A RELATIONAL-LEARNING PERSPECTIVE TO MULTI-LABEL CHEST X-RAY CLASSIFICATION

Anjany Sekuboyina*1,2  Daniel Oñoro-Rubio1  Jens Kleesiek3,4  Brandon Malone1

1 NEC Laboratories Europe, Germany  
2 Department of Informatics, Technical University of Munich, Germany  
3 Institute for AI in Medicine (IKIM), University Hospital Essen, Germany  
4 University of Duisburg-Essen, Germany

ABSTRACT

Multi-label classification of chest X-ray images is frequently performed using discriminative approaches, i.e. learning to map an image directly to its binary labels. Such approaches make it challenging to incorporate auxiliary information such as annotation uncertainty or a dependency among the labels. Building towards this, we propose a novel knowledge graph reformulation of multi-label classification, which not only readily increases predictive performance of an encoder but also serves as a general framework for introducing new domain knowledge.

Specifically, we construct a multi-modal knowledge graph out of the chest X-ray images and its labels and pose multi-label classification as a link prediction problem. Incorporating auxiliary information can then simply be achieved by adding additional nodes and relations among them. When tested on a publicly-available radiograph dataset (CheXpert), our relational-reformulation using a naive knowledge graph outperforms the state-of-art by achieving an area-under-ROC curve of 83.5%, an improvement of ~1% over a purely discriminative approach.

Index Terms— knowledge graph, relational learning, radiographs, chest, multi-label classification

1. INTRODUCTION

Chest radiography is one of the most common imaging modalities for diagnosing pathologies in lungs and heart, with multitudes of scans performed annually. Resulting high number of images require efficient screening. However, shortage of radiological experts, long working hours, manifestation of abnormalities with similar visual cues, etc. make accurate and efficient diagnosis difficult. Thus, there is a need for a computer-aided and explainable decision-support system in the clinical routine that can assist medical experts in reliably identifying pathological chest X-rays (CXR). This task is termed as multi-label CXR classification (multiple co-occurring findings can be drawn from one radiograph).

Thanks to large-scale CXR datasets [1, 2], numerous deep learning-based approaches have been proposed to tackle the problem multi-label CXR classification. Two major factions of solutions proposed in the literature include: (1) ad-hoc binary multi-label classification [3, 4] and (2) exploiting label-dependency-based classification [5, 6, 7]. Approaches from the latter faction are interesting as they aim to exploit domain knowledge in addition to an image-to-label mapping. Yao et al. [7] propose a recurrent neural network-based decoder to learn the inter-dependency among labels. Pham et al. [6] construct a label tree and propose a hierarchical training regime of learning to classify the leaf nodes followed by the non-leaf ones, which is equivalent to conditional training. Chen et al. [5], on the other hand, propose to exploit label co-occurrence. A graph based on the co-occurrence of the labels is used to enrich the feature representations learnt by a naive convolutional neural network (CNN) working on images.

Motivated by the idea of elegantly incorporating auxiliary information (e.g. domain knowledge) into the problem of

Fig. 1: An overview of the Radiological Knowledge Graph constructed with the chest X-rays and ‘findings’ as entities.
multi-label CXR classification, we propose to reformulate the problem using a relational-learning perspective. Primarily, such a formulation involves a knowledge graph (KG) composed of the CXRs and the diseases (or findings) as the nodes (or entities) and the associations between them denoted by directed edges (or relations). We argue that this formulation is more flexible and show that it results in a better discriminator. New information can be added either as a new node or as a new relation.

1.1. Background: Link prediction in knowledge graphs

In this section, we concisely introduce a knowledge graph (KG) and the link-prediction problem in the KG-realm. A knowledge graph $\mathcal{G}$ can formally be represented by a set of triplets $\{(s, r, o)\}$, where each triple denotes a ‘fact’ in the real world. A triple captures the relation, $r$, between a subject, $s$ and an object, $o$. The subject and object belong to a set of entities, $s, o \in \mathcal{E}$, and the relation belongs to a set of relations, $r \in \mathcal{R}$. For example, consider a sub-world with the following: $\mathcal{E} = \{\text{Kyoto, Japan, Murakami}\}$ and $\mathcal{R} = \{\text{bornIn, locatedIn}\}$. A KG corresponding to this worlds could have triples such as (Murakami, bornIn, Kyoto), (Murakami, bornIn, Japan), (Kyoto, locatedIn, Japan), etc.

Notice that such a KG can be incomplete due to reasons such as missing facts, corrupted data, etc., and it is of interest to complete the graph. Link prediction is one of the methods that address graph completion. It can be formalised as a triplet-ranking problem. It involves learning a scoring function, $\psi : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$, such that for a given triplet $x$, $\psi(x)$ takes a high value if $x$ is a true fact and a low value otherwise.

1.2. Contribution

We propose to look at multi-label CXR classification as a link prediction problem in a radiological knowledge graph (RadKG). Example entity- and relation-sets for such a KG would be $\mathcal{E} = \{\text{CXRs, diseases, other findings, patient reports, ...}\}$ and $\mathcal{R} = \{\text{hasDisease, hasFinding, childOf, ...}\}$. This formulation enables a seamless combination of multiple modalities of data as well as external domain knowledge such as disease ontologies [8] and patient graphs [9].

Specifically, we combine the domains of relational machine learning with multi-label classification to make three key contributions:

1. We reformulate multi-label CXR classification as a link-prediction problem in a radiological knowledge graph.
2. We show that our reformulation yields an area-under-ROC curve of 83.5% compared to 83.2% from prior state of the art [5], when tested on CheXpert [1]. This is without incorporating any label dependency, unlike [5].
3. We demonstrate the flexibility of our KG representation by including additional information: (a) uncertain labels in the ground truth annotations and (b) the dependency among CXR findings.

2. METHOD

We present our approach in two parts: First, we describe the construction of a simple radiological KG; second, we solve CXR multi-label classification using link-prediction in this KG.

2.1. Radiological Knowledge Graph (RadKG)

The entity set $\mathcal{E}$ in RadKG is composed of the CXRs and the findings, $\mathcal{E} = \mathcal{E}_X + \mathcal{E}_F$. The CXR entity-subset consists of $m$ CXRs denoted by $X$, $\mathcal{E}_X = \{X_i\}_{i=1}^m$. The findings-subset is denoted by $\mathcal{E}_F = \{F_j\}_{j=1}^n$. In CheXpert [1], $F_j$ corresponds to findings such as ‘Pneumonia’, ‘Edema’, ‘Fracture’ etc., with $n = 14$. The image-to-label annotations in multi-label classification are denoted by relations between $\mathcal{E}_X$ and $\mathcal{E}_F$, $\mathcal{R} = \{\text{hasFinding, probablyHasFinding}\}$. Fig. 1 illustrates RadKG. Specifically, hasFinding links a CXR and a finding when it is positively annotated, while probablyHasFinding captures the uncertain annotations\(^1\). Observe that lack of an edge between $X_i$ and $F_j$ implicitly implies a negative annotation.

2.2. Link prediction on RadKG

Formulating the problem. A link-prediction model primarily operates on ascertaining the validity of triples. If a triple, $t = (s, r, o)$, exists in the KG, the model is supposed to predict a TRUE value and a FALSE otherwise. Fig. 2 illustrates this model for an example triple from the RadKG. The proposed link-prediction model consists of three modules: An encoding, an embedding, and a scoring module. The encoding and embedding modules are responsible for mapping the entities in $t$ to an embedding space with representations

\(^1\)CheXpert contains uncertain annotations indicating findings where the NLP-based automatic annotator was not confident about its prediction.
denoted by \( e_s, e_o \in \mathbb{R}^d \) (\( d = 100 \), in this work). The scoring function, \( \psi(s, r, o) = \psi_r(e_s, e_o) \in \mathbb{R} \), then scores \( t \) at a high value if it is a positive triple and vice versa. In the following sections, we detail the specifics of the three modules and describe the training and inference procedures.

Encoding the entities. In a typical KG, the entities are abstract concepts. The encoding module is responsible for representing these concepts as a mathematical entity. In RadKG, this representation concerns \( \mathcal{E}_f \): a finding, \( F_j \), is encoded into a one-hot vector, \( \mathbb{I}_j \in \mathbb{R}^{|\mathcal{E}_f|} \), active at the \( j \)th index. In this work, \(|\mathcal{E}_f| = 14\) findings from CheXpert’s label set.

Additionally, RadKG also consists of CXRs (\( \mathcal{E}_X \)) which are visually-rich entities. Instead of one-hot representations as described above, we choose to encode \( X \)s using a convolutional neural network. Such domain-specific encoding not only avoids the ad-hoc, orthogonal one-hot codes but enables efficient capture of the visual information in CXRs and results in similar CXRs having similar encodings. For this purpose, we extract the 1024-length feature representations from a 121-layered DenseNet121 (DenseNet121) pre-trained for multi-label CXR classification.

Embedding the entities. The encoding module results in 1024-length codes for CXRs and \(|\mathcal{E}_f|\)-length codes for the findings. Note that these codes are static, especially since the DenseNet121 is pre-trained (offline) and the one-hot encoding for the findings has no learning component. Therefore, we incorporate an embedding module which maps these codes to embedding spaces more suited for the link-prediction task. Denoting the encoded representations with \( c_X \) and \( c_F \), the embedding module consists of two functions: \( g_X(c_X) : \mathbb{R}^{1024} \rightarrow \mathbb{R}^d \) and \( g_F(c_F) : \mathbb{R}^{|\mathcal{E}_f|} \rightarrow \mathbb{R}^d \). We denote the embedded representation by \( e \).

Scoring the triples. Once the subject and object entities are embedded into the embedding space, recall that the triple \( t \) needs to be scored for its validity. We investigate two scoring functions in this work, DistMult [10] and ConvE [11], formally denoted as follows:

\[
\begin{align*}
\text{DistMult: } &\quad \psi_r(e_s, e_o) = \langle e_s, r_r, e_o \rangle, \\
\text{ConvE: } &\quad \psi_r(e_s, e_o) = f(\text{vec}(f([e_s; r_r] + w))W)e_o,
\end{align*}
\]

where \( r_r \in \mathbb{R}^d \) denotes the embedding for the relation \( r \) and \( \langle \rangle \) denotes an inner product of the three vectors. The ConvE score function is relatively ‘convoluted’, involving the reshaped and concatenated entity and relation embeddings, \( \overline{e_s}, \overline{r} \in \mathbb{R}^{k_w \times k_h} \) (s.t. \( k_w \times k_h = d \); \( k_h = k_w = 10 \)). The concatenated embeddings are passed through non-linear functions \( f = \text{ReLU} \), convolved (kernel \( w \) of size 5) and finally vectorised and passed through a fully connected-layer with weight \( W \). In order to avoid clutter, we refer the reader to [10] and [11] for the minutiae of the two scoring functions.

### Learning to link

In the proposed model, the entity and relation embeddings, along with the parameters of the scoring function parameters, are to be learnt. They can be trained with supervision once the scoring function is defined. This entails sampling valid and invalid samples from the training data and learning to assign high and low scores to them respectively. For this, we apply a sigmoid function on the raw scores, \( p = \sigma(\psi_r(e_s, e_o)) \), and minimise a binary cross-entropy loss, as formulated below:

\[
\mathcal{L}(p, y) = -y \log(p) - (1 - y) \log(1 - p),
\]

where the target label, \( y \), is 1 if a triple is valid and 0 otherwise.

Once trained, we infer the labels of an unseen CXR, \( X_{test} \), by scoring every completion from RadKG of the form \((F_{test}, \text{hasFinding}, F)\). The query-triple is scored for every \( F_j \in \mathcal{E}_f \) as the object and positively annotated for those findings for which \( \sigma(\psi_r(\cdot)) > \tau \), where \( r \) is \text{hasFinding} and \( \tau \) is chosen to appropriately tradeoff sensitivity and specificity.

### Data

CheXpert [1] is a publicly-available chest radiograph dataset consisting of 224,315 lateral and frontal chest radiographs annotated for 14 findings with their presence, absence, or uncertain presence. The data is split into three-folds of train, validation, and a test sets with a 70 : 10 : 20 proportion stratified at patient-level. We work with the frontal CXRs padded to squares, resized to \( 320 \times 320 \) pixels, and z-score normalised using the train-set’s pixel mean and variance. Note that the test and validation sets do not contain CXRs with uncertain labels.

### 3. RESULTS

The results of our experiments are tabulated in Table 1. We use the area under ROC curve (AUC) as a metric to evaluate classification performance. All our experiments are evaluated over the three random data folds as described above and their mean is reported. We do not report standard deviation as it was less than 0.1% in all cases. Similar to prior work, we perform two sets of experiments: first, we regard the uncertain labels as positives and second, as negatives. We compare our approach to two prior works [1] and [5], the former using a typical CNN and the latter using a CNN whose feature are augmented with disease features learnt from a disease co-occurrence graph using a graph convolution network.

### Baselines

Recall that CXRs are encoded using a DenseNet121 trained for multi-label CXR classification. This naturally forms one of our baselines. Observe that the performance of our naive implementation is on par with the that of [5]. Once the DenseNet is trained, the 1024-length features are stored as CXR codes and used for the consequent experiments. We

---

2Official CheXpert test set is not publicly available
We incorporate new domain knowledge into RadKG by adding two relations (cf. Table 2):

\[ \text{probablyHasFinding} \] and \[ \text{coOccur} \]. Since the margins are insignificant, we report the standard deviations for a clearer picture.

| RadKG+ConvE | Uncertain \( \rightarrow \) +ve | Uncertain \( \rightarrow \) -ve |
|-------------|-------------------------------|-------------------------------|
| + probablyHasFinding | 0.8339\( \pm \)0.0004 | 0.8330\( \pm \)0.0013 |
| + coOccur | 0.8343\( \pm \)0.0000 | 0.8332\( \pm \)0.0009 |

Incorporating two additional relations, \[ \text{probablyHasDisease} \] and \[ \text{coOccurWith} \], did not yield an improvement. Non-success of \[ \text{coOccurWith} \] can be attributed to the relatively fewer finding-to-finding triples compared to an image-to-finding triples, i.e. \( F_i \) has a far higher degree of edges linked to CXRs than to other \( F_j \)'s. However, it is surprising that the effect of \[ \text{probablyHasDisease} \] is minor. We assume this is due to the uncertain annotations in CheXpert being noise rather than useful information. Observe, in Table 3 of [1], that the experiment considering the uncertain labels as a separate class significantly outperforms others in only one out of the five diseases evaluated on. However, we expect that adding more reliable information, e.g. medical knowledge, will boost the performance further.

Future work. Proposed formulation of classification as a KG-completion problem opens up possibilities to fuse multi-modal data such as text reports and patient-population graphs. However, a encoding component pre-trained towards a certain task (multi-label classification, in our case) could result in representations that do not generalise well across tasks. Therefore, it is of interest to adopt a task-agnostic representation learning framework. Combining the encoding and embedding modules resulting in a fully-end-to-end formulation is also a future research direction.

5. CONCLUSION

In this work, we present a relation-learning based reformulation of the multi-label CXR classification. A knowledge graph constructed with CXRs and ‘findings’ as entities forms the core of our approach. We then pose the classification as a link-prediction problem in this KG. We demonstrate a
superior performance of the proposed approach on publicly available CXR dataset, achieving an AUC of 83.5%, outperforming state-of-art methods.

6. ACKNOWLEDGEMENTS.

The authors do not wish to declare any conflicts of interest.

7. COMPLIANCE WITH ETHICAL STANDARDS.

This study was conducted retrospectively using human subject data made available in open-access as CheXpert [1]. Explicit ethical approval from our end was not instructed by the data source.

8. REFERENCES

[1] Jeremy Irvin et al., “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in Proc. AAAI, 2019.

[2] Aurelia Bustos et al., “PadChest: A large chest x-ray image dataset with multi-label annotated reports,” Med Image Anal, 2020.

[3] Pranav Rajpurkar et al., “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,” arXiv:1711.05225, 2017.

[4] Bingzhi Chen et al., “DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest X-rays,” Biomed Signal Process Control, 2019.

[5] Bingzhi Chen et al., “Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification,” IEEE Biomed Health Inform, 2020.

[6] Hieu H Pham et al., “Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels,” 2019.

[7] Li Yao et al., “Learning to diagnose from scratch by exploiting dependencies among labels,” arXiv:1710.10501, 2017.

[8] Olivier Bodenreider, “The unified medical language system (UMLS): integrating biomedical terminology,” Nucleic Acids Research, 2004.

[9] Carol Friedman et al., “A generalized relational schema for an integrated clinical patient database,” in Proc. Annu. Symp. Computer Appl. in Medical Care, 1990.

[10] Bishan Yang et al., “Embedding entities and relations for learning and inference in knowledge bases,” in Proc. ICLR, 2015.

[11] Tim Dettmers et al., “Convolutional 2d knowledge graph embeddings,” in Proc. AAAI, 2018.