9.6. [Learning notes] TieNet: Text-Image Embedding Network for common Thorax Disease Classification and Reporting in Chest X-rays

一. Aim: Extract the image & text representations

① multiple label disease prediction (= disease classification)

② use only image as input, ∴ output disease classification and a preliminary report together.

二. Method

1. End-to-End Trainable CNN-RNN Model

Input1: Text Report $\rightarrow \vec{w}_t$

①

$$S = \{\vec{w}_1, \cdots \vec{w}_T\},\ \vec{w}_t \in \mathbb{R}^v \quad \boxed{d_w \times T}$$

$\vec{w}_t$: vector standing for a d_w dimensional word embedding for the t-th word in report. $\boxed{d_w \times 1}$

V: size of vocabulary

T: length of the report.

Input2: Image $\rightarrow \vec{X}$

② LSTM:

$$H = (\vec{h}_1, \cdots, \vec{h}_T) \quad \boxed{d_h \times T}$$

$\vec{h}_t$: hidden layer in LSTM $\boxed{d_h \times 1}$

d_h: LSTM state space of dimension d_h

$\vec{X}$: the convolutional activations from the transition layer. $\boxed{D \times D \times C}$

$\vec{a}_t$: subsequent attention map.

$$\vec{h}_t = LSTM([w_t, a_t * X], h_{t-1}) \quad \boxed{d_h \times 1}$$

2. Attention Encoded Text Embedding

① weights $G = softmax(\vec{W}_{s2} \tanh(\vec{W}_{s1} \vec{H}))$ $\boxed{r \times T}$

r: number of global attentions we want to extract from the sentence.

$\vec{W}_{s1}$: $\boxed{s\text{-by-}d_h}$ matrix

$\vec{W}_{s2}$: $\boxed{r\text{-by-}s}$ matrix

s: a hyperparameter governing the dimensionality.

② embedding matrix $\vec{M} = \vec{G}\vec{H}$ $\boxed{r \times d_h}$

$\vec{g}^i$: how much each hidden state contributes to the final embedding representation of $\vec{M}$.

② $\hat{X}_{AETE}$. $\boxed{1 \times d_h}$ → ~~Im~~ Text

To provide a final global text embedding of the sentences in report, the AETE executes max-over-r pooling across M, producing an embedding vector $\hat{X}_{AETE}$ with size d_h.

Image

3. Saliency Weighted Global Average Pooling

$$\hat{X}_{SW\text{-}GAP}(c) = f(\vec{a}_t, \vec{g}_t, X) \quad \boxed{1 \times C}$$

$\hat{X}_{SW\text{-}GAP}$: representing the global visual information, guided by both text- and visual-based attention.

4. Joint Learning.

① Combine both text & image to produce final classification.

$$\hat{X} = [\hat{X}_{AETE} : \hat{X}_{SW\text{-}GAP}]$$

use a Fully-connected Layer to produce the output for multi-label classification

③ Balance the large difference between image loss & text loss.

5. Image Auto-Annotation

→ mine image classification labels.

Test: ① R: reports only.

Input: only reports. (no image).

Output: generated reports.

(image classification label)

② I+R: image + report pairs.

Input: Image + report pairs.

Output: image classification label.

Train: All use ChestX-ray 14 (Report + Image)

6. Automatic Classification & Reporting of Thorax Diseases

→ only have images, classify label.

Train: Input: use both text & image.

Test: Input: one image

Output: ① generated text/report.

② image classification label.

(multiple label prediction)

9.6. Explainable Prediction of Medical codes from Clinical Text.

medical codes { diagnoses
treatment (procedures).

Aim: predict medical codes from clinical (notes) text.

(ICD codes) ⇓ labels ⇓ Document.

Method. classification: Documents – labels

Application: decision support setting:

① explain why it predict eachother &.

② explain what parts of the text ~~are~~ most relevant to each code.

Method.

1. Convolutional Architecture:

to translate docs to H: $d_c \times N$ dimensional matrix.
for pre-label attention mechanism.

Input: $X = [x_1, x_2, \ldots, x_N]$ $\boxed{d_e \times N}$

N: lenth of document.

x_i: d_e: dimensional pre-trained embeddings for each word in the document

Output: H: $\boxed{\mathbb{R}^{d_c \times N}}$

d_c: size of the filter output.

N: length of document.

2. Attention. (Instead Pooling to reduce the Matrix to a vector)

aim: ① to assign multiple lables (i.e. medical codes) for each document.

② An additional benefit: it selects the k-grams from the text that are most relevant to each predicted label.

Input: H: $\boxed{\mathbb{R}^{d_c \times N}}$

Output:

① resulting vector: $H^T u_l$ ($u_l \in \mathbb{R}^{d_c}$) $\boxed{\mathbb{R}^{N \times 1}}$

② attention vector α:

$$\alpha_l = \mathrm{SoftMax}(H^T u_l). \quad \boxed{\mathbb{R}_{1 \times N}}$$

③ vector representations for each label

$$v_l = \sum_{n=1}^{N} \alpha_{l,n} h_n. \quad \boxed{d_c \times 1}$$

Baseline model: compute a single vector v for all labels.

$$v_j = \max_n h_{n,j}. \quad \boxed{d_c \times 1}$$

? z_l: a max-pooled vector,

3. Classification

aim: compute a probability for label l.

using linear layer & a sigmoid transformation.

$$\hat{y}_l = \sigma(\beta_l^T v_l + b_l)$$

Output: 1. (a probability)

4. Training.

minimize the binary cross-entropy loss.

L_2 norm of the ~~modern~~ model weights,

using Adam optimizer.

$\boxed{\text{CAML Done}}$

Optimization:

5. Embedding label descriptions.

Aim: Due to the dimensionality of the label space, many codes are rarely observed in the labeled data. To improve performance on these codes, we add the following regularizing objective to our loss L.

$L(X, y) = L_{BCE} + \dots\dots$

$\boxed{\text{Variant Description CAML} \quad \text{VD-CAML}}$

[4]: Dataset

mimic IV - Full

1. ICD codes (labels): 8921 { diagnose 6918, procedure 2003 }

documents (discharge summaries) ≈ 50,900

- training 47714
- validation 1632
- testing 3372

2. Secondary Evaluations:

mimic IV - 50

ICD top codes (labels): 50

documents (discharge summaries)

- training 8067
- 1574
- 1730

3. Mimic II Full.

labels: ~~20,5~~ 5031

Documents (discharge summaries)

- Training 20.533
- Testing 2282

五 Evaluation.

AUC (micro-R, Macro-R)

F_1 (Micro-R, Macro-R)

P@n.

六. Evaluation of Interpretability

4.2 Results:

Evaluator to judge. (Table 7).