Graph Convolutional Networks for Multiple Concept Detection from Medical Radiology Images

Erin Brown
Stanford University
browne@stanford.edu

Yuxi Ke
Stanford University
kyx@stanford.edu

Abstract

Multi-label image recognition, the task of predicting a set of labels representing objects in an image, is an important task in computer vision. We consider a related problem, multi-concept detection, which also involves more abstract and longer labels that does not necessarily correspond to a certain object directly. As such, standard multi-label image recognition models are not generally sufficient for this more complex task. Here, we extend the highly successful graph convolutional network (GCN) based multi-label image recognition framework for the case of multi-concept detection and investigate the efficacy of various mechanisms for concept word sequence representation within this framework. We implement this multi-concept detection scheme for the 2019 ImageCLEF Caption Challenge dataset on medical radiology images. After verifying the GCN model on MS-COCO dataset, we trained on the radiology data and achieved effective image feature extraction with a core ResNet-101 model, but encountered difficulty in learning on the graph network branch of our model. We experimented with different GCN hyperparameters to try to overcome the loss plateau problem.

1. Introduction

The interpretation of medical radiology images requires considerable time from highly trained experts and as such is often a limiting factor in the speed of the clinical diagnosis and treatment process. Automatic methods that can quickly and accurately identify relevant concepts in medical radiology images can speed the interpretation process significantly on their own and can ultimately be built into complete end-to-end image interpretation and summarizing systems that require minimal human expert intervention.

Deep learning architectures with convolutional neural networks (CNNs) has gained wide attention for its high performance in recognizing images. With latest advances in computer vision, biomedical image interpretation and summary have became a less daunting task. CNNs have shown promise in the context of radiology and the potential of helping radiologists achieve diagnostic excellence and to enhance patient diagnosis and treatment. While CNNs have demonstrated a great success in single-label image classification, it is crucial to realize that medical radiology images generally contain multiple labels, which could correspond to 1) many attributes of a particular disease or 2) cues to multiple diseases. Therefore, multi-label image recognition paves the way for enhancements of interpretation and diagnosis of the biomedical images. New methods associated multi-label image recognition enable the predictions of a set of object labels that present in an image. Comparing to multiple-class image classification, the multi-label is a more challenging task due to combinatorial attributes of the feature space. As multiple items co-occur in an image, the label dependencies can be implemented to improve the multi-label task.

For the course project, we are working on the 2019 ImageCLEF Caption Challenge dataset which tackles this problem [1]. The challenge comprises the task of identifying which concepts from a list of concepts are represented in a given image, all of which come from the PubMed Open Access subset of the Radiology Objects in Context (ROCO) dataset.

Given an image, we used a graph convolutional network (GCN) network to output a list of concepts with undefined length, which comes from the Unified Medical Language System (UMLS) set of Concept Unique Identifiers (CUIs). The concepts include not only basic object detection types such as “shoulder joint” and “jaw bone”, but also more refined diagnosis type concepts such as “lymph node hyperplasia” and “st segment depression” as well as concept modifiers including “not at all,” “possible,” and “improved,” adding to the difficulty of the task.

In general, multi-concept detection is a challenging problem that has not been as well studied as multi-label classification, which though quite similar, is sufficiently different to create issue. Nonetheless, the basis for multi-concept detection we can pull from successful methods in multi-label classification, specifically those with features that make them particularly amenable to the broader case of concepts as well as object identification. Multi-label classification methods have come far from the original naive multiple single-label classification technique that neglected the dependencies between labels, and there are some we can choose from that incorporate the dependency between labels into the classification process (e.g. [21]) and even features of the labels themselves (e.g. [4]). Such methods are promising for the task of multi-concept detection as the better understanding of the labels (or concepts) themselves and the interactions between them can allow for better handling of the abstraction associated with concepts over object labels.

2. Related Work

2.1. Previous ImageCLEF Challenges

The results of the 2018 ImageCLEF Caption Challenge are published and include discussions of the methods various groups used [10]. There were two main approaches used on the ImageCLEF concept detection task: multi-modal classification and retrieval [16].

ImageSem approached the problem with a retrieval methodology obtaining 0.0928 in terms of mean F1 scores [23]. Particularly, they employed topic modeling method to select more relevant concept for a given test image, and applied Gensim to modeling topic distribution on retrieved similar images and candidate Concept Unique Identifiers. They retrieved similar images from the training set and clustered concepts of those images [16]. In contrast, multimodal classification was more commonly used [15, 17, 19]. Superior results were achieved by UA.PT Bioinformatics [15]. Besides a traditional bag-of-visual-words algorithm, they experimented with logistic regression and k-Nearest Neighbors (k-NN) for the classification step. Morgan State University implemented a deep learning based approach via both image and text features of the training set for modeling [17]. However, instead of relying on the full 220K-image collection, they used a subset of 4K images on the Keras5 framework to generate deep learning based features. [19] utilized a k-NN based concept detection algorithm on order to automatically predict multiple concepts in medical images. The visual representation of images was based on the bag-of-visual-words and bag-of-colors models.

The top group achieved an F1 score of 0.1108 with a multi-label classification approach using an adversarial autoencoding convolutional neural network (CNN) architecture for unsupervised extraction of visual features [16] with batch normalization, Leaky ReLU activations, regularization loss in the final activations of the encoder, and two dropout layers. The next best group achieved an F1 score of 0.0928 with a CNN-based multi-label classification mechanism using a transfer learning model based on the pre-trained Inception-v3 [22]. These scores provide a baseline against which we can compare our model with.

2.2. Multi-Label Image Classification

One founding work of interest is [21], "CNN-RNN: A Unified Framework for Multi-Label Image Classification," which proposed a framework for learning a joint image-label embedding to characterize the semantic label dependency as well as the image-label relevance. In this work, labels recognized in the same image are predicted in a sequential manner utilizing the joint-embedding space. However, sequential relationship between concepts does not always hold true. The paper focuses on labeling the images with classically defined categories, for instance, ship and airplane, which are much simpler than the medical concepts we encounter in our data set.

Graph Convolutional Networks More recently, GCNs have improved upon the state-of-the-art for multi-label image classification tasks. Introduced in 2017 by Kipf and Welling for the task of semi-supervised classification [11], graph convolutional networks are a powerful neural network architecture for learning on graphs. A GCN is a variant of convolutional neural networks that operates directly on graphs and learns hidden layer representations that encode both graph structure and node features. One approach is to perform convolution on the Fourier transformed graph spectrum, where translation is invariant by operating with the delta function [5]. Here, we follow the conceptual framework put forward by Kipf and Welling [11] to learn a concept-specific descriptor vector, whose dot product with the image feature is calculated for score generation.

Chen et al. [4] adapted the GCN structure proposed in [11] for multi-label image recognition (ML-GCN), achieving state-of-the-art results on the Microsoft Common Objects in Context (COCO) [12] and the 2007 PASCAL Visual Object Classes (VOC 2007) challenges. The ML-GCN architecture considers the class labels as nodes and uses the word embedding of each label as a feature vector for the respective nodes. In doing so, the information implicitly contained in the natural language representation of the labels is incorporated into the model. An adjacency matrix is input to the GCN levels to provide the graph information between the nodes. Constructed from the training data from scratch, it represents the dependent relationship between the concept labels. In the paper, the dependency is illustrated in a conditional and directional manner. The presence of label A indicates a higher likelihood that label B will co-occur, but it does not have to hold true for B to A. The methods proposed in [4] form the foundation of our approach and are discussed further in our Methods section.

More recently, a faster GCN is reported to deal with the time and memory consumption problems brought by the large graphs [3]. They used a Monte Carlo approach for integral estimation of embedding functions. This accelerates GCN training significantly without compromising accuracy notably. While the original GCN learns representations for each label seen in the training set, an inductive network is put forward [7], where a function is able to provide embedding of labels unseen by the model previously. This is achieved by sampling the neighborhood of a node and aggregating feature information from its neighbors.

For our model, we require word sequence embeddings to take the place of single-word embeddings in the node features matrix for the GCN model. There exist both very simple mechanisms such as various pooling mechanisms and more complex mechanisms such as that described in [20], which involves a recurrent neural network approach to embedding generation from single-word embeddings, and other more simple mechanisms as described in [18]. Shen et al. found that simple pooling operations are quite effective at representing longer documents, but recurrent or convolutional compositional functions are more effective when constructing representations for short sentences, and hypothesized that this might be a result of the increased impact of word order on the sentence scale [18].

2.3. Attention Mechanisms

Another work from which we have taken inspiration is “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering” [2]. In their work, they propose a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. For their approach, the bottom-up mechanism (based on Faster R-CNN) proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings. As time allows, we will explore the incorporation of this visual attention mechanism into our proposed model. Results from previous ImageCLEF Caption Challenges indicate that visual attention mechanisms offer significant improvement in performance.

3. Methods

Our proposed method is based primarily on the multi-label image recognition with graph convolutional networks (ML-GCN) architecture proposed by Chen et al. [4]. The overall architecture is outlined in Figure 1.

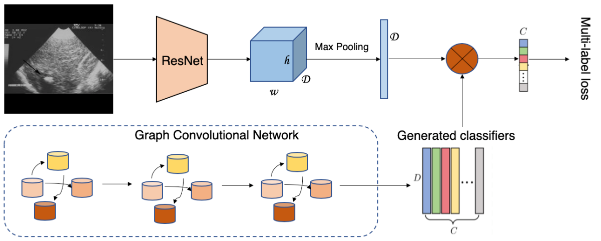


Figure 1: The overall architecture of our proposed method, reproduced with minor modification from [4].

A graph was used to model the inter dependencies between labels, which aims to capture the topological structure in the label space. Particularly, each node (label) of the figure is presented as word embeddings of the label, which are mapped directly into a set of interdependent classifiers with GCN, and then directly applied to an . image feature for classification. Two factors motivated this design of the GCN based model [4]. Firstly, as we are sharing the embedding-to-classifier mapping parameters across all classes, the weak semantic structures are expected to be retained by the learned classifiers in the word embedding space, where semantic related concepts are close to each other. Additionally, the gradients of all classifiers can impact the classifier generation function, which implicitly models the label dependencies. Secondly, a novel label correlation matrix based on their co-occurrence patterns is designed to explicitly model the label dependencies by GCN, with which the update of node features will absorb information from correlated nodes (labels).

3.1. GCN-based Concept Detection

Image Representation Learning The image representation learning module of ML-GCN can be done with any CNN-based model. For ML-GCN, Chen et al. use ResNet-101 [8], which performs well and is easily trainable. For our purposes, we use ResNet-50 as the CNN base. Following the CNN base, a global max-pooling layer is applied to yield image-level feature vector $\mathbf{x}$.

3.2. Concept Detection Learning

3.2.1 Graph Convolutional Network

GCN was used to perform semi-supervised classification. The key idea here is to update the node representations by propagating information between nodes. Let G be a graph with set of C nodes $\mathcal{C}$ and adjacency matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$. We store the d-dimensional node features in matrix $\mathcal{H} \in \mathbb{R}^{C \times d}$ as input to the GCN. Each GCN layer takes as input the node features along with the adjacency matrix and outputs an updated features matrix. In general, the GCN layers have form

$$\mathbf{H}^{l+1} = f(\mathbf{H}^{l}, \mathbf{A}), \tag{1}$$

where f is some nonlinear function. More specifically, per [11],

$$\mathbf{H}^{l+1} = h\left(\hat{\mathbf{A}}\mathbf{H}^{l}\mathbf{W}^{l}\right) \tag{2}$$

where h denotes some activation function. In [4] this is chosen to be LeakyReLU. The features matrix output by the final layer is an inter-dependent concept detector which we denote $\mathbf{W}$. By applying this concept detector to the output $\mathbf{x}$ from the image representation module, we obtain predicted scores $\hat{\mathbf{y}} = \mathbf{W}\mathbf{x}$.

3.2.2 Multi-label Recognition

In principle, we can use any CNN base models to learn the features of an image. In our experiments, we implemented ResNet-101 [9] in the base model in experiments. Therefore, assuming that an input image I is with the 448448 resolution, we can obtain 20481414 feature maps from the conv5 x layer. Then, global max-pooling is used to obtain the image-level feature x:

$$\mathbf{x} = f_{GMP}(f_{cnn}(I; \theta_{cnn})) \in \mathbb{R}^D) \tag{3}$$

where θ_{cnn} indicates model parameters and D = 2048.

3.2.3 GCN based classifier learning

The inter-dependent object classifiers are learned, i.e., $\mathbf{W} = \mathbf{W}_{i=1}^{C}$ from label representations via a GCN based mapping function, where C denotes the number of categories [4]. Stacked GCNs are implemented, where each GCN layer l takes the node representations from the previous layer (H_l) as inputs and outputs new node representations, i.e., H_{l+1}. For the first layer, the input is the $Z \in \mathbb{R}^{C \times d}$ matrix, where d is the dimensionality of the label-level word embedding. For the last layer, the output is $W \in \mathbb{R}^{C \times D}$ with D representing the dimensionality of the image representation. By applying the learned classifiers to image representations, the predicted scores is retrieved as

$$\mathbf{y} = \mathbf{W}x \tag{4}$$

3.2.4 Loss Function

Let $y \in \mathbb{R}^C$ be the ground truth label of an image, where $y_i = 1$ if concept i is represented within the image and $y_i = 0$ otherwise. Let C denote the number of concepts, D the dimensionality of the image representation, and d the dimensionality of the label-level word embedding. We use the following multi-label classification loss function

$$\mathcal{L} = \sum_{c=1}^{C} y^c \log\left(\sigma\left(\hat{y}^c\right)\right) + \left(1 - y^c\right) \log\left(1 - \sigma\left(\hat{y}^c\right)\right) \tag{5}$$

where σ is the standard sigmoid function.

3.2.5 Node Features

The ML-GCN model uses standard GloVe word embeddings [14] for node features, which it may do easily since the labels for each class in the COCO and VOC 2007 datasets are single word entities. For the ImageCLEF 2019 dataset, however, some concepts are given by a single word, e.g. "fibrolipoma," while others are given by many words, e.g. "proximal muscle weakness due to defect at the neuromuscular junction." In [4], one-hot encodings were found to work equally well when used as node features. For simplicity, we use a variation on this and take uniformly random vectors as our node features as described further in Section 4.3.

3.2.6 Adjacency Matrix

We construct an adjacency matrix $\mathbf{A}$ for our graph G from the training data according to the mechanism outlined in [4]. Consider first the co-occurrence rates of the concepts, and define co-occurrence matrix $\mathbf{M} \in \mathbb{R}^{C \times C}$ such that $M_{i,j}$ denotes the number of training set images for which both the ith and jth concepts are tagged. From this we construct the conditional probability matrix $\mathbf{P} \in \mathbb{R}^{C \times C}$, where $P_{i,j}$ is the probability that the jth concept is represented in an image given the occurrence of the ith concept, by normalizing row-wise by the number of times n_i that the ith concept occurs as $\mathbf{P}_i = \mathbf{M}_i / n_i$. Then, to prevent overfitting the correlation matrix to the training set, we consider the binary correlation matrix $\mathbf{B} \in \mathbb{R}^{C \times C}$, where for some threshhold τ,

$$B_{i,j} = \begin{cases} 0, & \text{if } P_{i,j} < \tau \\ 1, & \text{if } P_{i,j} \geq \tau \end{cases} . \tag{6}$$

This binarization reduces noise from rare incidental co-occurrences and strengthens generalizability to unseen datasets. However, it also results in over-smoothing. We can remedy this by re-weighting to obtain our desired adjacency matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$, given by

$$A_{i,j} = \begin{cases} p / \sum_{i \neq j} B_{i,j} & \text{if } i \neq j \\ 1 - p & \text{if } i = j \end{cases} \tag{7}$$

for some parameter $0 \leq p \leq 1$. In practice, we set τ to 0.4 and p to 0.2, which Chen et al. found to be the optimal values on both the MS-COCO and VOC 2007 datasets [4].

3.3. Evaluation

F1 score We used the F1 score to evaluate the concept prediction result. F1 score is the harmonic average of precision and recall, and is given by

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

where

$$\text{precision} = \frac{\sum \text{True Positive}}{\sum \text{True Positive} + \sum \text{False Positive}} \tag{9}$$

and

$$\text{recall} = \frac{\sum \text{True Positive}}{\sum \text{True Positive} + \sum \text{False Negative}} \tag{10}$$

.

The F1 score is the official evaluation metric for the ImageCLEF Caption Challenge.

Mean Average Precision Mean average precision (mAP) is a frequently used metric for object detection and labeling tasks such as COCO and VOC. It is given by the mean of the average precision over all classes:

$$\text{mAP} = \frac{\sum_{i=1}^{n} \text{AP}_i}{n} \tag{11}$$

where AP_i is the average precision of class i and n is the number of classes. We track mAP as training progresses to evaluate model effectiveness.

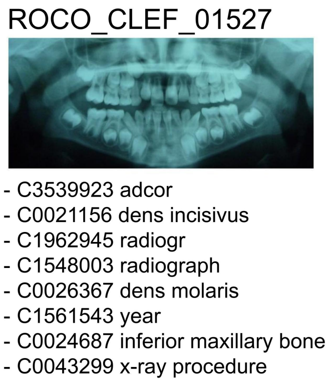
4. Dataset and Features

4.1. Dataset Overview

Our primary dataset of focus is that provided by the 2019 ImageCLEF Caption Challenge. The challenge provides a corpus of 56,629 training and 14,157 validation images as well as a list of 5,528 string concepts that are represented throughout the corpus. We split 10,000 images from the training set to use for testing. For each image, a list of the "golden" associated concepts is provided. For the ImageCLEF challenge, systems are evaluated via F1 scoring of predicted concepts against the ground truth concept lists, with source code for the official F1 scoring tool is provided through the challenge website [1]. We found that the mAP was more sensitive than the F1 score, however, and thus instead used the mAP for development.

4.2. Dataset Pre-processing

The images provided vary significantly in size and aspect ratio. To account for this, we computed statistics over the dataset and determined that almost all images had minimum heighth/width between 150 and 900, so we elected to resize images to dimension $(448, 448)$ to fit ResNet-50. We preserved aspect ratios, scaling the larger axis to size 448 using `PIL.Image.resize()` with antialiasing and padding out the smaller axis to size 448 with black. The figure below shows an example of the information available in the training set.

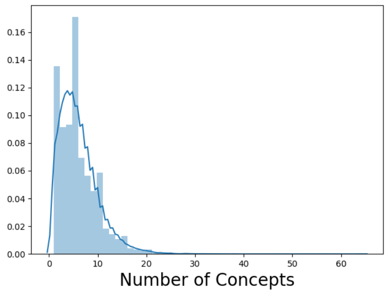


ROCO_CLEF_01527

- C3539923 adcor
- C0021156 dens incisivus
- C1962945 radiogr
- C1548003 radiograph
- C0026367 dens molaris
- C1561543 year
- C0024687 inferior maxillary bone
- C0043299 x-ray procedure

Figure 2: One example of a radiology image with the information provided in the training set.

We firstly analyzed the annotated concept frequency distribution for better understanding the task images. The distribution is important for multi-label training object selection. As seen in the distribution, most of the medical images have fewer than 20 associated concepts.



4.3. Concept Representation

Chen et al. reported only minimal improvement in model performance from semantic-based word embeddings such as GloVe over one-hot vector representations, suggesting that the basis of the efficacy of the proposed ML-GCN model resulted from the overall architecture rather than information implicit in the label representations. Given the length of the concept representations and the increased frequency of rare and out-of-vocabulary words compared to COCO or similar tasks, we expect that sophisticated natural language processing mechanisms outside the scope of this course project would be required to preserve significant semantic information for concept representation in the the ImageCLEF task. As such, we eschewed standard phrase embeddings such as GloVe or Word2Vec with pooling in favor of a random vector representation for each concept. The simple one-hot mechanism used in Chen et al. for the COCO task, which includes 80 label classes, is computationally infeasible for the ImageCLEF task with 5528 concept classes. Instead, we generated a 300-dimensional vector (same dimension as the GloVe vectors used in [4]) with entries sampled uniformly from $(-1, 1)$ for each class. While we cannot achieve the linear independence of the one-hot representation within such reduced dimension, this mechanism of random sampling ensures maximum separation between representations.

	Mean PWD	Mean L
GloVe	8.61 ±0.86	6.51 ±0.66
Random	14.13 ±0.48	10.0 ±0.26

Table 1: Word vector statistics, including mean length and mean pairwise distance under the L2-norm, for GloVe vectors compared against uniformly randomly generated word vectors used for ML-GCN on ImageCLEF.

To ensure viability of the representation within the ML-GCN framework, we verified that the generated vectors had magnitude and pairwise distance similar to that of the GloVe vectors used for the COCO class label representations. These statistics, as included in Table 1, demonstrate that the randomly generated vectors have minimal variability in magnitude and pairwise distance as desired. We further confirmed that the minimum pairwise distance between any two vectors remained close to the mean so as to prevent inadvertent confusion between concept labels in representation. As the vector representations are not tied to the semantic content of the concept labels, any differences in relative correlation between vectors would suggest meaning where there is none. It should be noted that while the actual value of the vector magnitudes is not inherently meaningful, preserving similarity of scale between the GloVe and randomly generated vectors prevents the introduction of scaling factors that could impact the ease of model training.

4.4. Concept Adjacency

The other core preprocessing step was the assembly and analysis of the co-occurrence matrix $\mathbf{M}$ described in Section 3.2.6. Once computed for ImageCLEF, where $\mathbf{M}_{\text{ImgCLEF}}$ has dimension 5528×5528, we examined its sparsity structure in comparison with that for COCO, where the $\mathbf{M}_{\text{COCO}}$ is of dimension 80×80. The following figure shows the full sparsity structure of $\mathbf{M}_{\text{ImgCLEF}}$, while the figure after shows a closer look at the sparsity structure of $\mathbf{M}_{\text{ImgCLEF}}$ on the same scale as $\mathbf{M}_{\text{COCO}}$ for comparison.

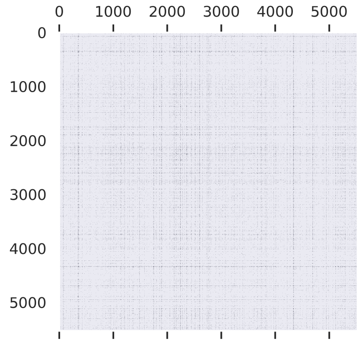


Figure 3: Sparsity pattern of the adjacency matrix for the ImageCLEF training dataset.

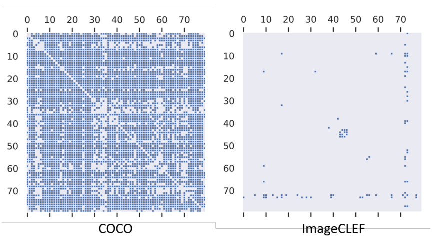
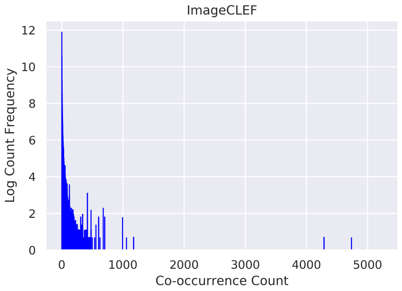
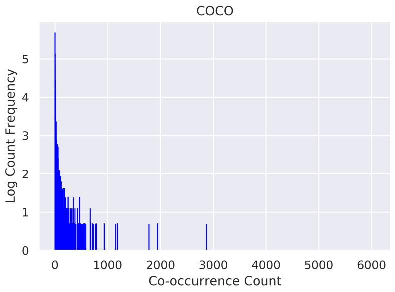


Figure 4: Sparsity pattern of the adjacency matrix for the COCO training dataset (left), and representative subset of the sparsity pattern of the adjacency matrix for the ImageCLEF training dataset (right).

It is clear from these figures that $\mathbf{M}_{\text{ImgCLEF}}$ is signifcantly more sparse than $\mathbf{M}_{\text{COCO}}$.

We also check the profile of the values contained within the co-occurrence matrix for both $\mathbf{M}_{\text{ImgCLEF}}$ and $\mathbf{M}_{\text{COCO}}$ as shown in Figures 5 and 6.





5. Experiments

5.1. ResNet Baseline

We firstly implemented a 50-layer ResNet architecture as our primary baseline since it forms the foundation of our proposed GCN model. We pre-trained the multi-label image classification model with a label matrix of binary flags where rows indicate images and a total of 5528 columns present all the labels. On each row of the label matrix, positions of 1 indicate the labels for that particular image. Random flipping, cropping and distortions were used for data augmentation during the training.

Secondly, we also implemented a large NASNet architecture for the multi-label image classification. The cross entropy loss of this model is overall lower than that of Resnet-50. The loss for these models can be seen in Figures 8 and 9.

5.2. Our Implementation of GCN

Our implementation of GCN is based off of that described in [4] and uses their published Github repository as a source for starter code. This code was primarily preserved as is for evaluation over MS-COCO, and then modified and used in conjunction with new scripts for our ImageCLEF implementation.

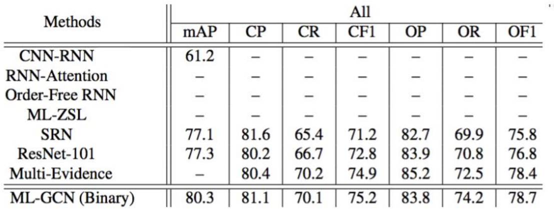
With MS-COCO We verified the results reported in [4] on the MS-COCO dataset for multi-label image recognition. MS-COCO has been the benchmark for multu-label image recognition, which contains 82081 images as the training set and 40504 images as the validation set. The number of categories were only 80 for MS-COCO with approximately 3 objects per image. As in [4], the method performance was evaluated on the validation set as the ground-truth labels for the test set are not available.

With Radiology Images Our ML-GCN consists of two GCN layers with dimensions of 1024 and 2048, respectively. For label representations, we adopt 300-dim uniformly and randomly generated vectors comparable to GloVe [14] in statistics. For the correlation matrix, unless otherwise stated, we set the matrix smoothing threshold τ in Eq. (6) to be 0.4 and the re-weighting parameter p in Eq. (7) to be 0.2, as reported in Chen et. al. [4]. In the branch where the image representation is learned, we choose LeakyReLU [13] with a negative slope of 0.2 to be the non-linear activation function. The negative slope leads to faster convergence in the experiments. For the feature extraction backbone, we use ResNet-50 [8] pre-trained on ImageNet [6]. During training, the input images are random cropped and resized into 448 $\times$ 448 with random horizontal flips for data augmentation. For network optimization, we use SGD as the optimizer. The momentum is set to be 0.9 and weight decay to 104. Except where stated otherwise, the initial learning rate is 0.1, and decays by a factor of 10 every 40 epochs. We implement the model based on PyTorch.

6. Results

6.1. ML-GCN with MS-COCO

As described above, we replicated the implementation of ML-GCN for MS-COCO described in [4], which demonstrated it as a great model for multi-label image recognition. Our results matched those reported in [4] precisely, which we have included for ease of reference in Table 2, reproduced from [4]. This table contains the quantitative results comparing the ML-GCN with state-of-the-art methods, such as CNN-RNN, RNN-attention. Note that we only implemented ML-GCN, not any of the models against which it is being compard.



Methods	All						
	mAP	CP	CR	CF1	OP	OR	OF1
CNN-RNN	61.2	–	–	–	–	–	–
RNN-Attention	–	–	–	–	–	–	–
Order-Free RNN	–	–	–	–	–	–	–
ML-ZSL	–	–	–	–	–	–	–
SRN	77.1	81.6	65.4	71.2	82.7	69.9	75.8
ResNet-101	77.3	80.2	66.7	72.8	83.9	70.8	76.8
Multi-Evidence	–	80.4	70.2	74.9	85.2	72.5	78.4
ML-GCN (Binary)	80.3	81.1	70.1	75.2	83.8	74.2	78.7

6.2. ML-GCN with Radiology Images

Our ML-GCN model for the 2019 ImageCLEF Caption Challenge failed to yield successful learning despite tuning over a range of learning rates and additional hyperparameters. It should be noted that in addition to implementing this system for MS-COCO, closely replicating the results of [4] without issue, we have further implemented checks throughout to ensure that our implementation indeed works as it should.

As none of the learning rates tested yielded successful learning, we then fixed our learning rate at `lr = 0.1`, which [4] found to be optimal for their implementation on COCO and VOC, while testing different correlation matrix thresholding τ values.

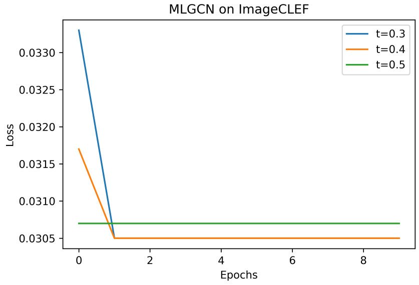


Figure 5: For each value of threshold τ, the loss quickly plateaus.

The threshold value $\tau = 0.4$ achieved the lowest loss, though the difference between those depicted is so small as to be negligible.

We observed a best mAP of 0.185, which occurred in the first epoch with learning rate 0.1 and a threshold τ of 0.4.

While our ML-GCN model did not train successfully in ImageCLEF, our implementations of ResNet-50 and NASNet exhibited the appropriate loss behavior, thus suggesting that the problem with our ML-GCN architecture lies in the graph convolutional aspect of it rather than with the ResNet-101 model at its core. The loss for these models is shown in the figures below.

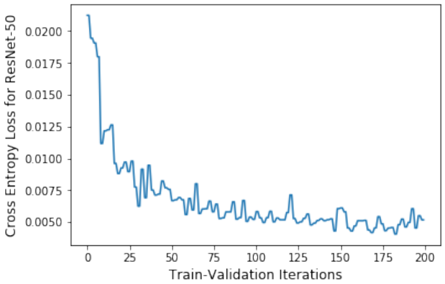


Figure 6: The ResNet-50 cross entropy loss on ImageCLEF decreases over the train-validation steps with small fluctuations.

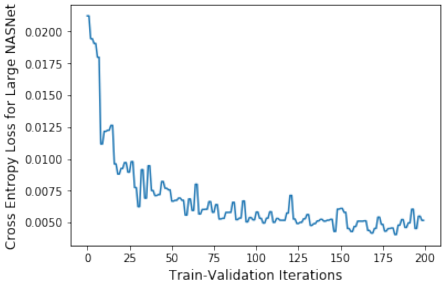


Figure 7: The NAS-Net cross entropy loss on ImageCLEF decreases over the train-validation steps with small fluctuations.

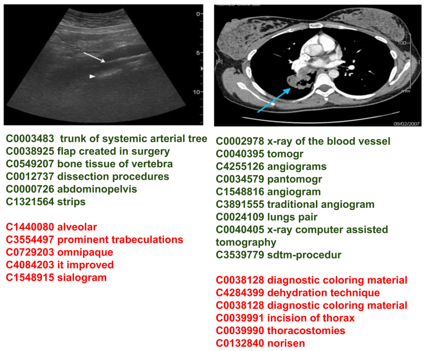

C0003483 trunk of systemic arterial tree
C0038925 flap created in surgery
C0549207 bone tissue of vertebra
C0012737 dissection procedures
C0000726 abdominopelvis
C1321564 strips
C1440080 alveolar
C3554497 prominent trabeculations
C0729203 omnipaque
C4084203 it improved
C1548915 sialogram
C0002978 x-ray of the blood vessel
C0040395 tomogr
C4255126 angiograms
C0034579 pantomogr
C1548816 angiogram
C3891555 traditional angiogram
C0024109 lungs pair
C0040405 x-ray computer assisted tomography
C3539779 sdtm-procedur
C0038128 diagnostic coloring material
C4284399 dehydration technique
C0038128 diagnostic coloring material
C0039991 incision of thorax
C0039990 thoracostomies
C0132840 norisen

Figure 8: True and False Predictions

A few examples with true and false label predictions are illustrated as above. Green indicates correct concept labels while red points to incorrect predictions. As we have noticed that in some of the true labels might be incomplete or misleading, which could cause undesirable training accuracy compared to the MS-COCO dataset. In addition, the true labels are rare medical terms which can be ”out-of-vocabulary” and lead to false cues to the GCN models beyond the image feature extractions.

7. Conclusion

The ML-GCN framework is challenging to train for the ImageCLEF task, and further work is required to determine appropriate hyperparameters for learning. The extreme sparsity of the concept adjacency matrix and the scale disparity seen in the concept co-occurrence frequency profile combine to induce ill-conditioning within the system via the correlation matrix, leading to catastrophic zeroing of the updates and preventing effective learning. Our ability to cross-validate across first different hyperparameters (e.g. learning rate) and then different scaling mechanisms for the correlation matrix was greatly limited by the computational demands of the model in the face of our restricted time and computational resources, and as such we are still in the process of finding a mechanism that effectively preserves the information of the correlation matrix for the graph convolutional network while avoiding the introduction of undesirable numerical effects.

We hypothesize that the binary threshold step performed on the adjacency matrix may over smooth the already-sparse edges between concepts. Other matrix smoothing methods that are less dramatic than the binary operation right now should be tried, so as to maximally preserve sparse net information while discarding noise from extremely rare co-occur events and maintaining numerical stability within the model.

7.1. Future Work

First and foremost, more work is required to alter the construction of the correlation matrix in such a way that our model can learn effectively for the 2019 ImageCLEF Caption Challenge task with its associated difficulties. For example, inductive methods [7] could be used to provide features for concepts barely seen or absent in the training set. It would be useful to visualize the learned concept features. Not only will this be indistinctly understandable, but also help us find out the problems happening during training. After that, we would like to explore different mechanisms for concept representation via word phrase embedding to make use of the structure implicit in the labels themselves. This is an especially important area to improve upon given the complexity and abstraction of the notion of concept labeling compared to its more common counterpart, concept detection.

Furthermore we would like to use attention mechanisms [2] to relate concepts to image regions. It would also be useful for visualizing the concepts learned.

References

[1] Imageclef 2019 caption. `https://www.imageclef.org/2019/medical/caption`.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. pages 6077–6086, 06 2018.

[3] J. Chen, T. Ma, and C. Xiao. Fastgcn: Fast learning with graph convolutional networks via importance sampling. *CoRR*, abs/1801.10247, 2018.

[4] Z. Chen, X. Wei, P. Wang, and Y. Guo. Multi-label image recognition with graph convolutional networks. *CoRR*, abs/1904.03582, 2019.

[5] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *CoRR*, abs/1606.09375, 2016.

[6] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[7] W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. *CoRR*, abs/1706.02216, 2017.

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

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] B. Ionescu, H. Müller, M. Villegas, A. G. S. de Herrera, C. Eickhoff, V. Andrearczyk, Y. D. Cid, V. Liauchuk, V. Kovalev, S. A. Hasan, Y. Ling, O. Farri, J. Liu, M. Lungren, D.-T. Dang-Nguyen, L. Piras, M. Riegler, L. Zhou, M. Lux, and C. Gurrin. Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Avignon, France, September 10-14 2018. LNCS Lecture Notes in Computer Science, Springer.

[11] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[12] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[13] A. L. Maas. Rectifier nonlinearities improve neural network acoustic models. 2013.

[14] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[15] E. Pinho and C. Costa. Feature learning with adversarial networks for concept detection in medical images: Ua. pt bioinformatics at imageclef 2018. In *CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France*, pages 10–14, 2018.

[16] E. Pinho and C. Costa. Feature learning with adversarial networks for concept detection in medical images: Ua.pt bioinformatics at imageclef 2018. 09 2018.

[17] M. M. Rahman. A cross modal deep learning based approach for caption prediction and concept detection by cs morgan state. In *CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France*, pages 10–14, 2018.

[18] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *ACL*, 2018.

[19] L. Valavanis and T. Kalamboukis. Ipl at imageclef 2018: A knn-based concept detection approach. In *CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France*, pages 10–14, 2018.

[20] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, and L. Carin. Joint embedding of words and labels for text classification. In *ACL*, 2018.

[21] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[22] X. Wang, Y. L. Zhang, Z. Guo, and J. Li. Imagesem at imageclef 2018 caption task: Image retrieval and transfer learning. In *CLEF*, 2018.

[23] Y. Zhang, X. Wang, Z. Guo, and J. Li. Imagesem at imageclef 2018 caption task: Image retrieval and transfer learning. CLEF, 2018.

8. Contributions and Acknowledgments

Erin Brown Implemented models and obtained most of the results.

Yuxi Ke Proposed GCN as the model for our image concept recognition task. Report writing. Poster making. Discussed the model along implementations.

Acknowledgments We thank also **Lu Yang**, who is not enrolled in CS231n but contributed in discussion. Lu Yang is a Bioengineering PhD student at Stanford.