Pulmonary Disease Classification Using Globally Correlated Maximum Likelihood: an Auxiliary Attention mechanism for Convolutional Neural Networks

Journal:	*IEEE Transactions on Artificial Intelligence*
Manuscript ID	Draft
Manuscript Type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Verenich, Edward; Clarkson University Martin, Tobias; Clarkson University Velasquez, Alvaro; AFRL Khan, Nazar; University of the Punjab Hussain, Faraz; Clarkson University, Electrical and Computer Engineering
Keywords:	Artificial intelligence in medicine, Artificial neural networks, Convolutional neural networks, Deep learning

Pulmonary Disease Classification Using Globally Correlated Maximum Likelihood: an Auxiliary Attention mechanism for Convolutional Neural Networks

Edward Verenich, Tobias Martin, Alvaro Velasquez, Nazar Khan, and Faraz Hussain

***Abstract*—Convolutional neural networks (CNN) are now being widely used for classifiying and detecting pulmonary abnormalities in chest radiographs. Two complementary generalization properties of CNNs, translation invariance and equivariance, are particularly useful in detecting manifested abnormalities associated with pulmonary disease, regardless of their spatial locations within the image. However, these properties also come with the loss of exact spatial information and global relative positions of abnormalities detected in local regions. Global relative positions of such abnormalities may help distinguish similar conditions, such as COVID-19 and viral pneumonia. In such instances, a global attention mechanism is needed, which CNNs do not support in their traditional architectures that aim for generalization afforded by translation invariance and equivariance. Vision Transformers provide a global attention mechanism, but lack translation invariance and equivariance, requiring significantly more training data samples to match generalization of CNNs. To address the loss of spatial information and global relations between features, while preserving the inductive biases of CNNs, we present a novel technique that serves as an auxiliary attention mechanism to existing CNN architectures, in order to extract global correlations between salient features.**

***Impact Statement*—We improve sensitivity of Convolutional Neural Networks (CNNs) using an auxiliary global attention mechanism (GCML) that enables CNNs to utilize global spatial information similar to Vision Transformers (ViTs). Our technique retains the benefits of spatial invariance and equivariance inherent to CNNs, while allowing spatial information of features to be used as discriminators. GCML retains these inductive biases in data starved environments, which ViTs lack due their architecture, and hence require significantly more training data to achieve a similar level of generalization. Finally, we show impirically, that GCML improves the sensitivity of standard CNNs when classifying pulmonary conditions in chest X-rays. We provide all associated code, data, and models for reproducibility and improvement through further research.**

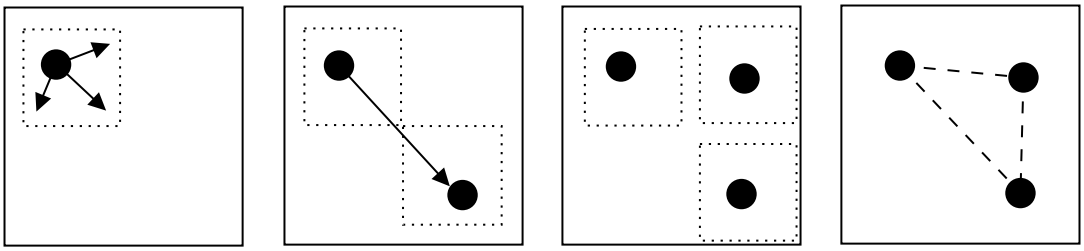
***Index Terms*—COVID-19 detection, convolutional neural networks, global attention mechanism, data starved environment.**

I. Introduction

With the emergence of the COVID-19 pandemic, the use of Convolutional Neural Networks (CNNs) to detect presence of pulmonary diseases in medical imagery has quickly gained momentum, where a significant majority of approaches center around fine-tuning pre-trained CNNs on new data, as reported by Roberts et al. [1]. The benefits of utilizing such techniques are intuitive, including the ability of CNNs to detect features that are difficult for humans to identify, learning certain correlations between positive cases, and the speed with which such predictions can be made in order to aid in timely diagnosis. CNNs have been shown to outperform individual radiologists in detecting pneumonia in chest X-rays as reported by Rajpurkar et al. [2]. In that work, X-ray based pneumonia diagnoses made by a group of four radiologists were compared to a custom CNN, using the F1 metric, where only one radiologist performed better than the model. *This result prompted us to explore a possible mechasism of visual analysis of X-rays by a radiologist who outperformed the CNN, and whether that might translate into more accurate CNNs for classification of pulmonary diseases.*

Our hypothesis is that while human radiologists may not be as effective as CNNs at identifying individual salient features, their ability to quickly consider *global spatial relations* between those features may play a factor. Our intution for this came from recent work by Borghesi et al. [3], where an experimental scoring system for chest X-rays was used by radiologists to quantify and monitor disease progression in COVID-19 patients. The scoring worked by dividing the frontal X-ray of the lungs into six zones, where each zone was assessed with a quantitative score ranging from 0, representing no lung abnormalities, through 3, representing the highest severity of abnormalities. To obtain the final score, each region's scores were summed, and used as a measure of COVID-19 severity on the lungs. *Our idea is to enable the CNN model to track abnormalities and their relations to each other across regions, and base our predictions on these learned spatial relationships and not just a cummulative score,* similar to what a human subject matter expert might do to distinguish different diseases.

In the remainder of this section, we briefly describe how image classification using CNNs works, including their strengths and limitations, with the goal to introduce the notion of *additive* classification that we argue standard CNNs perform. We then discuss how certain positive generalization properties of CNNs limit their ability to account for global spatial correlations between regions of an image, thereby lacking the ability to utilize positional information as a discriminating factor.



(a) CNN Equavariance (b) CNN Invariance (c) CNN Multi-object (d) GCML Attention

Fig. 1: Examples of inductive biases inherent to Convolutional Neural Networks *(a,b,c)*. Small translational shifts in a target object (black circle) within the receptive field of the kernel (dotted square) will shift the activation equally, and will still classify the target correctly due to *(a)* translation equivariance. Major translational shift of the target object (black circle) to a new spatial position within the image will also result in the correct classification, regardless of its global position, due to *(b)* translational invariance. Due to the scoring function, presence of multiple target objects *(c)* within the image will result in a classification that the object is present in the image, regardless of quantity or spatial locations of the objects within the image. Our GCML attention mechansim *(d)* uses spatial information of features, and their global interrelations, to distinguish between classes that can exhibit the same features in different locations, similar to Vision Transformers (ViTs) [4], but with significantly fewer training samples needed by ViTs to achieve similar generalization. GCML achieves this by not performing tokenization of input images, but rather tokenizing class activation maps generated by the CNN with respect to the class of input images.

A. Image Classification with CNNs

Yamashita et al. [5] provide an excellent overview of CNNs and their application in radiology, while we provide a general summary. In a feed forward Convolutional Neural Network, image classification is performed by propogating input through a series of convolutional layers, where filters that were previously trained to recognize specific features of an image get activated and are used as input to subsequent layers. Typically, this process involves downsampling or reducing the mapping resolution of activated output with respect to input, as image data moves further through the layers. This is done by utilizing convolutional filters of size $k > 1$, but significantly less than the size of the input. By learning local correlation structures within the bounds of the window defined by k, particular convolutional layers learn specific local features. During inference, this allows the network to recognize objects or larger features, that are compositions of smaller features, which were recognized by earlier layers. After the last convolutional layer, 2 dimensional activation maps are flattened to 1 dimensional data and are passed to a fully connected layer that tallies the contributions of activated filters with respect to classes represented in the final connected layer. A classification can then be made by taking the maximum tally, or top K tallies for top K classification. Additional layers, such as a layer representing a *SoftMax* function, can be utilized after the final connected layer to obtain class probabilities from the joint distribution of represented classes, but the core process described above is the same.

The use of convolutions in CNNs provides two important generalization properties with respect to image classification, *translation equavarience and invariance*, where a convolution is equavarient to translation if an object in an image is spatially shifted, convolution's output is equally shifted. Invariance is a result of a position-independent pooling operation that follows a convolution, and results in a loss of absolute location, but enables the visual inductive prior of convolutional operators [6]. At a higher level, they enable features to be reliably detected regardless of their spatial location within an input image.

CNNs became the go-to method for image classification after Krizhevsky et al. [7] published their results on *ImageNet* [8]. Within the medical domain, CNNs started to be employed for diagnostic assistance of pulmonary diseases through classification of medical imagery, including CT scans and X-rays [9] [10]. With the rise of the COVID-19 pandemic, much work has been done in applying CNNs for image classification of medical imagery in order to rapidly detect pulmonary manifestations of COVID-19 in chest X-rays [11]. Much of this work, which we further discuss in Section II, utilizes CNNs for classifying chest X-rays in a manner we described above. For the purposes of this work, we refer to this method of image classification as *standard* or *additive*, because classification is performed by summing contributions of various filters at the final connected layer of the network.

B. Limitations of Image Classification with CNNs

Even with state-of-the-art results in image classification and object detection, CNNs have certain limitations relevant to image classification in certain domains. For example, Hosseini et al. [12] report degraded image classification performance on images with reversed brightness, or *negative* images. This may be mitigated by proper pre-processing of data, but it may signal that color channel information or texture may be learned by the network, along with shapes. Although not a limitation in itself, it introduces additional considerations when evaluating network performance on new and out-of-distribution data.

The main limitation that we address in this work has to do with the *localized* perceptive fields of convolutional filters. Although CNNs exhibit translational invariance and equivariance, which improves generalization in many instances, it can

also hurt generalization through loss of spatial information within the pooling layers of the CNN [13]. The canonical example of this problem relates to classifying an image of a human face, where features such as nose, eyes, lips are identified, yet placing them in different parts of the image in a manner that does not resemble a face can still yield a face classification [14]. Relating this to the chest radiographs domain, small manifestations of pulmonary disease are identified by a CNN, but their spatial interrelationships are mostly lost. In the natural imagery and radiology domains, the loss of spatial information, due to the contraints imposed by convolutional filter size and pooling, can result in incorrect classifications. Figure 1 provides a visual intuition to the inductive biases that CNNs exhibit in terms of identifiying target objects or certain features, for example abnormalities in X-rays. While CNNs are particularly good at detecting abnormalities regardless of their spatial locations, the scoring function used for classification in standard CNNs, shown in Eq 1, does not consider spatial locations nor spatial relations between features as discriminators. We show in this work, that accounting for global spatial interrelation of features improves classification of pulmonary conditions.

C. Contribution

Our main goal in this work is to provide empirical evidence to the hypothesis that accounting for global correlations between activated regions of an image improves classification of pulmonary conditions in chest radiographs. This improvement also extends to image classification domains where discriminating a target class involves accounting for global spatial correlations between features. Our secondary goal is to show that the classification approach, using our novel Globally Correlated Maximum Likelihood (GCML) auxiliary attention mechanism, is competitive to standard CNN classification, while utilizing significantly fewer model parameters, less computational resources, and much fewer training samples. To that end, our contributions are the following:

- *We developed a novel auxiliary attention mechanism, GCML, that is utilized with existing Convolutional Neural Network architectures in order to account for global spatial correlations between salient features of represented classes.*
- *Using GCML, we show competitive results for image classification on a benchmark dataset, CIFAR-10, using significantly less model parameters and computation, while not utilizing any pre-training on samples in relevant domains.*
- *We show that by utilizing the GCML attention mechanism, we improve image classification, specifically increasing model sensitivity or recall of pulmonary diseases in chest radiographs. This also includes effective generalization on a previously unseen dataset, suggesting that the GCML attention mechanism improves and complements visual inductive priors learned by CNNs by accounting for spatial relations.*
- *Our results show that standard CNN and our GCML technique for image classification have particular strengths, and show potential utility when utilized as ensemble methods.*
- *Finally, we provide an open source[1] reference implementation of our technique, allowing further research into its improvement and utility.*

The remainder of this paper is structured as follows. Section II describes work related to image classification of pulmonary diseases using medical imagery. Section III outlines our approach, while briefly discussing work relevant to attention mechanisms. We report results of our experiments in Section IV, including a benchmark dataset of natural imagery and two separate datasets of chest radiographs. Finally we conclude by discussing our results, limitations of our approach, and future research directions.

II. Related Work

In this section we describe work related to pulmonary disease classification using chest X-rays or CT scans. Some approaches mentioned also include forms of weakly supervised localization, which in contrast to object detection in imagery, does not require that training labels be accompanied with spatial coordinates of objects to be detected. These types of training labels are also referred to as *image level* labels, as they only provide information on whether certain target classes are present in the image, not their spatial locations. To the best of our knowledge and at the time of this work, CNN attention mechanisms have not been used for the purpose of improving image classification of pulmonary diseases.

Rahaman et al [15] employed transfer learning to fine-tune several CNN architectures to classify chest X-ray images into three classes: COVID-19, Healthy, and Pneumonia. A total of 860 images were used in their study, which reported the VGG19 [16] architecture having the best performance achieving an accuracy of 89.3 percent, average precision of 0.90, a recall of 0.89, and F1 score of 0.90. Khan et al. [17] proposed CoroNet, a network based on the Xception [18] architecture for diagnosis of COVID-19 from chest X-rays. They used a dataset of X-ray images [19] to train their network, achieving 95 percent accuracy when classifying images for COVID-19, Normal, and Pneumonia. Tamal et al. [20] used their radiology classification model for COVID-19 classification with low severity, as scored by radiologists, on new data from patients at a local hospital, showing generalization to out-of-distribution data and achieving an overall accuracy of 90 percent. Kim et al. [21] reported relevant findings on the effect of dataset composition practices on classification performance. The authors reported that higher classification performance was observed on datasets where data composing each class came from different sources. Heidari et al. [22] showed that their image preprocessing scheme improved pulmonary disease classification in X-ray images. Similar diagnostic aid approaches to classifying COVID-19 were reported in [23], [24], and [25], where the use of transer learning using a pre-trained Convolutional Neural Network architecture to perform multi-class classification was the common factor.

[1] https://gitlab.com/verenich/gcmlpub

In addition to classifying images as representing pulmonary conditions, work has been proposed to provide some explainability of those classifications. Tsiknakis et al. [26] utilized the Inception [27] architecture to first classify X-ray images into specific diseases, and then applied a weakly supervised localization technique GradCAM [28] to identify regions within the image responsible for a particular classification. Wang et al. [29] proposed a similar approach to diagnose and localize disease manifistation in X-rays. Verenich et al. [30] proposed a method to reduce aleatoric uncertainty in weakly supervised localization that can arise from significant class overlap between features associated with similar pulmonary diseases. Gupta et al. [31] proposed an approach that classifies and performs weakly supervised localization using standard Class Activation Maps [32].

Recently, Transformers, originally proposed by Vaswani et al. [33], have shown state-of-the-art performance on natural language processing tasks. These ideas have been utilized in Vision Transformers (ViT) [4] to introduce positional attention mechanisms for image classification, attaining results comparable to state-of-the-art CNNs. However, ViTs lack translational invariance and equavarience of CNNs, thus require significanly more training data to generalize [4]. To the best of our knowledge, and at the time of this work, ViTs have not been used for classification of pulmonary diseases using X-rays. One possible reason for this is insufficient amount of training data, to achieve same levels of generalization as CNNs with the data that is currently available.

III. Approach: GCML Attention Mechanism

The goal of our attention mechanism is to preserve spatial interrelationships of activated regions in a given image I relative to target classes C learned by a convolutional neural network G. Our hypothesis is that localized pulmonary abnormalities, detected by the convolutional neural network in different regions, can yield additional discriminative power when their global interrelationships are considered. The main intuition for this is that translation invariance and equivariance properties of the CNNs result in class discrimination that is additive, with respect to activation strengths of convolutional filters, but not their spatial relation to each other. This section describes the architecture of our approach, that accounts for spatial interrelations of salient features.

A. Attention Type

Our technique explores complementing convolutional neural networks with attention mechanisms. The main difference between prior work [34]–[37] and GCML is that *we do not alter the architecture of the CNN* using self-attention layers, but instead provide a separate auxiliary structure that is created using a pre-trained network. In addition, the stochastic structure of GCML does not require it to be trained simultaneously with the CNN using *backpropagation* as in [37], thus it does not have to be differentiable and is significantly faster to train. Finally, tokenization of input images into patches is also not required, instead we use downsampled output of the last convolutional layer, scaled by class weights from the final connected layer, to generate class activation maps, as proposed by Zhou et al. [32]. These class activation maps are used as input to our GCML attention mechanism. In addition to being effective at performing weakly supervised localization, class activation maps were also used to distinguish overlapping regions of images that may belong to different but overlapping classes [30].

The work that is in closest alignment with our approach is the UL-Hopfield model [38], which uses pre-trained CNNs with an associative memory bank to perform image classification. Our approach differs in several ways: the UL-Hopfield auxiliary memory learns class-specific *core patterns* from a pre-trained CNN in an unsupervised manner, while we use the CNN to generate a pattern using training labels. Second, the type of input to the attention function that is used to train their memory structure is extracted from the last pooling layer without weighting activated feature maps by a specific class, which is done in our approach during training. Finally, for experiments we use a CNN with 42x less parameters than the CNN they use for feature extraction, as we discuss in Section IV.

B. Input Features

Formally, to compute the attention function input tensor M_c, where c is a class represented in the CNN, we do the following: given an input image, let $f_k(x, y)$ be the activation of filter k at the last convolutional layer of of a CNN G and (x, y) be the spatial location. During classification, for each filter k, a pooling layer outputs a global average F_k defined as $\sum_{x,y} f_k(x, y)$, which is then used as input to the fully connected layer, giving us the class score S_c in Eq 1, where w_k^c is a scalar weight indicating the importance of F_k to class c.

$$S_c = \sum_k w_k^c F_k \tag{1}$$

To compute each spatial element of the class activation map [32], or 2-dimensional tensor M_c, we use

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \tag{2}$$

The resulting tensor M_c effectively splits the input image I into regions or tokens, where entries $M_{i,j}^c$ represent activation intensities of those regions for class c.

Our technique would be equally applicable to be used with class activation maps generated by another weakly supervised localization technique called Grad-CAM, proposed by Selvaraju et al. [28]. It requires that we compute the gradients of output of the network with respect to feature map activations for each class c, making it slower than the standard CAM method for the purposes of training the GCML structure on large datasets.

C. Attention Function

Similar to the original work on transformers [33], our attention function is a mapping of a query Q, based on a

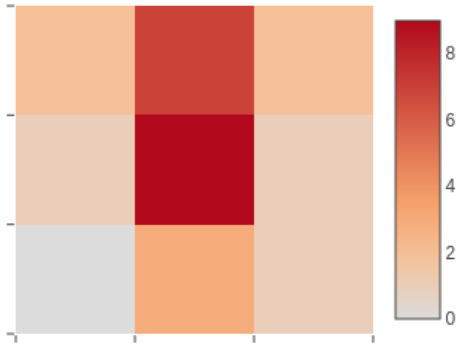


Fig. 2: Example of a 2-dimensional tensor M, which is the same as the class activation map, that is used as an input to the attention function Q. The dimensions of M are dictated by the mapping resolution of the convolutional layer that we use to compute the class activation map. Values at $M_{i,j}$ represent activation intensities for class C after an image I is passed through the network G.

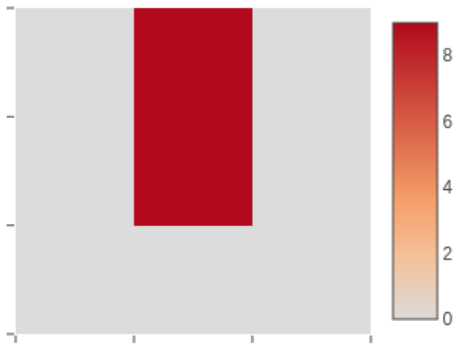


Fig. 3: Example of a 2-dimensional tensor M in its intermediate state after normalization and application of threshold τ as computed within attention function Q. Hyperparameter τ is also optimized during the training stage of the GCML structure and its optimized value is persisted for inference.

datastore key K to a retrieved value V. In our case, the query function Q takes as input a real tensor M of size $H \times W$, which maps it to a datastore key K, to retrieve likelihood value V. In training mode, when learning the GCML structure, key K is used to update the value at that index, while during inference mode, it is used to retrieve V at position K. In both cases, attention function Q remains the same. The states of input, intermediate states, and output of Q are illustrated in Figures 2, 3, and 4. Parameter τ is used as a threshold to convert input M_c to its intermediate state shown in Fig 3 using Equation (3).

$$f_\tau(M_{i,j}) = \begin{cases} 1, & \text{if } M_{i,j} \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The full algorithm for obtaining datastore key K from tensor M_c using the attention function Q is described in Algorithm 1.

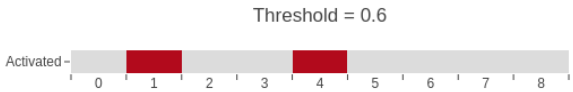


Fig. 4: Intermediate state of tensor M is flattened to a vector B where entries B_i represent bits that are set to 0 or 1. Converting this vector of bits to an integer yields our datastore key K, which would be 144 in the vector displayed. This value is the output of the attention function Q. Note that given the length L of vector B, the number of possible entries K_i is 2^L. Also, either big or little endianness can be used for bit arrangment, as long as it is consistent throughout training and inference.

Algorithm 1 Attention function Q

procedure Q(M_c) ▷ Computes K from M_c
  $M_c' \leftarrow normalize(M_c)$
  $M_c'' \leftarrow f_\tau(M_c')$ ▷ Threshold τ
  $B \leftarrow flatten(M_c'')$ ▷ From 2-d to 1-d
  $K \leftarrow binToInt(B)$
  return K ▷ Datastore key K
end procedure

D. Training GCML

The GCML datastore S is a tensor with dimensions $C \times P$, where C is the number of classes represented in network G and P is of size 2^L, where L is the length of the flattened vector B from which key K is derived. For a given class y, and possible activations x of $M^y_{i,j}$, being in on or off states, *each row* of S represents a discrete *Conditional Probability Distribution* of class y given activations x, as shown in Equation (4).

$$S[y] = P(y|x_0, y|x_1, \ldots, y|(x_0, x_1, \ldots, x_L)) \tag{4}$$

To compute these likelihood distributions for all classes, we utilize a fully trained convolutional neural network G along with the training data set D_T, which was used to train G.

Algorithm 2 GCML training procedure (1 epoch)

procedure UPDATE(S, D_T, G) ▷ Update likelihoods in S
  for i, l_c in D_T **do** ▷ Image i, with label l_c
    $M_c \leftarrow G_m(i, l_c)$ ▷ Compute M_c via G and Eq 2
    $K \leftarrow Q(M_c)$ ▷ Compute key K
    $S[l_c][K] \leftarrow +1$ ▷ Increment state
  end for
  $S \leftarrow normalize(S)$ ▷ Optional normalization
  return S ▷ GCML store S
end procedure

Algorithm 2 shows the training procedure for the GCML datastore. For simplicity, we show the procedure for single images sequentially. In practice however, we implement these procedures on batches of images provided by a dataloader. We also note that none of the weights in network G are being updated during this procedure, and the network is set to evaluation mode. The procedure $G_m(i, l_c)$ involves propagating

the image through network G, and computing M_c for that image using class label l_c. Finally, the normalization procedure of S is marked as optional because this normalization can also happen before inference using GCML is performed. One reason to hold off on normalization, is to allow S to be further trained on additional data, as we will discuss later in this paper.

Another consdideration when training the GCML structure is to account for data transformations that were performed while training the original network G. For example random crops or flips along a horizontal or vertical axes may alter the class activation map, thus using more epochs with random transformations should improve the GCML structure in some domains, such as natural imagery of CIFAR-10. We note however, that in a domain such as chest radiology, major transformations such as flips will actually degrade performance as orientation of X-ray images is consistent and generalizing to such tranformations is not needed and is actually harmful. Therefore, in order to improve generalization of attention mechanisms, appropriate transformations should be considered based on the target domains.

E. Inference with GCML

To perform inference using the GCML attention mechanism we utilize a trained Convolutional Neural Network G along with the GCML datastore S. As mentioned earlier, the attention function Q remains the same during training and inference, the main difference is that during inference we compute M_c for *all* classes represented in G instead of just the class label provided with training data. Additionally, before inference is performed, we must make sure that the datastore S is normalized, as the normalization step during training shown on line 7 in Algorithm 2 is optional, in order to enable further training.

Algorithm 3 GCML inference procedure

1: **procedure** PREDICT(I, G, S) ▷ Predict class on image I
2: $\vec{M_c} \leftarrow G(I)$ ▷ Tensors M_c for all classes in G
3: $\vec{K_c} \leftarrow Q(\vec{M_c})$ ▷ Compute K for all classes in G
4: $\vec{V_c} \leftarrow lookup(S, \vec{K_c})$ ▷ Get class likelihoods
5: $C_P \leftarrow argmax(\vec{V_c})$ ▷ Get Max Likelihood class
6: **return** C_P ▷ Return predicted class
7: **end procedure**

Algorithm 3 shows the inference procedure of our approach, which expects datastore S to be normalized. For each image I, we propagate it through the convolutional neural network G, where we compute inputs $\vec{M_c}$ to the attention function Q, where each item in $\vec{M_c}$ is a class activation map computed for each class represented in G, given input I. In other words, single input I will generate N inputs to the attention function, where N is the number of classes in G. The next step is to compute vector $\vec{K_c}$, where each entry K_i is a key to a likelihood value representing class i. Class likelihood values $\vec{V_c}$ are then retrieved from S and maximum value is taken to determine the class that input I belongs to.

Figure 5 shows a full view of the inference procedure using the GCML attention mechanism, as well as the standard inference process using the convolutional neural network. As shown in the diagram, in addition to two classifications, labeled (a) for standard CNN classification and (b) for classification using the attention mechanism, Q function input tensors $\vec{M_c}$ can also be used to perform weakly supervised localization (c) by upsampling them to the same size as the input image to produce a heatmap of relevant regions for a given class.

F. On Attention Input Size

The dimensions of the input M_c to the attention function Q are determined by the final mapping resolution of the last convolutional layer of network G. For example, ResNet50, a version of a widely used convolutional architecture utilizing residual layers [39], has a final mapping resolution of 7×7, when used with input images that are 224×224. This would result in GCML datastores with 2^{49} entries for each class, as this is the number of possible binary combinations of threshold activated regions within M_c of that size. Instead, we downsample the final resolution layer to a more manageable size, 4×4 and 5×5 as we will discuss in Section IV.

IV. Experiments

In this section we report the results of using our GCML attention mechanism to perform image classification. We performed experiments on three datasets: (1) dataset of natural imagery CIFAR10 [40], (2) COVID-19 radiology dataset [41] containing X-ray imagery of patients diagnosed with COVID-19, viral pneumonia, and no findings, (3) dataset of COVID-19, pneumonia, and no findings images taken from [19]. The purpose of the third dataset is to assess generalization of our method to new data.

The CIFAR10 dataset contains 60000 32×32 images representing 10 mutually exclusive classes, meaning each image belongs to only one class, which are: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Dataset authors note that automobile and truck classes do not overlap. The main reason we use CIFAR10 for our experiments is that the authors of the approach most similar to ours [38] provide extensive evaluation results, as well as being the largest dataset evaluated in that work.

As the main data for our experiments, the COVID-19 radiology dataset [41] contains a total of 2905 images, where COVID-19 cases are only represented in 219 of them, with the rest evenly distributed between viral pneumonia and images with no findings. This data set is particularly interesting to us as it represents both class imbalance and a data starved environment, as in such cases vision transformers do not outperform state-of-the-art CNN models [4] without pre-training on very large datasets in the similar domain.

To remain within reasonably accessible hardware constraints, all of our experiments were performed on a single machine with the following hardware characteristics: Intel Core i7 CPU, 64 GiB of RAM, NVIDIA RTX 2080 GPU. We used Ubuntu 20 as the operating system, NVIDIA CUDA 11 GPU acceleration library, and Pytorch 1.9.0 as our model implementation framework.

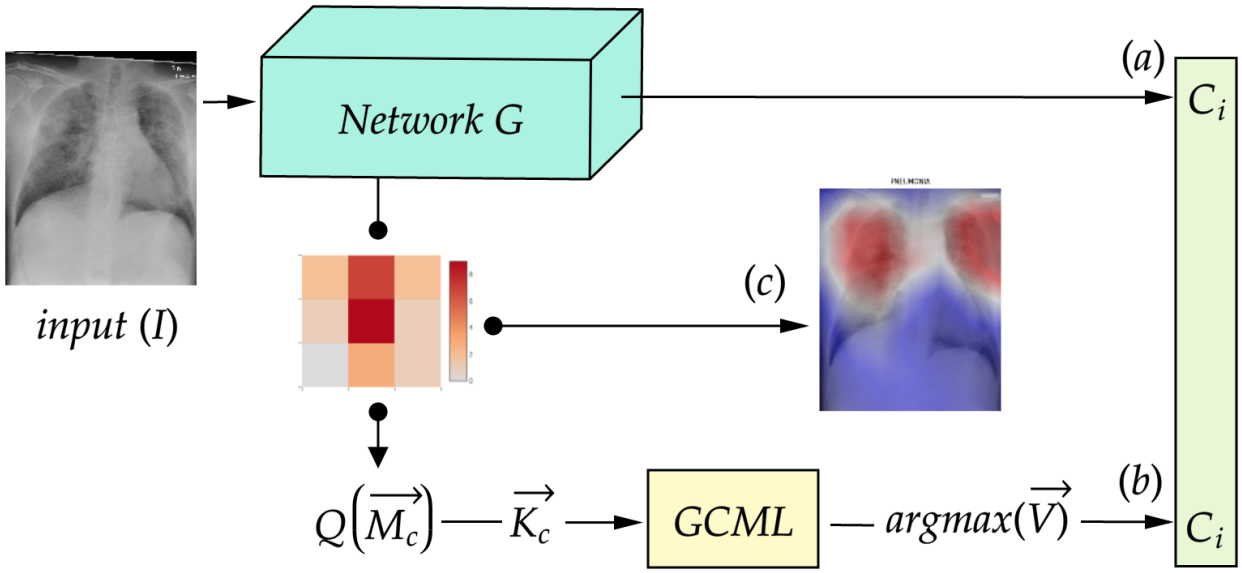


Fig. 5: Inference flow using a Convolutional Neural Network G and the auxiliary attention mechanism GCML. For a given input image (I), outputs (a) and (b) represent classifications made by the Convolutional Neural Network through its standard classification layer (a), and using the GCML attention mechanism (b). Both classifications can be utlized as an ensemble in a single forward pass during inference. In addition, input tensor M_c, to the attention function Q, can be upsampled using bilinear sampling to obtain a heatmap that performs weakly supervised localization of relevant regions for a given class as shown with output (c). We note that we include path (c) as an example of weakly supervised localization that can be performed using M_c, which we do not perform in our experiments. The input to the attention function Q is marked as a vector because 2-d tensors M_c are computed for every class represented in network G, hence a vector of 2-d tensors is passed to Q to compute a vector of class keys $\vec{K_c}$ that is then used to retrieve a vector of class likelihood probabilities $\vec{V}$ as shown in Algorithm 3. We also note that the block labeled as $GCML$ represents GCML datastore tensor S as shown in Algorithm 2.

A. CIFAR-10 Results

Here we describe our experimental settings and results on the CIFAR-10 benchmark dataset. As mentioned earlier, the model we use as our feature extractor is significantly smaller than that used in [38], where a pretrained ResNet50 architecture was used. We implemented a smaller version of a residual network architecture, which contains only 787,482 parameters compared to 33,554,432 for ResNet50. Additionally, we did not perform resizing of input images to 224×224, but instead used the original input size of 32×32. The mapping resolution of the last convolutional layer in our network is 4×4 compared to 7×7 utilized in [38]. The authors [38] did not fine-tune their feature extractor on the CIFAR-10 dataset, while we trained ours from scratch for 80 epochs on a single GPU (Nvidia RTX 2080) taking around 15 minutes or about 0.25 GPU hours. The only data augmentation we performed was padding input images by 4 pixels and randomly cropping at the original size of 32×32. The training phase of the GCML auxiliary attention mechanism took 28 minutes on the same machine with several τ values. Training the $GCML$ structure with different values of τ does not require any retraining of the CNN feature extractor. The test partition of the CIFAR-10 dataset (10,000 images) was not used for either of training phases. We also did not utilize a validation set during training of our feature extractor in order to replicate the training environment of [38] as much as possible.

Table I shows our results marked with (*) along with reported results in [38] and the state-of-the-art performance on the CIFAR-10 dataset reported in [4]. Only predictions made using the GCML data structure, marked (b) in Figure 5, were used. Our methods outperforms [38], but we note that the training of their memory module was unsupervised, even though their pre-trained feature extractor was significantly larger. One observation in these results is the significant reduction in parameters and extra training costs shown by our method, where our model has 800x fewer parameters than ViT-H/14 and 42x fewer parameters than UL-Hopfield while requiring no

TABLE I: CIFAR-10 Results for image classification using our GCML approach, UL-Hopfield network [38], and state-of-the-art vision transformer model [4]. We include vision transformer models to illustrate efficiency of our approach in terms of number of parameters and computational cost. The two transformer models were pre-trained on 300M images using cloud TPUv3 with 8 cores for 8 days [4], while our method took less than 1 hour on a single GPU. The numbers of trainable parameters are reported for each method according to standard practice.

Model	% Correct	Parameters	Extra Training Data
ViT-H/14 [4]	99.5	632M	300M
ViT-L/16 [4]	99.42	307M	300M
UL-Hopfield [38]	83.1	33.6M	1.28M
Res4Cif GCML*	**85.5**	0.79M	0

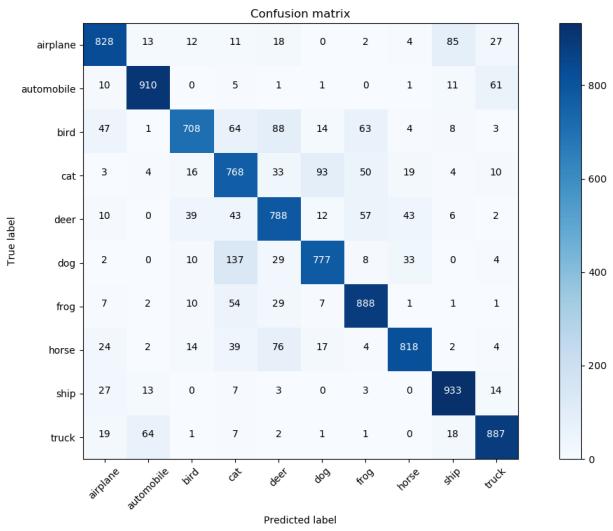


Fig. 6: Confusion matrix on the CIFAR-10 dataset reported in [38] We note that they report true labels on the y-axis of the confusion matrix.

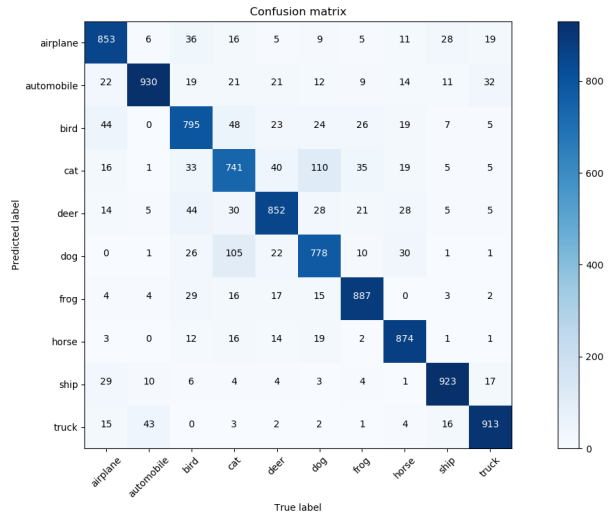


Fig. 7: Confusion matrix on the CIFAR-10 dataset using our Res4Cif and GCML with parameter $\tau = 0.001$. Our method outperforms UL-Hopfield networks [38] with significantly fewer parameters and without pre-training. We also note that in another experiment using $\tau = 0.009$, classification improved, as shown in Table II, implying further potential improvement.

extra data to pre-train. Both ViT-H/14 and ViT-L/16 are vision tranformer models, which are pretrained on large datasets. ViT-L/16 model, is pre-trained on the ImageNet-21k dataset, and is trained on cloud TPUv3 with 8 cores for approximately 8 days [4]. Our method had a combined training time of less than 1 hour on a single machine with 1 GPU.

Table II shows the performance of our method on CIFAR-10 at different values of the τ hyperparameter. Optimal values of this parameter depend on the the type of data normalization that is performed on the original data, and normalization methods of activated feature maps that are used to compute

TABLE II: CIFAR-10 Results for classification using GCML and different values of the τ hyperparameter. This threshold parameter is used to train the GCML structure as well as perform inference.

Hyperparameter τ value	Percent correct	Training epochs
0.3	76.59	80
0.1	82.21	80
0.05	83.6	80
0.009	**85.51**	80
0.001	85.5	80

TABLE III: COVID-19 Dataset [41] partitions. The validation partition is used as a model selection criteria during fine-tuning of the feature extractor model.

Data Class Samples	Training	Validation	Test
COVID-19	131	44	44
No Finding	804	269	268
Viral Pneumonia	807	269	268

input to our attention function Q. In this experiment we did not perform any normalization on the input data and used simple *min-max* normalization on activated feature maps. Figure 6 shows the confusion matrix on the CIFAR-10 dataset reported in [38] and Figure 7 shows our method's confusion matrix on the same data.

B. COVID-19 Radiology Results

The main goal of this work was to evaluate the potential benefit of attention mechanisms in identifying pulmonary conditions in chest radiographs that may be missed by spatial invariance of convolutional neural networks. Vision transformers are an active area of research and convolutional neural nets are currently the main approach for diagnostic assistance in the chest radiology domain [42]–[46], thus we focus on these techniques. Data in the radiology domain is not as well curated as in the natural imagery domain, with new diseases like COVID-19 presenting new classes to identify. This presents both, class imbalanced and data starved environments. Here we present our results of applying our method to the publicly available COVID-19 Radiology dataset [41].

For this experiment we randomly split the COVID-19 Radiology dataset into train, validation, and test partitions as shown in Table III. The validation set is used to select the best set of weights during the model G fine-tuning process, which is a standard practice in model training. Since validation data is never used to update weights of G during training, in the second part of our experiment we utilize the validation set to further refine the GCML structure to investigate the effect of updating just the attention mechanism without fine-tuning the feature extractor model on this data.

Since we have relatively few training points, we utilized transfer learning by fine-tuning a ResNet50 architecture that was pre-trained on the ImageNet dataset. The standard modification that is done during transfer learning is to remove the last connected layer that represents original classes and replace it with new target classes before fine-tuning with new

TABLE IV: COVID-19 dataset [41] combined results for accuracy, F1 score, and sensitivity using standard CNN, our GCML approach, and further tuned GCML^T structure using the test partition of the COVID-19 Radiology dataset [41].

METRIC (95% CI)	CNN	GCML	GCML^T
Accuracy	0.948 ±0.018	0.938 ±0.020	0.942 ±0.019
F1 Score	0.973 ±0.013	0.968 ±0.014	0.970 ±0.014
Sensitivity	0.962 ±0.016	0.955 ±0.017	0.958 ±0.016

data. One other modification that we performed was to reduce the final mapping resolution of the network from 7×7 to 5×5. This is done by adding a single convolutional layer with a kernel (filter) size of 3, padding of 0, and unit stride, while preserving the same number of in and out channels as the previous convolutional layer. This follows from simple ouput resolution calculus for a given dimension of a CNN convolutional layer, where for any equilateral input i, kernel size k, padding p, and stride $s = 1$, the output resolution is given by $o = (i - k) + 2p + 1$.

We then fine-tuned all parameters of this architecture using the training partition of the COVID-19 dataset [41] for 30 epochs, while using the validation partition to keep track of the best performing weights, selecting them as our final feature extractor model. Fine-tuning was done using the following settings and hyperparameters:

- input was resized to 224×224 through a random resize crop and normalized using *ImageNet* mean and standard deviation values
- *Stochastic Gradient Descent* was used as our optimization function
- *Cross Entropy Loss* was our training criterion
- *learning rate* was set to 0.001
- *momentum* was set to 0.9

For our next step we trained the GCML structure using the training partition of [41] and our convolutional neural network from the previous step using several *cam activation points* or τ values for 15 epochs each, with the best performing value on the test portion being $\tau = 0.05$. Training for multiple epochs was done because we used the same *random resize crop* data transform that was done during training of our feature extractor model, presenting slight positional variations. Finally, we run our method on the test partition of the COVID-19 Radiology dataset keeping track of both classifications, standard CNN and GCML attention mechanism, as shown in paths (a) and (b) of Figure 5.

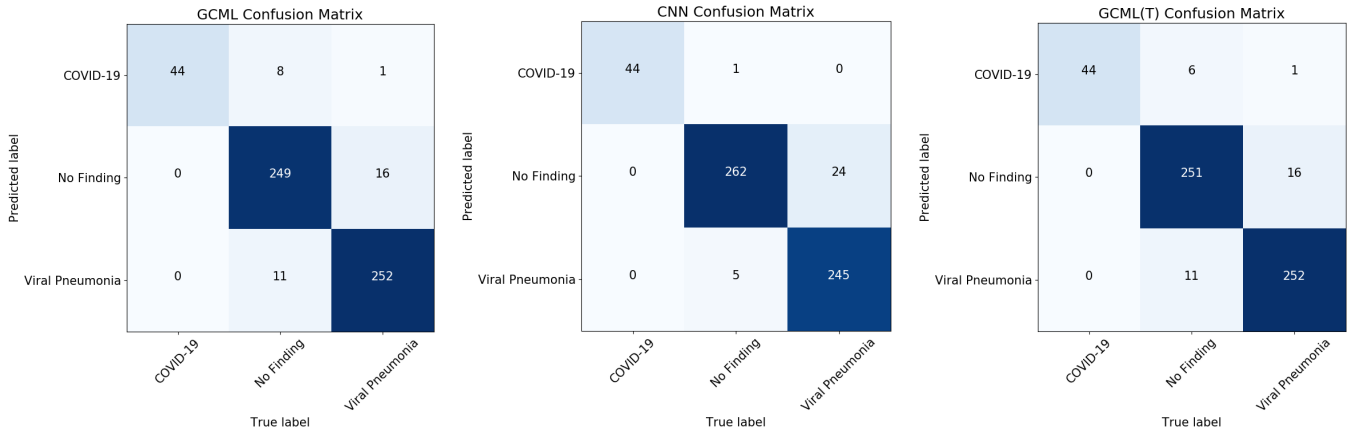
During the training phase of the GCML structure, we explicitly made the normalization step, shown on line 7 of Algorithm 2, optional. This allows us to further train a given GCML structure on new data without having to track distribution statistics of the original training data. For our next experiment we evaluated the effect of further training the GCML structure on additional data points without further tuning the convolutional neural network used as the feature extractor. We used the validation portion of [41] to further train only the GCML structure with the same threshold $\tau = 0.05$. We used 44 samples of COVID-19, 269 samples of No Findings, and 269 samples of Viral Pneumonia images to train for 5 epochs. As we described earlier, forgoing the optinal normalization step in Algorithm 2 allows for this functionality. We then ran the test partition, which neither the convolutional neural network nor the GCML structure have been trained on, to obtain results for combined accuracy, F1 score, and sensitivity. Table IV shows our results using three models, *CNN* or standard classification using our convolutional neural network, *GCML* using our attention mechanism, and finally GCML^T, where we further trained the attention mechanism using the validation partition.

TABLE V: COVID-19 per-class classification accuracy (95% CI) for standard CNN, our GCML approach, and further tuned GCML^T structure using the test partition of the COVID-19 Radiology dataset [41]. We also note that GCML^T structure was further tuned on the validation partition, while our feature extractor used with GCML^T was not further trained with the validation partition.

METHOD	COVID-19	No Finding	Viral Pneumonia
CNN	1.0	0.978 ±0.018	0.911 ±0.034
GCML	1.0	0.929 ±0.031	0.937 ±0.029
GCML^T	1.0	0.937 ±0.029	0.937 ±0.029

The combined results in Table IV show several encouraging results. First, the attention mechanism performed similarly well with a combined accuracy of 0.938 as compared to the traditional classification method of convolutional neural networks with a combined accuracy of 0.948. Second, by further tuning the GCML structure, to obtain GCML^T, with a small number of data points we were able to improve all metrics, bringing combined accuracy to 0.942. This is encouraging because tuning this structure is significantly faster than tuning a convolutional neural network. Finally, once we drilled down to class level accuracy, we observe that GCML improves the standard CNN classification on the Viral Pneumonia class by about 3 percent, as shown in Table V. This was significant for two reasons, first *it empirically verified our main hypothesis that our auxiliary attention mechanism could identify cases of pulmonary conditions that the standard CNN classification missed*. Second, these results support our intuition of the additive nature of standard CNN classification by showing that this method outperformed the attention method on the *No Findings* class, as spatial relations between activated features matter significantly less, as expected.

Due to the absense of object level labels, or bounding boxes labeling localized manifestations of pulmonary conditions in training data, it is difficult to identify feature specific intersections or lack thereof between images that CNN and GCML classified correctly or incorrectly. But even with only image level labels for training, where we know that the condition is present but have no localized information of its manifistation, the confusion matrices in Fig 8 clearly show GCML's improved performance on the *Viral Pneumonia* class, while both methods performed the same on the *COVID-19* class.



(a) (b) (c)

Fig. 8: Confusion matrices for evaluation results using the test partition of the COVID-19 Radiology dataset [41]. The three results represent the following: (a) classification using our GCML attention mechanism, (b) classification using the final connected layer of our fine-tuned Convolutional Neural Network, and (c) classification using our GCML^T attention mechanism further trained using the validation partition of [41]. Our first observation is that our attention mechanism (a) is able to identify pulmonary conditions that the standard CNN (b) missed. This provides empirical evidence to our hypothesis for the utility of attention mechanisms in the chest radiology domain. Second, by further training our GCML structure using the validation partition, which was only used as a model selection criteria during the training of our feature extractor, performance was further improved (c). This was done without further training of the CNN feature extractor. These results suggest that hybrid CNN plus attention-based ensemble techniques, that utilize different inductive biases, provide a promising set of approaches to incorporating global interactions between dispersed features of pulmonary diseases that are manifested in chest radiographs.

TABLE VI: Generalization test dataset for COVID-19 Radiology Data derived from [19]. Neither our feature extractor G, nor the GCML structure S were trained on this data.

COVID-19	No Finding	Pneumonia (Viral and Bacterial)
133	949	390

C. Radiology Generalization Performance

For our final experiment we examined our method's ability to generalize to a separate dataset [19] containing X-ray samples of patients diagnosed with COVID-19, No Findings, and Pneumonia, both viral and bacterial. We created this dataset by extracting images from a publicly available repository [19] in order to create a test partition that was comparable in size to the train partition we used in training our method. The main motivation for this experiment was an observation that many instances of work related to diagnosing COVID-19 and other pulmonary conditions using chest radiographs omit external generalization experiments, as reported in Roberts et al. [1].

Table VI shows the generalization test partition that we created using [19]. Neither our feature extractor nor the GCML structure were trained using this data. We then tested this data using GCML classification, using $\tau = 0.05$. The only transformations that we performed during inference were resizing the image to 224×224, and normalization of the tensorized image to *ImageNet* values for mean and standard deviation.

Table VII shows our results, including per class accuracy, while Fig 9 shows the confusion matrix. Similar to our initial results on [41], we saw our attention mechanism perform well at identifying pulmonary conditions, both COVID-19 and Pneumonia, and slightly worse on the No Findings class. Further examining per-class performance we see that sensitivy rates for both COVID-19 and Pneumonia are high, making the approach effective at detecting infected patients, represented in data never seen by the model during training. These generalization results also increase our confidence in that pertinent features of pulmonary conditions are being discovered by the model, as opposed to spurious artifacts such as X-ray markings related to the origin or medium of radiographs.

TABLE VII: Generalization dataset [19] results at 95% confidence interval. Combined accuracy, F1 score, and sensitivity are shown using our GCML classification approach. Last row shows per-class accuracy for the same classifier.

(95% CI)	GCML		
Accuracy	0.940 ±0.012		
F1 Score	0.969 ±0.009		
Sensitivity	0.955 ±0.011		
	COVID-19	No Finding	Pneumonia
Class Accuracy	0.985 ±0.021	0.9294 ±0.016	0.951 ±0.021

D. Confidence Level on Hypothesis Test

To evaluate our main hypothesis regarding our attention mechanism outperforming the *standard* CNN when classifying pulmonary conditions, we perform a hypothesis test [47] on the two classifers as they pertain to pneumonia classification. Both classifiers performed equally well on COVID-19 classification. Let $p_1 = 0.911$ be CNN accuracy and $p_2 = 0.937$

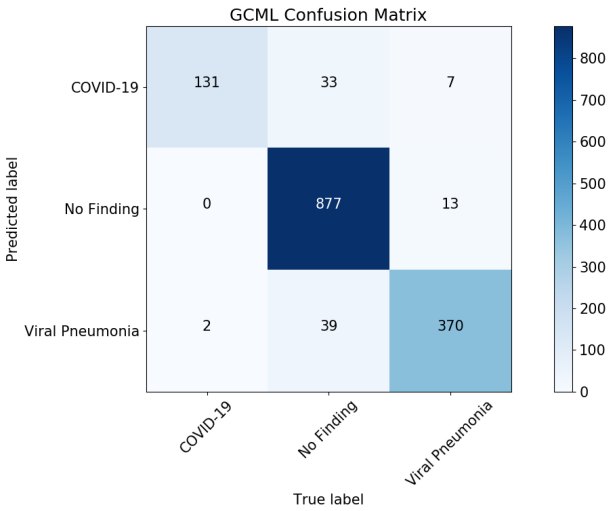


Fig. 9: Confusion matrix for predictions made on the radiology generalization test dataset, created from [19], using the GCML auxiliary attention mechanism. As noted, no training was performed using this dataset of neither the feature extractor nor the GCML attention mechanism.

be GCML accuracy shown in Table V. Let $n = 268$ be the number of samples of viral pneumaonia images in the test partition of [41]. Then, we calculate the test statistic Z as:

$$Z = \frac{p_1 - p_2}{\sqrt{2\hat{p}(1-\hat{p})/n}} \tag{5}$$

where $\hat{p} = (245 + 252)/2(268)$, with 245 and 252 being numbers of correct classifications from both classifiers, as shown in the confusion matrices in Figure 8. To show that $p_1 < p_2$, or that p_2 is better than p_1, we need to show $Z < -z_\alpha$, where z_α is obtained from a standard normal distribution pertaining to a significance level α. Given our sample size, we compute $Z = -1.1607$. Our test statistic is slightly better than the Z value of 1.15035 for a 75% confidence interval. Thus, we can state with 75% confidence level that GCML is more accurate than standard CNN for classifying pneumonia cases.

V. Conclusion

Detection of pulmonary disease using Artificial Intelligence techniques is an emergent field [11], driven not just by academic curiosity, but a real need to improve accessibility and speed of diagnosis. Well studied approaches to image classification need to be analyzed and improved to be effective in real world applications, where data can be scarce, noisy, and novel. Emergence of new diseases, such as COVID-19, exposed weaknesses in such applications of machine learning methods, but also highlighted potential benefits and promise to society if they were to be successfuly addressed [1].

In this work, we developed a method to improve classification performance of pulmonary diseases in chest X-rays by expanding the perceptive awareness of convolutional neural networks via GCML, our new attention mechanism. We showed that our auxiliary attention mechanism improved sensitivity of pulmonary disease classifers by accounting for spatial interrelations of features globally. Our initial results show a promising direction of research towards improving existing methods of image classification for diagnostic assistance of pulmonary diseases.

There are two main limitations of our initial approach. First, the current GCML datastore learns a conditional discrete probability distribution of a given class, more specifically the use of the τ parameter to threshold activation values of the input to attention function Q, results in a binomial distribution. This results in some information loss compared to a continuous distribution. Second, as described in Section III, we must reduce the final mapping resolution of our CNNs to manage the size of learned distributions, which essentially increases the area of individual image patches for which we learn global correlations. This can have a smoothing effect on multiple disease manifestations detected in a single region, causing it to be treated as a single manifistation. To address these limitations, we plan on improving the attention mechanism by utilizing normalized continuous values of class activation maps without a threshold parameter by learning a continuous multivariate distribution, such as a Dirichlet distribution. This will allow for image patches with smaller areas, resulting in more patches, to be used in learning global relations among even smaller manifestations.

Additionally, we can improve positional probability distributions learned by GCML structure by reducing class noise due to class overlapping data in the chest X-ray domain. Similar classes result in noisy class activation maps that contain many overlapping activations [30]. By better extracting class relevant activations, we can learn tighter conditional probability distributions for appropriate classes of diseases.

We believe that improving well established techniques for image classification towards domain-specific applications, such as pulmonary disease classification, is well worth the effort. By utilizing a strong technical foundation of CNNs, progress towards a useful diagnostic aid can be accelerated, and outcomes of patients, especially in regions where access to sophisticated laboratory diagnostics is limited, can be improved. To that end, we make the reference implementation of our attention mechanism freely available, including source code and models [48].

References

[1] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, AIX-COVNET, J. Rudd, E. Sala, and C.-B. Schonlieb, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021.

[2] P. Rajpurkar, J. A. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. Lungren, and A. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *ArXiv*, vol. abs/1711.05225, 2017.

[3] A. Borghesi and R. Maroldi, "Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression," *La radiologia medica*, vol. 125, pp. 509–513, 05 2020.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International*

Conference on Learning Representations, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[5] R. Yamashita, M. Nishio, R. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, 06 2018.

[6] O. Semih Kayhan and J. C. van Gemert, "On translation invariance in cnns: Convolutional layers can exploit absolute spatial location," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 262–14 273.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 01 2012.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[9] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] B. Sahiner, A. Pezeshk, L. M. Hadjiiski, X. Wang, K. Drukker, K. H. Cha, R. M. Summers, and M. L. Giger, "Deep learning in medical imaging and radiation therapy," *Medical Physics*, vol. 46, no. 1, pp. e1–e36, 2019. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13264

[11] O. Albahri, A. Zaidan, A. Albahri, B. Zaidan, K. H. Abdulkareem, Z. Al-qaysi, A. Alamoodi, A. Aleesa, M. Chyad, R. Alesa, L. Kem, M. M. Lakulu, A. Ibrahim, and N. A. Rashid, "Systematic review of artificial intelligence techniques in the detection and classification of covid-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," *Journal of Infection and Public Health*, vol. 13, no. 10, pp. 1381–1396, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187603412030558X

[12] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran, "On the limitation of convolutional neural networks in recognizing negative images," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, pp. 352–358.

[13] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/2cad8fa47bbef282badbb8de5374b894-Paper.pdf

[14] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107731420190921X

[15] M. Rahaman, C. Li, Y. Yao, F. Kulwa, M. Rahman, Q. Wang, S. Qi, F. Kong, X. Zhu, and Z. X, "Identification of covid-19 samples from chest x-ray images using deep learning: A comparison of transfer learning approaches," *Journal of X-ray science and technology*, vol. 28, no. 5, 2020.

[16] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 730–734.

[17] A. Khan, J. Shah, and M. Bhat, "Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images." *Comput Methods Programs Biomed.*, vol. 196, no. 105581, 2020.

[18] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[19] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[20] M. Tamal, M. Alshammari, M. Alabdullah, R. Hourani, H. A. Alola, and T. M. Hegazi, "An integrated framework with machine learning and radiomics for accurate and rapid early diagnosis of covid-19 from chest x-ray," *Expert Systems with Applications*, vol. 180, p. 115152, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421005935

[21] G. Kim, J. Kim, C. Kim, and S.-M. Kim, "Evaluation of deep learning for covid-19 diagnosis: Impact of image dataset organization," *Journal of Applied Clinical Medical Physics*, 06 2021.

[22] M. Heidari, S. Mirniaharikandehei, A. Zargari, G. Danala, Y. Qiu, and B. Zheng, "Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms," *International Journal of Medical Informatics*, vol. 144, p. 104284, 09 2020.

[23] A. Z. Khuzani, M. Heidari, and S. A. Shariati, "Covid-classifier: An automated machine learning model to assist in the diagnosis of covid-19 infection in chest x-ray images," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/05/18/2020.05.09.20096560

[24] M. Alruwaili, A. Shehab, and S. Abd ElGhany, "Covid-19 diagnosis using an enhanced inception-resnetv2 deep learning model in cxr images," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–16, 06 2021.

[25] K. Purohit, A. Kesarwani, D. R. Kisku, and M. Dalui, "Covid-19 detection on chest x-ray and ct scan images using multi-image augmented deep learning model," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/10/19/2020.07.15.205567

[26] N. Tsiknakis, E. Trivizakis, E. Vassalou, Evangelia, Z. Papadakis, Georgios, A. Spandidos, Demetrios, A. Tsatsakis, J. Sánchez-García, R. López-González, H. Karantanas, Apostolos, and K. Marias, "Interpretable artificial intelligence framework for covid-19 screening on chest x-rays," *Experimental and therapeutic medicine*, vol. 20, no. 2, pp. 727–735, 2020.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[29] Z. Wang, Y. Xiao, Y. Li, Z. Jie, F. Lu, M. Hou, and X. Liu, "Automatically discriminating and localizing covid-19 from community-acquired pneumonia on chest x-rays," *Pattern Recognition*, vol. 110, p. 107613, 08 2020.

[30] E. Verenich, A. Velasquez, N. Khan, and F. Hussain, "Improving Explainability of Image Classification in Scenarios with Class Overlap: Application to COVID-19 and Pneumonia," in *Proceedings of the 19th IEEE International Conference on Machine Learning and Applications*, 2020.

[31] A. Gupta, A. Anjum, S. Gupta, and R. Katarya, "Instacovnet-19: A deep learning classification model for the detection of covid-19 patients using chest x-ray," *Applied Soft Computing*, vol. 99, p. 106859, 10 2020.

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[34] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1971–1980.

[35] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu, "Disentangled non-local neural networks," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 191–207.

[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[37] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *NeurIPS*, 2019.

[38] Q. Liu and S. Mukhopadhyay, "Unsupervised learning using pretrained CNN and associative memory bank," *CoRR*, vol. abs/1805.01033, 2018. [Online]. Available: http://arxiv.org/abs/1805.01033

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[40] A. Krishevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[41] T. Rahman, M. Chowdhury, and A. Khandakar, "Covid-19 radiology database," Online, 2020, accessed Jan 2021. [Online]. Available: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

[42] C. Butt, J. Gill, D. Chun, and B. A. Babu, "Deep learning system to screen coronavirus disease 2019 pneumonia," *Applied Intelligence*, p. 1, 2020.
[43] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng *et al.*, "A deep learning algorithm using ct images to screen for corona virus disease (covid-19)," *MedRxiv*, 2020.
[44] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.
[45] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli *et al.*, "Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound," *IEEE Transactions on Medical Imaging*, 2020.
[46] B. Hurt, S. Kligerman, and A. Hsiao, "Deep learning localization of pneumonia: 2019 coronavirus (covid-19) outbreak," *Journal of Thoracic Imaging*, vol. 35, no. 3, pp. W87–W89, 2020.
[47] R. Johnson and J. Freund, *Miller and Freund's Probability and Statistics for Engineers*. Prantice Hall International, 2011.
[48] E. Verenich, "Gcml artifacts repository," Online, 2021, accessed Jul 2021. [Online]. Available: https://gitlab.com/verenich/gcmlpub