space provided for special notes if necessary.

Prescription Information Each submission must include:
Prescription receipts/labels **or** a patient history printout from your pharmacy, **signed** by the dispensing pharmacist, which include all information listed below:

- Pharmacy name and address
- Date filled
- Drug name, strength and NDC number
- Rx Number
- Quantity
- Days Supply
- Price
- Patient's name

(Please note that Claims received missing any of the above information may be returned or payment may be denied.)

It is preferable to have receipts unattached or taped to a separate piece of paper. *Please* DO NOT staple or glue.

Reason for claim submission or special notes
This section can be used for special notes or comments.

Questions? Call Express Scripts Customer Service Department at 1-800-451-6245

Please return this claim to: **Express Scripts, Inc.**
P.O. Box 390873
Bloomington, MN 55439-0873
ATTN: Claims Department

S2ABLMC.CLF 05/06/03

(a) ID instance: claim pages (b) OOD instance: irrelevant pages (c) Scanned form example: `CMS1500`.

Figure 2: Health insurance claim forms example: ID and OOD examples from prescription drug claim form [1] and scanned claim forms [14].

[1] `https://huggingface.co/datasets/rvl_cdip`

4 Method

In this section, we first introduce the overall model framework and the algorithm of the learning process (Sec. 4.1), then we detail the contrastive learning technique (Sec. 4.2) and the adopted OOD scoring functions(Sec. 4.3). Finally, we discuss the computation issues for large-scale machine learning models and present our implementations (Sec. 4.4).

4.1 Model Framework

We illustrate the model framework in Figure 3 using a flow chart:

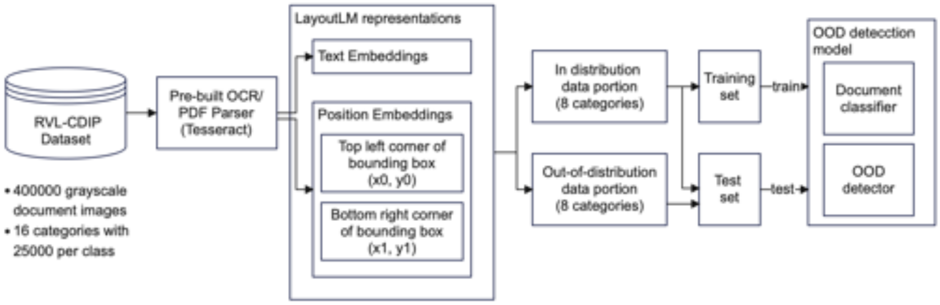

RVL-CDIP Dataset
• 400000 grayscale document images
• 16 categories with 25000 per class
Pre-built OCR/ PDF Parser (Tesseract)
LayoutLM representations
Text Embeddings
Position Embeddings
Top left corner of bounding box (x0, y0)
Bottom right corner of bounding box (x1, y1)
In distribution data portion (8 categories)
Out-of-distribution data portion (8 categories)
Training set
Test set
train
test
OOD detecction model
Document classifier
OOD detector

Figure 3: Flow chart illustrating the model structure and data flow for the contrastive out-of-distribution (OOD) detection method.

To summarize, the model performs two tasks simultaneously, namely image classification and OOD detection. The training process of the document image classifier involves the following steps: First, the raw image data is processed using optical character recognition (OCR) techniques to extract the text and position embeddings, which are then used as inputs for the pre-trained LayoutLM. Second, the model is fine-tuned to optimize both the main task image classification loss and the contrastive learning loss. We also incorporate an unsupervised OOD detection framework, which involves contrastive representation learning to distinguish between ID and OOD instances. The framework includes defining metric functions that map the representations into a score. The ideal model should have a low contrastive loss while maintaining classification performance.

Algorithm 1 Learning Process

```
Data: ID training set $\mathcal{D}_{\text{train}}$ and ID validation set $\mathcal{D}_{\text{val}}$.
Result: A trained classifier and an OOD detector.
Initialize the LayoutLM embeddings
for $t = 1 \ldots T$ do
    Sample mini-batches from $\mathcal{D}_{\text{train}}$.
    Calculate the classification loss $\mathcal{L}_{\text{ce}}$.
    Calculate the contrastive loss $\mathcal{L}_{\text{cont}}$.
    Calculate total loss $\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{cont}}$.
    Update model parameters w.r.t. $\mathcal{L}$.
    if $t$ % $evl.\ steps = 0$ then
        Fit the OOD detector on $\mathcal{D}_{\text{val}}$
        Evaluate scores both the classifier and OOD detector on $\mathcal{D}_{\text{val}}$.
    end
end
Return the best model checkpoint.
```

Algorithm 1 provides a detailed overview of our framework's learning process. In the training phase, both the training and validation datasets are used, and batches from the training set are employed to train the model using both classification and contrastive losses. The optimal model is selected based on its performance on the validation data. After training, the OOD detector is evaluated with the validation data used as an ID test set and the OOD data portion used as an OOD test set, where we compute the OOD scores of both ID and OOD instances. Ultimately, our framework generates a classifier for internally consistent data and an OOD detector for detecting OOD instances during inference.

4.2 Contrastive Learning

In practical, the distribution of OOD data is often unknown and unavailable during the training process. To address this challenge, we can leverage data from different classes to encourage the model to learn more discriminative representations. This can be achieved by defining a contrastive learning loss, which optimizes the compactness of same-class instances (ID) and encourages vectors of different-class instances to be far from each other. The contrastive loss aims to optimize the closeness of embeddings from positive pairs and the uniformity of the induced distribution of the normalized features on the hyperspace. These properties have been shown to have positive effects on downstream tasks such as OOD detection [16]. Meanwhile, the contrastive loss should have no negative impact on our in-domain classification task.

We consider the following two alternatives of the contrastive losses:

- **Supervised Contrastive Loss** [9]

  For a multi-class classification problem with C distinct classes, given a batch of training instances $\{\boldsymbol{x}_i, y_i\}_{i=1}^{M}$, the supervised contrastive loss is defined as:

$$L_{scl} = \sum_{i=1}^{M} \frac{1}{M|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\boldsymbol{z}_i^T \boldsymbol{z}_p}/\tau}{\sum_{a \in A(i)} e^{\boldsymbol{z}_i^T \boldsymbol{z}_a}/\tau}$$

  Here, the index i is called an anchor, and $A(i) = \{1, 2, \ldots, M\} \backslash \{i\}$ is the set of anchors and $P(i) = \{p \in A(i) : y_i = y_p\}$ is the set of indices of all positives (from the same class as i) in the batch distinct from i, and $|P(i)|$ is its cardinality. The hyper-parameter include the temperature τ. $\boldsymbol{z}$ is the L_2-normalized `[CLS]` LayoutLM embedding [19] used for sequence classification. The L_2-normalization is for avoiding extreme values in the dot product that may lead to unstable updates [20].

- **Margin-based Contrastive Loss** [20]

$$L_{margin} = \frac{1}{dM} \left(\sum_{i=1}^{M} \frac{1}{|P(i)|} \sum_{p \in P(i)} \|\boldsymbol{h}_i - \boldsymbol{h}_p\|_2^2 + \frac{1}{N(i)} \sum_{n \in N(i)} (\xi - \|\boldsymbol{h}_i - \boldsymbol{h}_n\|_2^2)_+ \right)$$

  Here, $N(i) = \{n \in A(i) y_i \neq y_n\}$ is the set of indices of all negatives (different class as i) in the batch distinct from i, and $|N(i)|$ is its cardinality. $\boldsymbol{h}$ is the un-normalized `[CLS]` embedding, d is the dimension of $\boldsymbol{h}$. The margin ϵ is empirically chosen as the maximum distance between pairs of instances from the same class in the batch [20]:

$$\epsilon = \max_{i=1}^{M} \max_{p \in P(i)} \|\boldsymbol{h}_i - \boldsymbol{h}_n\|_2^2$$

4.3 OOD Scores

In order to construct the OOD detector, we utilize various OOD scoring functions that were employed in our experiments. The main objective of these functions is to map the contrastive representations of ID and OOD instances to an OOD score, which is a measure of the likelihood that the instance is an OOD example. Higher values of the OOD score indicate a higher probability of the instance being OOD. We explore several commonly used scoring functions that have been utilized in the OOD literature.

- **Maximum Softmax Probability (MSP)**

$$g = 1 - \max_{j=1}^{C} \boldsymbol{p}_j$$

which uses the maximum class probability among C training classes in the softmax layer to indicate the presence of OOD.

- **Energy Score**

$$g = -\log \sum_{j=1}^{C} \exp(\boldsymbol{w}_j^T \boldsymbol{h})$$

where $\boldsymbol{w}_j \in \mathbb{R}^d$ is the weight of class j in the softmax layer and $\boldsymbol{h}$ is the input to the softmax layer.

- **Mahalanobis Distance (Maha)**

$$g = -\min_{j=1}^{C} (\boldsymbol{h} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{+} (\boldsymbol{h} - \boldsymbol{\mu}_j)$$

where a Gaussian distribution is fitted to the ID validation set using the input representaion $\boldsymbol{h}$ in the penultimate layer of the model with fitted mean vector $\boldsymbol{\mu}_j$ for class j and covariance matrix $\boldsymbol{\Sigma}$. Note that $\boldsymbol{\Sigma}^{+}$ is the pseudo-inverse of $\boldsymbol{\Sigma}$.

- **Cosine Similarity**

$$g = -\max_{i=1}^{M} cos(\boldsymbol{h}, \boldsymbol{h}_i^{(val)})$$

where cos is the angular similarity function of input representations.

4.4 PyTorch Lightning

Distributed, parallelized training is essential for large-scale machine learning models such as the one discussed in this paper. Distributed training enables the workload to be divided among multiple devices, increasing memory overhead and reducing the likelihood of maxing out CUDA memory. This is particularly important for models like ours that are heavy on GPU memory usage. When running the basic model on a single Nvidia RTX 6000, the GPU's 24GBs of CUDA memory are maxed out after just one epoch. Our model is large enough that when running on a Nvidia A40 with 48GB of memory, an out-of-memory (OOM) error is reached after only three epochs. Our model includes a training loop, evaluation loop over the validation and test datasets, and predict loop for each epoch were the test dataset is evaluated against the out-of-distribution dataset. With this setup a simple implementation of running each of the four loops on a separate GPU allowed for OOM errors to be overcome. However, this simple multi-GPU strategy is not resource efficient and leads to idle devices for much of the main training loop. Distributed training enables us to overcome OOM issues while effectively allocating resources and speeding up training significantly.

Several implementations of distributed training exist that can be implemented to achieve this. Distributed Data Parallel (DDP) parallelizes a model by splitting the input across specified devices. The model is replicated on each machine and device, each handling a portion of the input. Then, gradients from each device/node are averaged [**?**]. This is possible since the majority of transformations in a neural network do not involve data from separate samples. Therefore, the sum of per-parameter gradients calculated using mini-batch (x) subsets $(x^0, ..., x^n)$ will match the per-parameter gradients over the whole input batch ($\frac{\partial \mathcal{L}(x;w)}{\partial w} = \frac{\partial \mathcal{L}(x^0;w)}{\partial w} + ... + \frac{\partial \mathcal{L}(x^n;w)}{\partial w}$)[11]. Fully Sharded Data Parallel (FSDP) is different from DDP in that is shards model parameters, gradients, and optimizer states across GPUs and additionally offers the option to offload sharded model parameters to the CPU [**?**]. These are a few of the distributed training options that we were able to test briefly.

By using PyTorch Lightning Fabric, one is able to implement distributed training with such strategies for complex training loops with only small tweaks to existing code. The framework also enables mixed precision training, which reduces the memory footprint and improves the speed of our model. We chose PyTorch Lightning Fabric over PyTorch Lightning because it is more flexible and compatible with our complex training loop. By using PyTorch Lightning Fabric, we can focus on the core logic of our model and leave the scaling details to the framework. Our preliminary results show we were able to significantly speed up model training and evaluation by 2-3x fold. This not only saves time but also allows for more efficient use of computational resources. This allowed us to overcome the limitations of a complex training loop and achieve faster and more efficient model training. Overall, the use of distributed and parallelized training in PyTorch using PyTorch Lightning is crucial for large-scale machine learning models.

5 Experiments

5.1 Experimental Settings

To incorporate the multi-modal information to the learning process, we created a cross-modal matching branch, which is trained to match the multi-modal information and the label from the set of target categories. Given a pair of an input sample $\boldsymbol{x}$ and a category y, we first extract a feature vector $\boldsymbol{h} = f_\theta(\boldsymbol{x})$ for $\boldsymbol{x}$. y is transformed into a one-hot vector and then transformed into an embedding vector via a linear layer that match the dimension of the multi-modal feature vector $\boldsymbol{l} = f'_\phi(y)$. Afterwards, $\boldsymbol{l}$ and $\boldsymbol{h}$ are concatenated and fed into a multi-layer perceptron generating a matching score $s(\boldsymbol{x}, y)$. When training the cross-modal matching head, positive examples can be collected from labelled instances, while negative samples are signthesized by making pairs of instances and categories which are not identical to the ground-truth label.

We adopt two metrics that are widely adopted to measure the OOD detection performance in machine learning literature [4, 13, 20].

- **AUROC**: the area under the receiver operating characteristic curve with the true positive rate (TPR) plotted against the false positive rate (FPR). Higher AUROC values indicate better OOD detection performance.
- **FPR95**: the FPR when the TPR is 95%. Lower FPR95 values indicate better OOD detection performance.

We evaluate all configurations of contrastive losses, OOD scores, and the presence of image embeddings. Those include 16 settings composed on 2 alternative setups for contrastive losses, 4 alternatives of OOD scoring functions, and 2 cases with or without image embeddings.

We implement our model using the Transformer python package [17] and build the document image classifier based on LayoutLM [19] in the main experiement. All models are optimized with Adam [10] with a learning rate of $1e - 5$, batch size 32 for the heath insurance claim data and learning rate of $1e - 4$, batch size 64 for the RVL-CDIP data. For both datasets, we use a linear learning rate decay towards zero and fine-tune the model for 10 epochs. For training process, we use the validation split for hyper-parameter tuning. The resulting hyper-parameters we used are $\tau = 0.3$ and $\lambda = 1.1$ for health insurance claim data and $\tau = 0.3$, $\lambda = 2$ for RVL-CDIP data.

5.2 Experimental Results

This section presents the main results of our framework, including the document classification accuracy and OOD detection performance. Due to constraints in time and computing resources, we provide proof-of-concept results based on the health insurance claim data. For experimental results on RVL-CDIP, please refer to Appendix B. Further results on the health insurance claim data can be found in Appendix A.

ID Classification Test accuracy

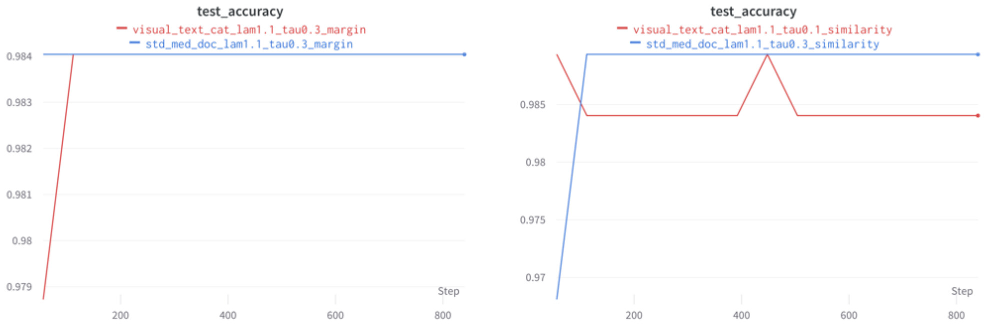


Figure 4: The classification accuracy on test datasets by two different contrastive loss: margin-based contrastive loss (left), supervised contrastive loss (right).

Note that the incorporation of OOD detection techniques does not seem to negatively impact the main task classification performance. This can be observed in Figure 4, where the blue lines represent test accuracy trained on the text-layout data only, and the red lines represent test accuracy after incorporating image information. We also found that adding image embeddings does not significantly affect test accuracy when using margin-based contrastive loss. However, incorporating multi-modal information with supervised contrastive loss appears to have a negative impact on test accuracy.

5.2.1 OOD Peformance

Table 1 and 3 present selected results on OOD detection performance metrics for both supervised and margin-based contrastive losses. The complete results can be found in Appendix A. Specifically, for the Mahalanobis distance and the MSP-based OOD scores, it can be observed that incorporating the image data enhances the OOD detection performance by simultaneously achieving a higher AUROC and a lower FPR95.

Data	*Maha AUROC*	*Maha FPR95*	*MSP AUROC*	*MSP FPR95*
text-layout	0.9807	0.05286	0.9962	0.007526
image-text-layout	**0.9833**	**0.01035**	**0.9995**	**0**

Table 1: Supervised Contrastive Loss

Data	*Maha AUROC*	*Maha FPR95*	*MSP AUROC*	*MSP FPR95*
text-layout	0.9598	0.05738	0.9886	0
image-text-layout	**0.9649**	**0.01505**	**0.9995**	0

Table 2: Margin-based Contrastive Loss

6 Conclusion and Discussion

In this project, we have validated the hypothesis that incorporating multi-modal information can improve OOD generalization for document classification based on pre-trained Transformers. We conducted a systematic investigation of the combination of contrastive losses, scoring functions, and the presence of multi-modal information. We presented meaningful proof-of-concept results based on one motivating dataset and demonstrated our efforts in overcoming computational challenges by re-implementing our model with PyTorch Lightning. Our findings offer valuable insights into the computational burdens when using distribution-based OOD scores, such as Mahalanobis Distance, for large-scale data. Additionally, our proposed framework is generalizable to other pre-trained transformers, such as LayoutLM2 [18] and LayoutLM3 [8], which have built-in model architectures for dealing with multi-modal information. This can be particularly beneficial when dealing with more complex NLP tasks, such as NER.

Individual Contribution

Shuxian Fan: problem formulation, coming up with the algorithm, coding up the algorithm, running tests, tabulating final results, and writing up the proposal and project report. Hao Li: writing up the proposal and project report, drafting slides, and brainstorming alternative losses. Warren Paris-Moe: implementing model, optimizing model, running tests, and tabulating final results. Yilun Xing: writing up the proposal, data processing, and drafting slides.

Acknowledgments

We thank the department of Computer Science & Engineering at the University of Washington for providing this research opportunity (CSE 547: Machine Learning for Big Data). We also thank for the help from Professor Tim Althoff and TA, Esteban Safranchik.

References

[1] American Family Insurance. Esi claim form, 2023. [Online; accessed Mar 12, 2023].

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

[4] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[5] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.

[6] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[7] Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: harvesting ood data with cross-modal matching for open-set semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8310–8319, 2021.

[8] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[9] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Mattias Langer, Zhen He, Wenny Rahayu, and Yanbo Xue. Distributed training of deep learning models: A taxonomic perspective. *IEEE Transactions on Parallel and Distributed Systems*, 31(12), 2020.

[12] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*, 2019.

[13] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31:2802–2818, 2018.

[14] Roots Automation. Iris cms-1500 document reader, 2023. [Online; accessed Mar 12, 2023].

[15] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022.

[16] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

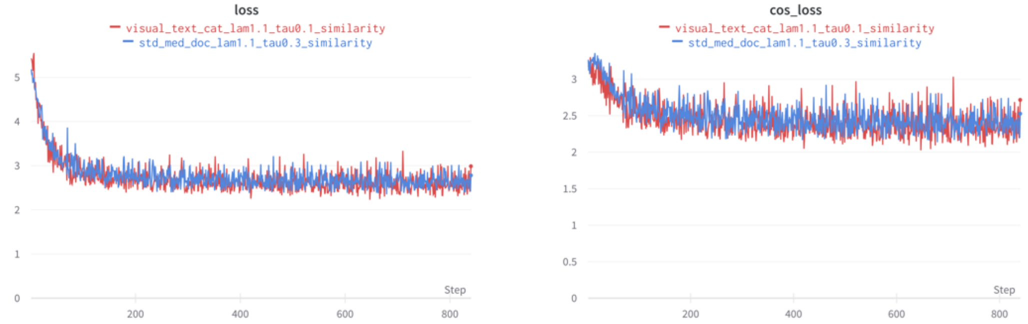
[18] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020.

[19] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020.

[20] Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive out-of-distribution detection for pretrained transformers. *arXiv preprint arXiv:2104.08812*, 2021.

A Additional Results for Health Insurance Claim Forms

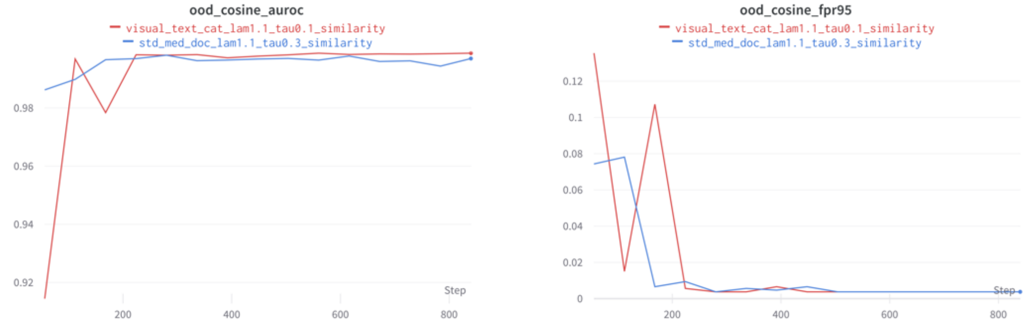
A.0.1 Supervised Contrastive Loss



(a) Total and contrastive losses.

(b) Validation and test classification accuracies.

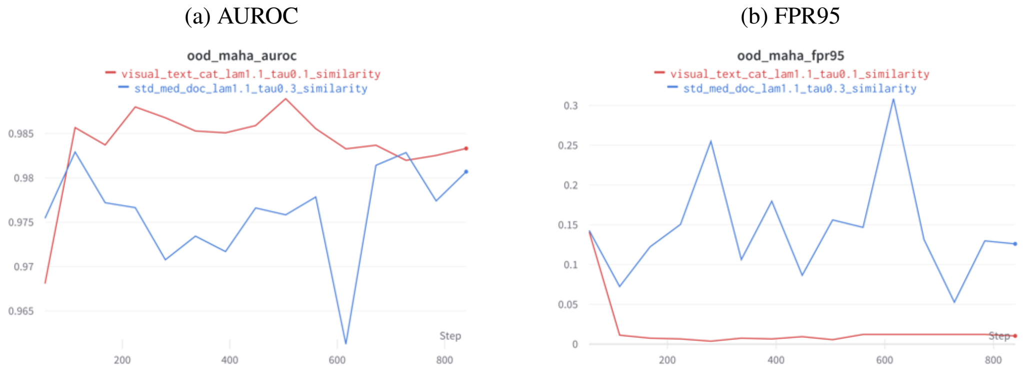
Figure 5: Results of the ID classification performance trained on text-layout data (blue) and multi-modal text-layout-image data (red) with supervised contrastive loss.



(a) AUROC

(b) FPR95

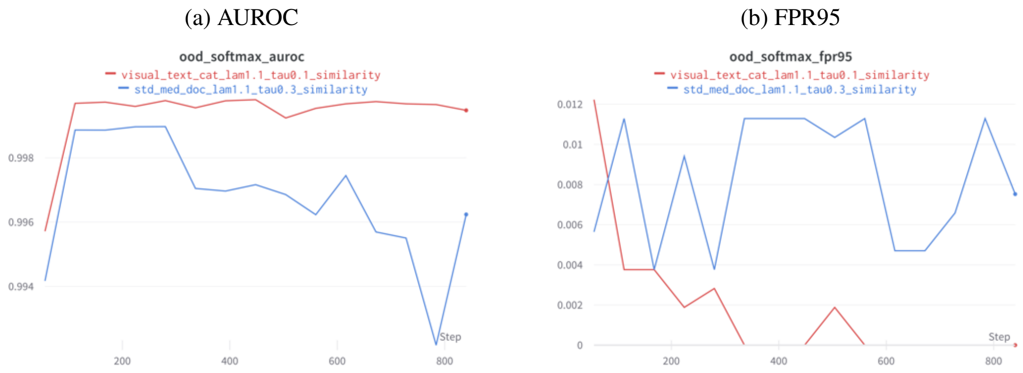
Figure 6: OOD performance for Cosine-similarity-based score with supervised contrastive loss.



(a) AUROC

(b) FPR95

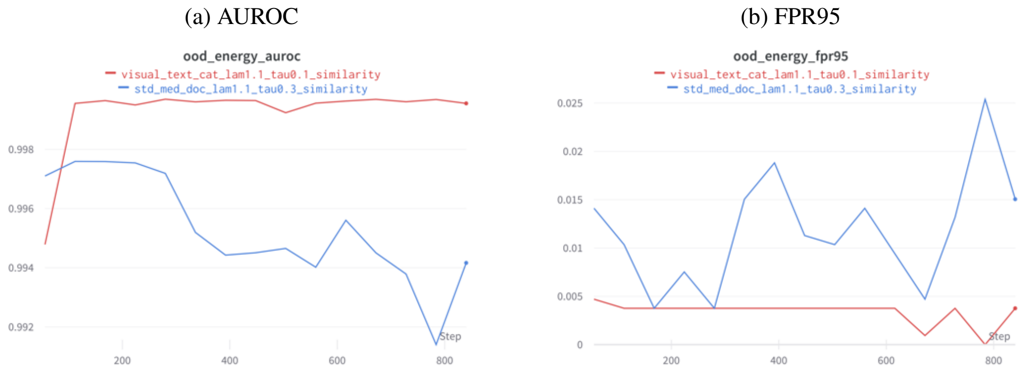
Figure 7: OOD performance for Maha-distance-based score with supervised contrastive loss.



(a) AUROC

(b) FPR95

Figure 8: OOD performance for maximum softmax probability score with supervised contrastive loss.



(a) AUROC

(b) FPR95

Figure 9: OOD performance for energy score with supervised contrastive loss.

A.0.2 Margin-based Contrastive Loss

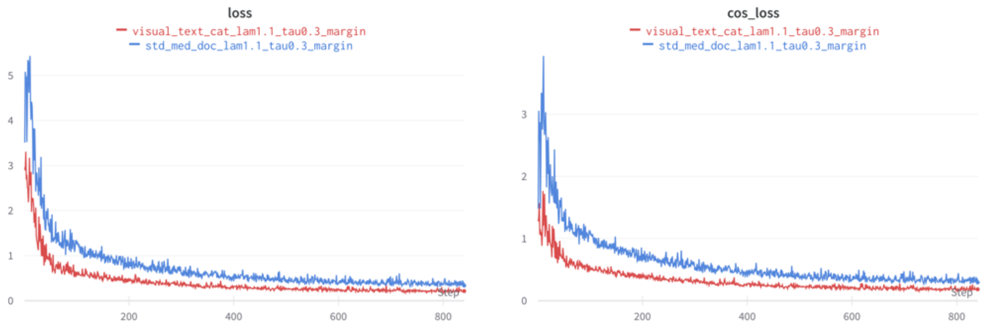
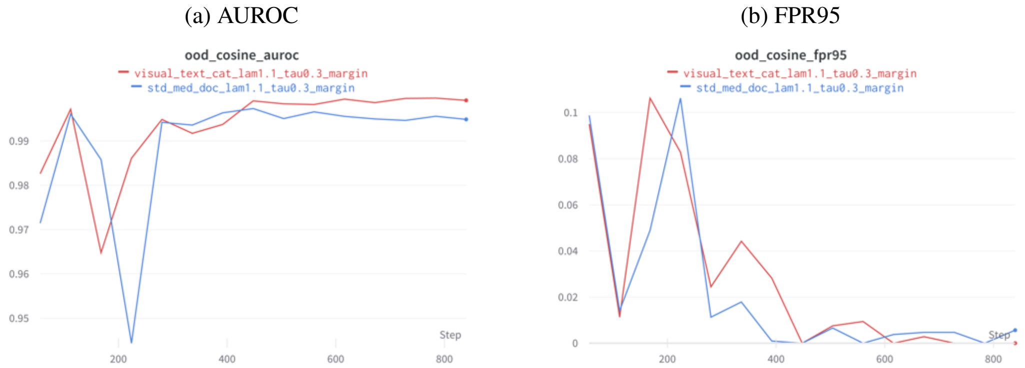


Figure 10: The total and margin-based contrastive loss values for training on the text-only data (blue) and multi-modal image-text data (red).

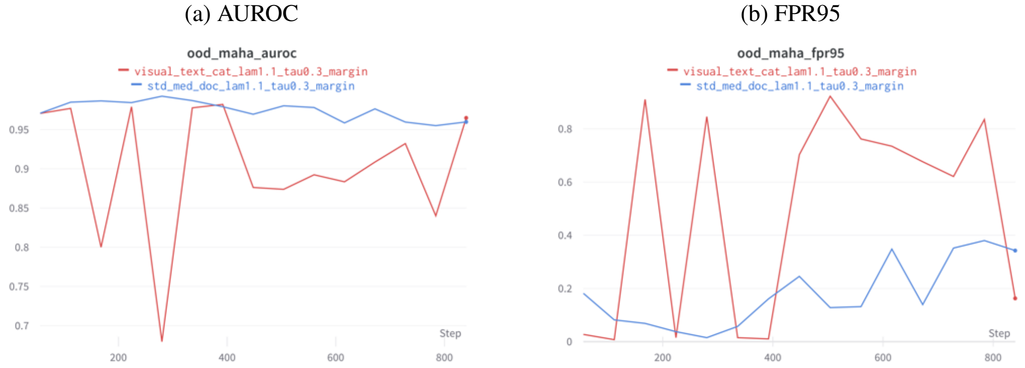
Performance on OOD Detection



(a) AUROC

(b) FPR95

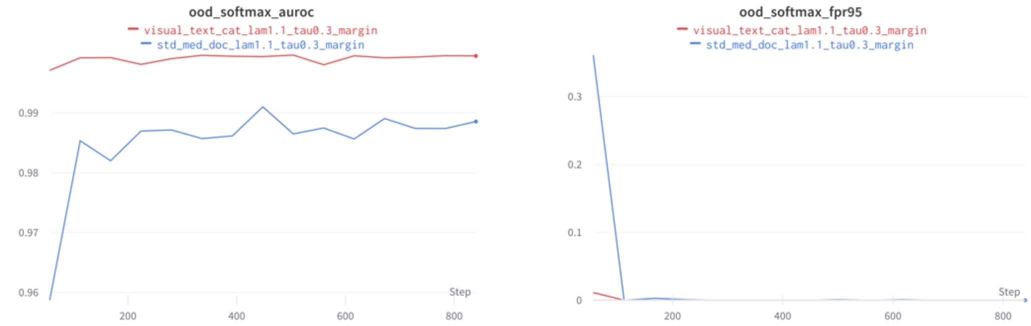
Figure 11: OOD performance for Cosine-similarity-based score with margin-based contrastive loss.



(a) AUROC

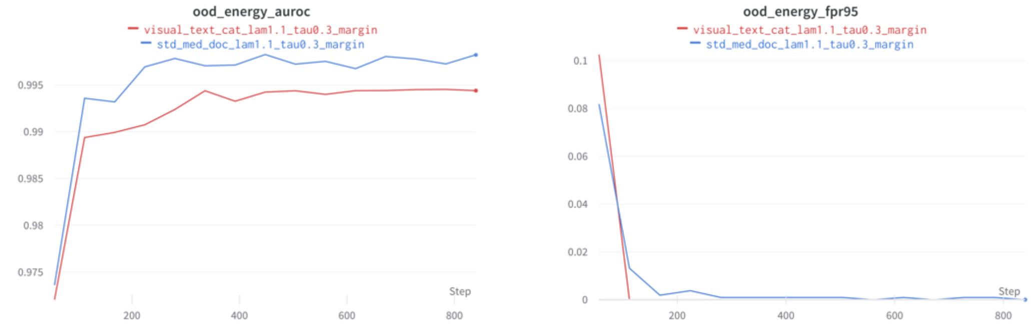
(b) FPR95

Figure 12: OOD performance for Maha-distance-based score with margin-based contrastive loss.


(a) AUROC

(b) FPR95

Figure 13: OOD performance for maximum softmax probability score with margin-based contrastive loss.


(a) AUROC

(b) FPR95

Figure 14: OOD performance for energy score with margin-based contrastive loss.

B Results for RVL-CDIP Data

Data	*Maha AUROC*	*Maha FPR95*	*MSP AUROC*	*MSP FPR95*
RVL-CDIP	0.9284	0.3638	0.8724	0.8890

Table 3: RVL-CDIP Dataset with 16 classes.