On the Importance of Image Encoding in Automated Chest X-Ray Report Generation

Otabek Nazarov, Mohammad Yaqub, Karthik Nandakumar ✉ otabek.nazarov@mbzuai.ac.ae

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

BMVC 2022

Abstract

- Compare different image encoding approaches (direct, fine-grained, CLIP, and Cluster-CLIP) along with multiple decoders to understand the relative importance of encoder and decoder components.
- Propose a novel cluster CLIP visual encoder (CCVE) that aims to generate more discriminative and explainable representations.

Introduction

- **Clinical problem**. Shortage of radiologists for on time chest X-ray diagnosis.
- **Problem statement.** Given an image of chest X-ray, generate a report capturing abnormalities.
- **Existing works.** Primarily focus on improving decoder and training method, but image encoding is neglected; mainly simply pretrained CNN is used.
- **Dataset.** MIMIC-CXR: ~200,000 image-report pairs

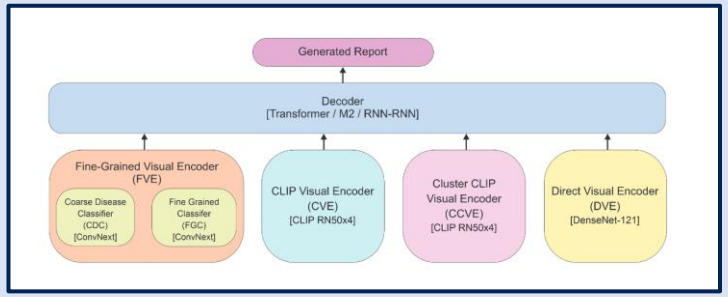

Figure 1. General experimental setup.

Method

- **Direct Visual Encoder (DVE)**. DenseNet-121 trained end-to-end along with decoder.
- **Fine-Grained Visual Encoder (FVE).** Two ConvNext-small classifiers (coarse with 14 classes and fine-grained with 410 classes).
- **CLIP Visual Encoder (CVE).** Contrastive language-image pretraining (CLIP) model trained on reports' impression section.
- **Cluster CLIP Visual Encoder (CCVE).** Novel encoding method designed to produce distinct class embeddings. Image passed through convolutional filter prior to CLIP encoding; filters are selected based on image label during training stage; all filters are used during inference (see Figure 2).
- **Decoders.** Three different decoders are used: transformer, M2, and hierarchical RNN.

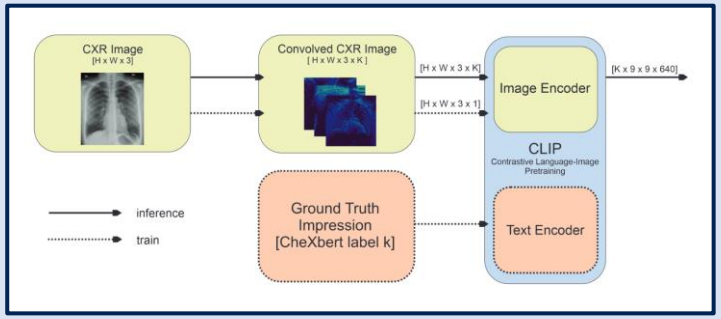

Figure 2. Cluster CLIP Visual Encoder.

Results and Discussion

Model	B1	B2	B3	B4	RG	MTR	CDR	P	R	F1
Transformer Decoder										
DVE	0.286	0.172	0.115	0.083	0.231	0.116	0.109	0.320	0.179	0.169
CCVE	0.267	0.159	0.104	0.074	0.224	0.107	0.091	0.246	0.142	0.108
CVE	0.276	0.165	0.110	0.079	0.221	0.110	0.092	0.382	0.142	0.129
FVE	**0.299**	**0.182**	**0.124**	**0.090**	**0.238**	**0.123**	**0.136**	**0.443**	**0.212**	**0.220**
M2 Decoder										
DVE	0.297	0.181	0.123	0.089	0.238	0.123	0.129	**0.418**	0.205	0.211
CCVE	0.266	0.159	0.105	0.073	0.224	0.108	0.090	0.249	0.146	0.130
CVE	0.278	0.167	0.112	0.081	0.227	0.112	0.103	0.206	0.345	0.116
FVE	**0.298**	**0.183**	**0.124**	**0.090**	**0.242**	**0.125**	**0.137**	0.402	**0.232**	**0.236**
RNN-RNN Decoder										
DVE	**0.289**	**0.171**	**0.114**	**0.081**	0.228	0.112	0.112	0.296	0.163	0.153
CCVE	0.246	0.147	0.097	0.068	0.225	0.104	0.096	0.243	0.137	0.103
CVE	0.254	0.152	0.101	0.071	0.226	0.106	0.096	**0.344**	0.140	0.100
FVE	0.277	0.167	0.112	0.080	**0.235**	**0.116**	**0.116**	0.309	**0.187**	**0.172**

- **FVE** showed the best performance; thus, semantic information extraction is a key for effective image encoding.
- **CLIP-based** performed poorly
  - CCVE gave ROC-AUC of 0.71, while FVE gave ROC-AUC of 0.83
  - Contrastive training might be focusing on wrong words during training
- **Future work.**
  - CLIP-based methods need proper training method for medical data
  - Explore other methods that better extract semantic information.