Multi-modality Self-supervised learning

- MIMIC-CXR Database
- Related work survey
- Our Strategy

Jong Hak Moon & Hyungyung Lee

Aug. 1st. 2020

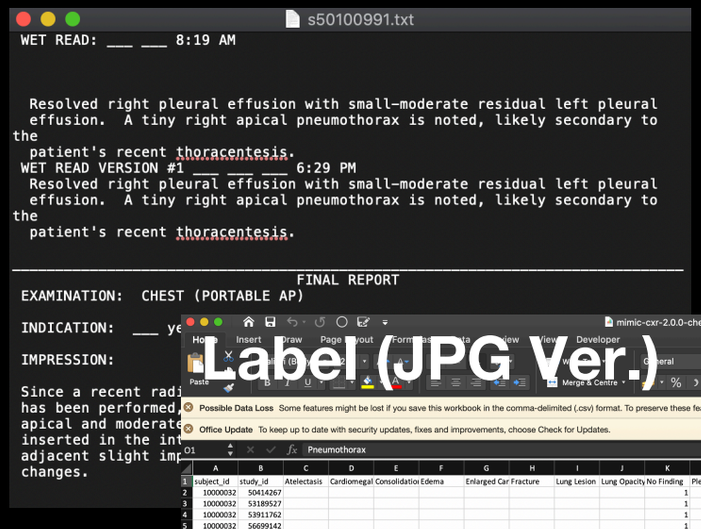
Open source dataset

Chest X-ray Dataset

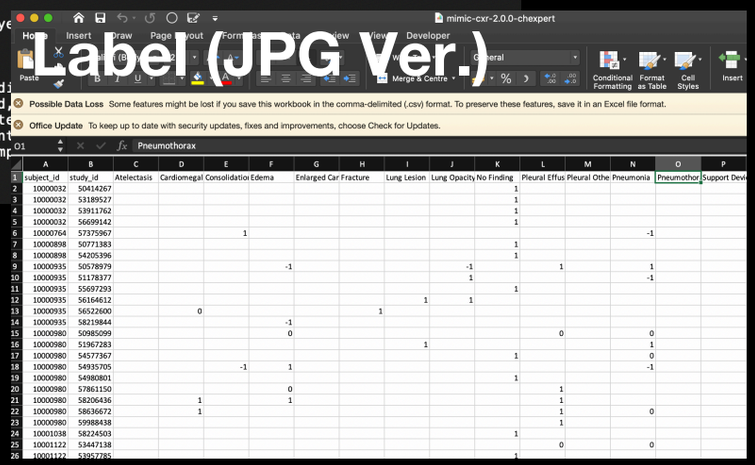
Dataset	Source Institution	Disease Labeling	# Images	# Reports	# Patients
CheXpert	Stanford Hospital	Automatic (CheXpert labeler)	224,316	None	65,240
Chest-Xray8	National Institutes of Health	Automatic (Dnorm+MetaMap)	108,948	None	32,717
Chest-Xray14	National Institutes of Health	Automatic (Dnorm+MetaMap)	112,120	None	30,805
PadChest	Hospital Universitario de San Juan	Expert + Automatic (Neural Network)	160,868	206,222 (Written in Spanish)	67,625
Open-I	Indiana Network for Patient Care	Expert	8,121	3,996	3,996
✓ **MIMIC-CXR**	**Beth Israel Deacones Medical Center**	**Automatic (CheXpert labeler)**	**473,057**	**206,563**	**63,478**

MIMIC-CXR Dataset

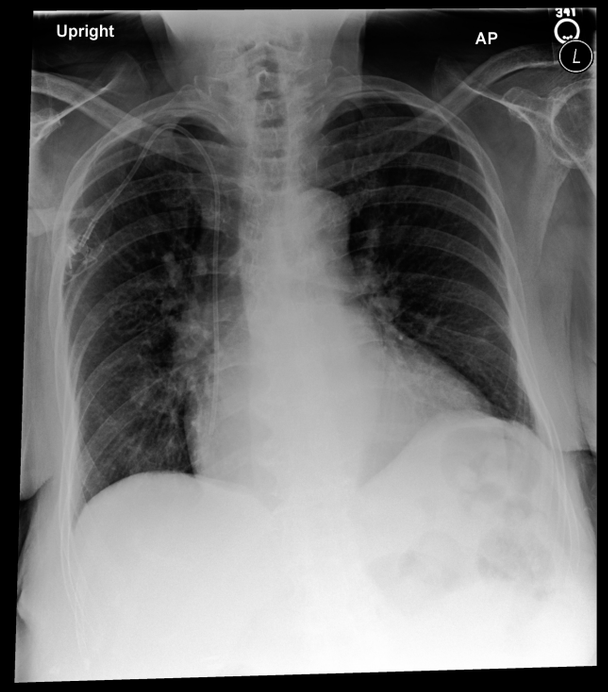
Free-text (DICOM Ver.)



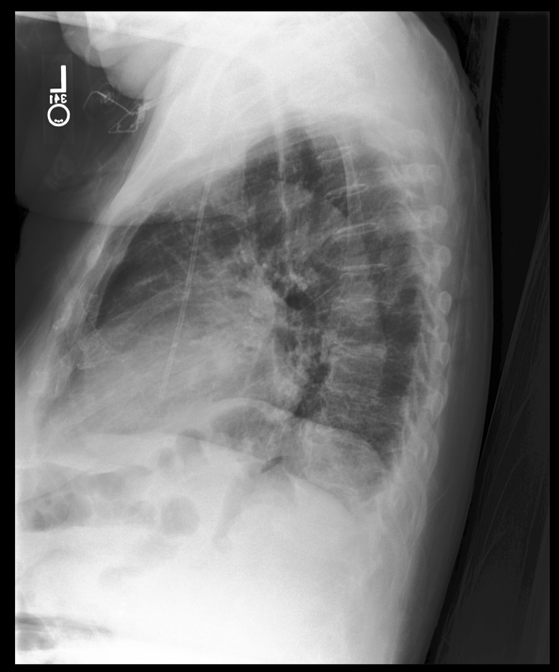
Label (JPG Ver.)



Frontal



Lateral



- ✓ Data : Image (DICOM,JPG), Free-text reports(DICOM ver.), Labels (JPG ver. Chexpert & Negbio) [1:positively mentioned, 0: negatively mentioned, -1:ambiguous]
- ✓ DICOM Image (Pixel value over 255), JPG Image (Pixel value normalized 0-255)
- ✓ Dataset consists of 377,110 images corresponding to 227,827 radiographic studies, and electronic health record (EHR), In 47% (107,186 images) of all data, a single class is shown.
- ✓ Multimodal (using free-text & Image) deep learning via self-supervised approach

Related work survey

Key words

✓ 1. Multi-modal deep learning approach

✓ 2. Self-supervised approaches in vision

3. Self-supervised approaches in natural language

Related work survey <Multi-modality>

✓ CVPR 18> TieNet- Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays On the Automatic Generation of Medical Imaging Reports

MICCAI 18> Multimodal Recurrent Model with Attention for Automated Radiology Report Generation

✓ NeurIPS 18> Unsupervised Multimodal Representation Learning across Medical Images and Reports

<Reinforcement> NeurIPS 18> Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation

<Graph Transformer> CVPR 19> Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation

✓ NeurIPS 19> Baselines for Chest X-Ray Report Generation

IEEE Access 19> Multi-Attention and Incorporating Background Information Model for Chest X-Ray Image Report Generation

MICCAI 19> Automatic Radiology Report Generation based on Multi-view Image Fusion and Medical Concept Enrichment

✓ <Reinforcement> CVPR 19> Clinically Accurate Chest X-Ray Report Generation

Related work survey <Self-supervised in vision>

NeurIPS 19> Learning Representations by Maximizing Mutual Information Across Views

CVPR 19> Self-Supervised Learning of Pretext-Invariant Representations

CVPR 20> Momentum contrast for unsupervised visual representation learning

Under review> Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning

Under review> Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

ICML 20> Data-efficient image recognition with contrastive predictive coding

ICML 20> A Simple Framework for Contrastive Learning of Visual Representations

Related work survey <BERT based Self-supervision>

Clinical NLP

NAACL 19> Publicly Available Clinical BERT Embeddings

MLHC 20> CheXpert++: Approximating the CheXpert labeler for Speed, Differentiability, and Probabilistic Output

Arxiv 20> CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT

Multi-modality

NIPS 19> ViLBERT- Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

ICLR 20> VL-BERT- PRE-TRAINING OF GENERIC VISUALLINGUISTIC REPRESENTATION

AAAI 20> Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training

Related Work Survey

Methods for Radiology Report Generation using chest X-ray Image and Free-text

Objective : Classification & Automatic Generation of Medical Reports

1>

1. TieNet- Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays On the Automatic Generation of Medical Imaging Reports (CVPR 18)

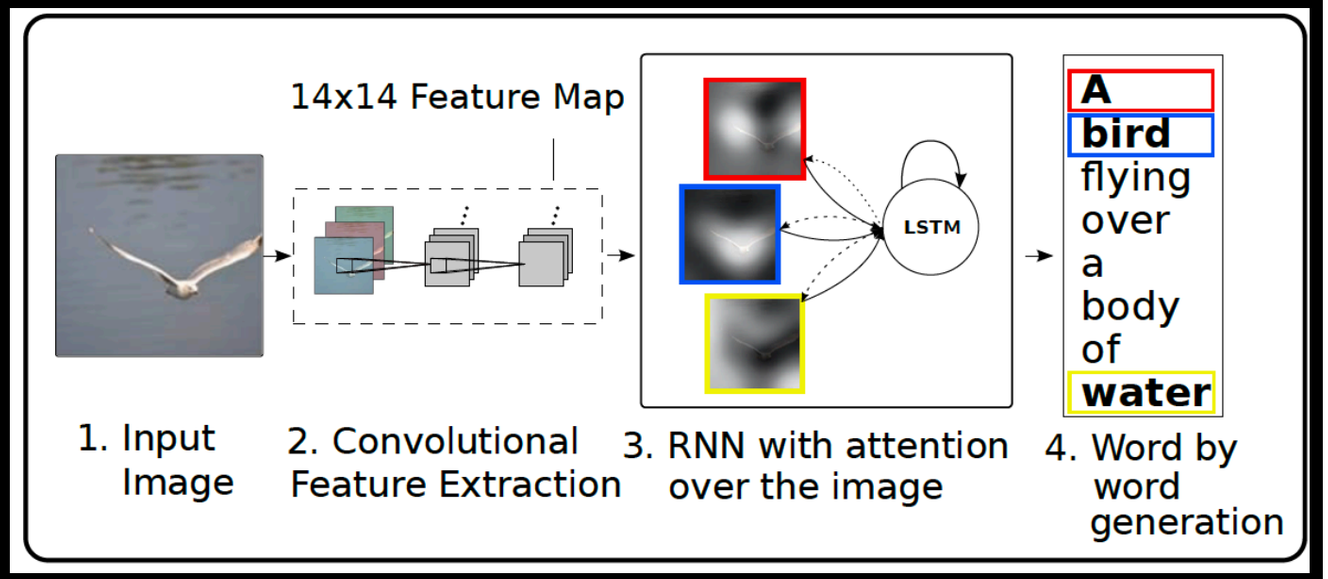
Dataset : ChestX-ray14

Motivation : Image-text attention based approach

Objective : Classification & Automatic Generation of Medical Reports

Main Idea

1> End-to-End Trainable CNN-RNN Model
2> Attention Encoded Text Embedding
3> Saliency Weighted Global Average Pooling
4> Joint Learning



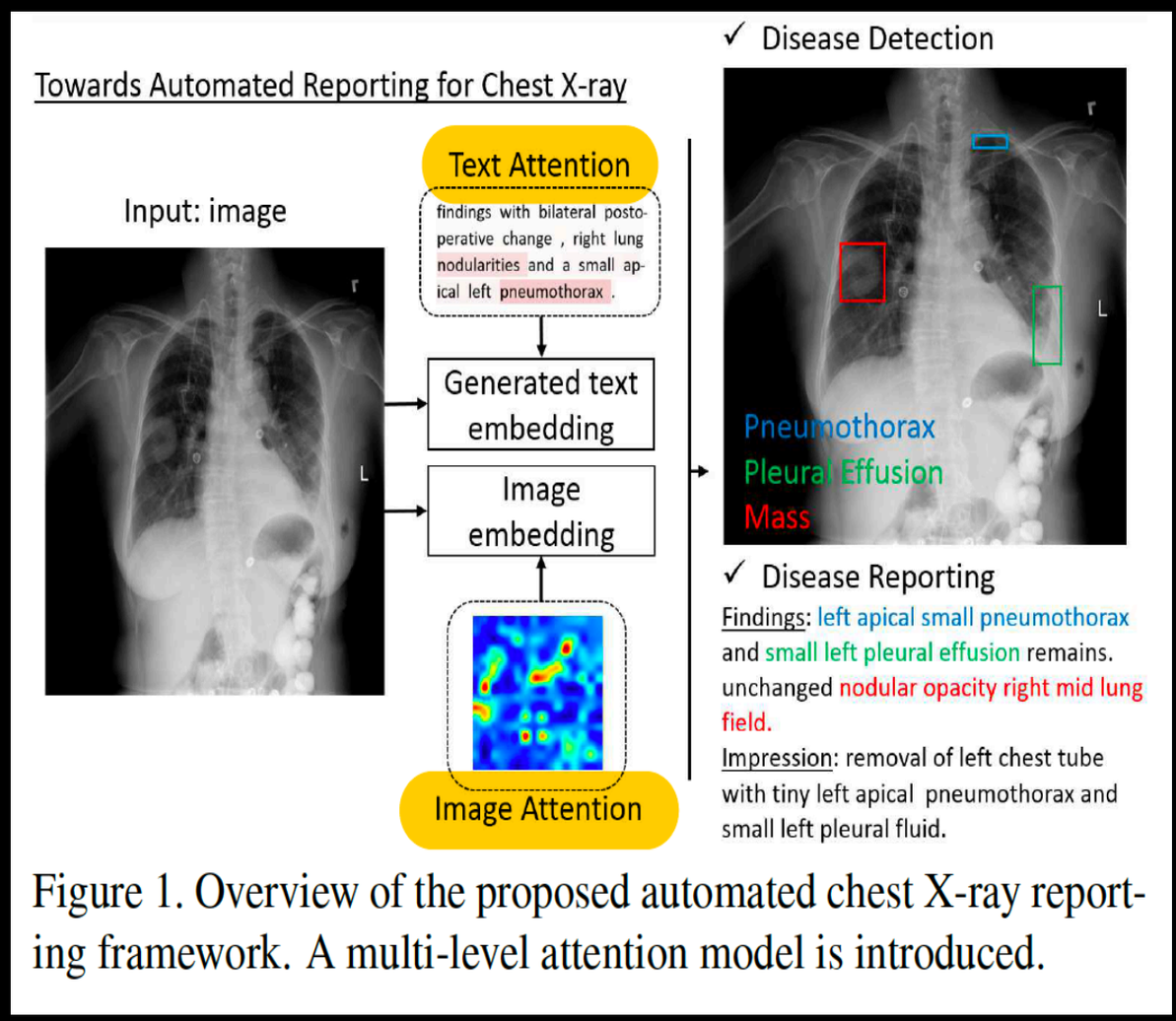


Figure 1. Overview of the proposed automated chest X-ray report-ing framework. A multi-level attention model is introduced.

1. TieNet- Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays On the Automatic Generation of Medical Imaging Reports (CVPR 18)

Method –Frame work

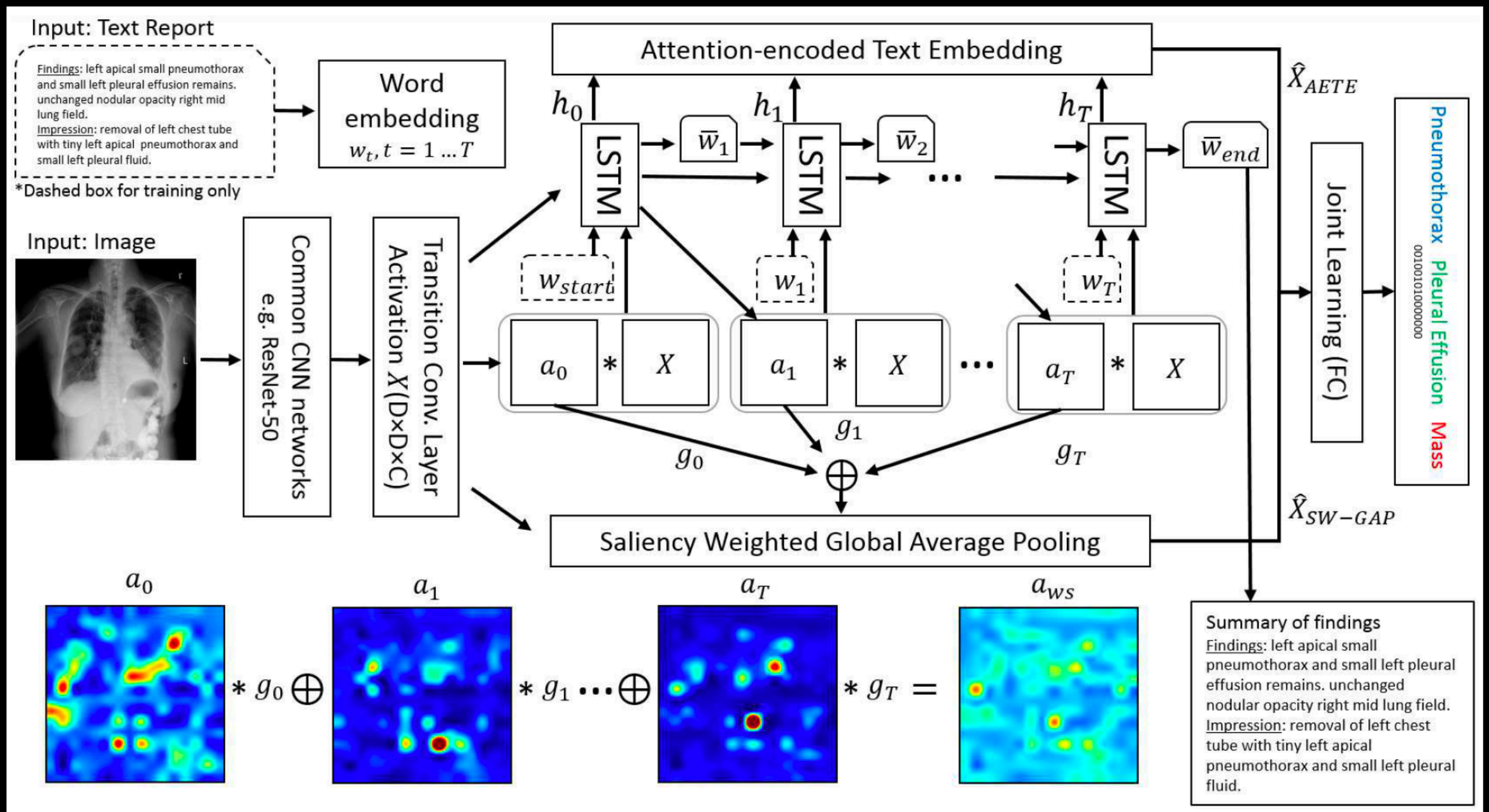
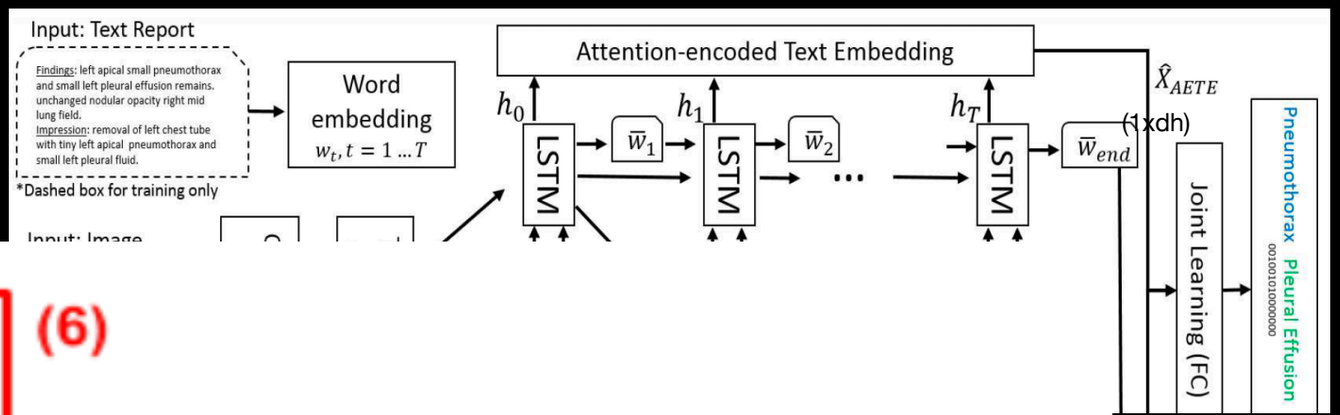


Figure 2. Framework of the proposed chest X-ray auto-annotation and reporting framework. Multi-level attentions are introduced to produce saliency-encoded text and image embeddings.

1. TieNet- Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays On the Automatic Generation of Medical Imaging Reports (CVPR 18)

Method –Frame work



1. Atte

- T

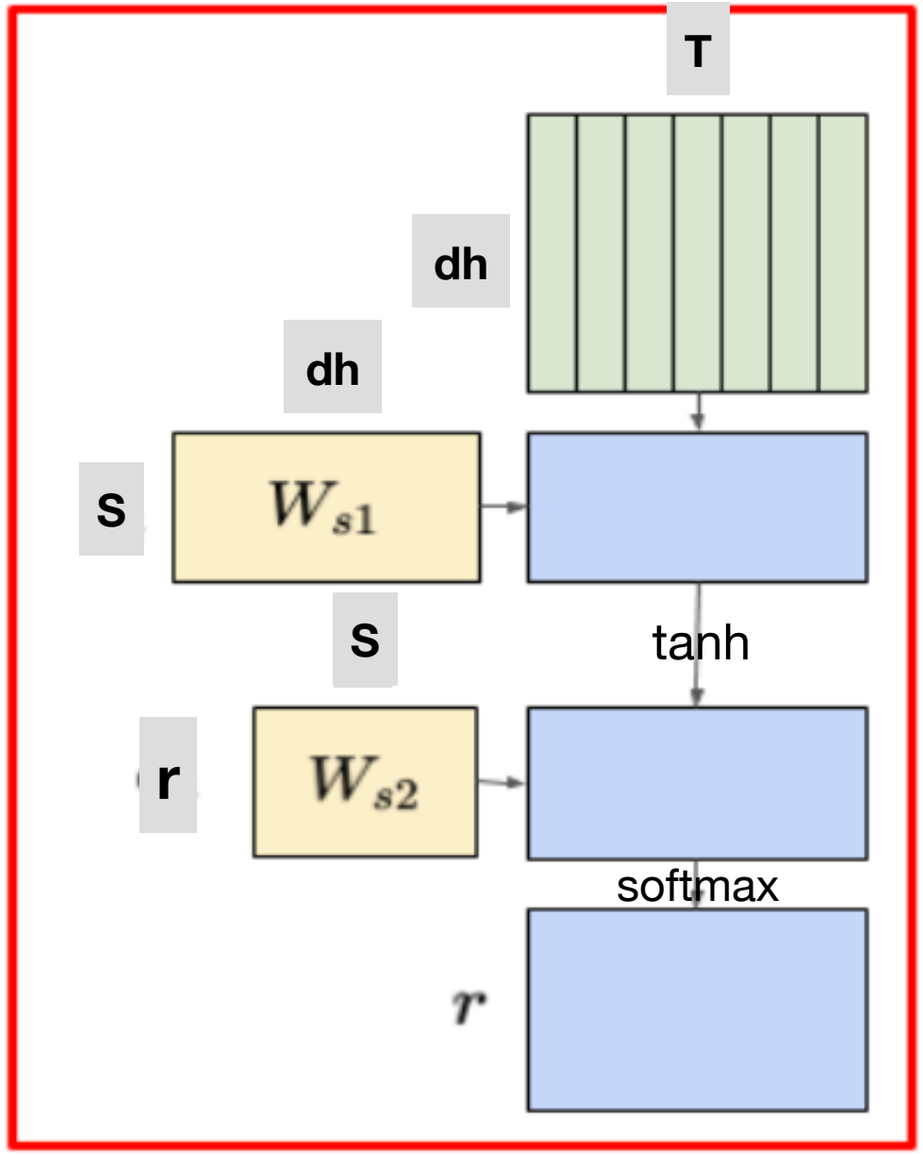
po

- T

- r

- H

- M



(6)

$$H = (h_1, h_2, \cdots, h_n)$$

$$A' = \tanh(W_{s1} H^T) : d_a \times n$$

$$A = Softmax(W_{s2} A') : r \times n$$

ICLR 2017) A Structured Self-Attentive Sentence Embedding

1. TieNet- Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays On the Automatic Generation of Medical Imaging Reports (CVPR 18)

Method –Frame work

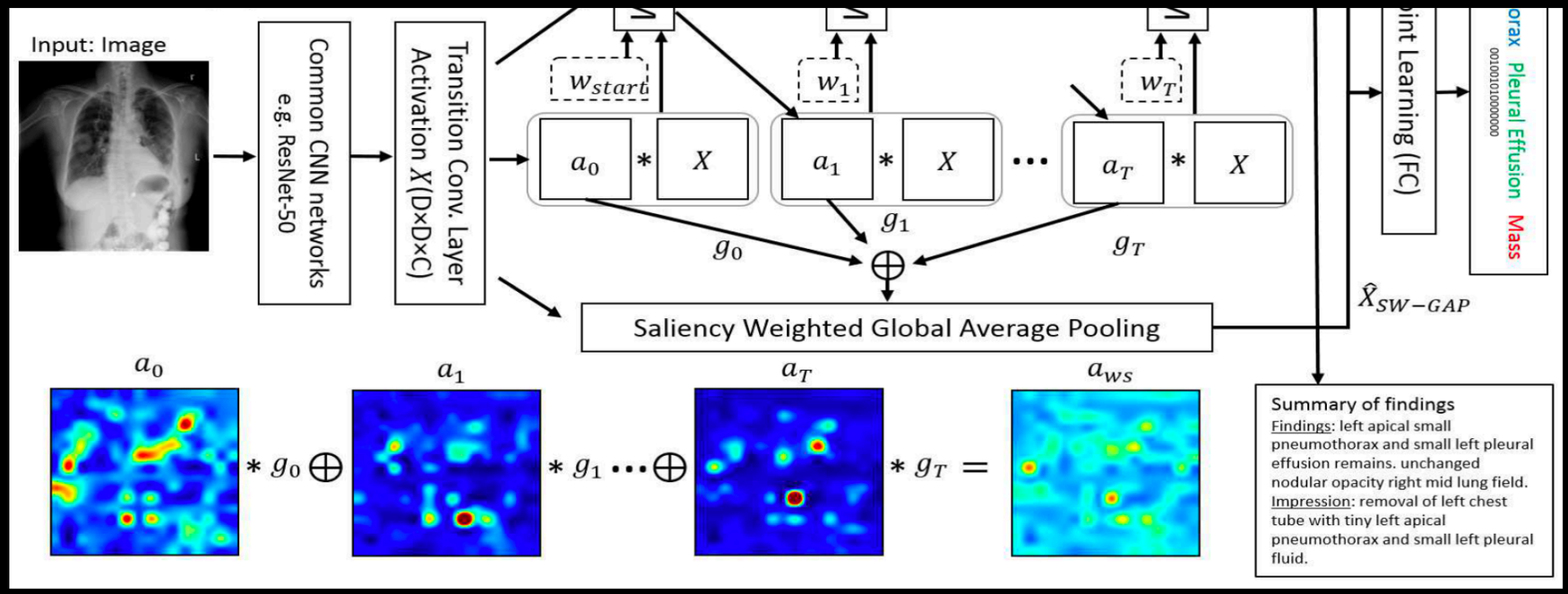
1. Saliency weighted Global Average Pooling (SW-GAP)

- To provide a more meaningful text embedding, we re-use the attention mechanism G, we perform a max-over-r operation to produce a sequence of saliency values, gt(t=1,...,T), for each word, wt. These saliency values are used to weight and select the spatial attention maps, at, generated at each time point:

$$\mathbf{a}_{ws}(x,y) = \sum_t \mathbf{a}_t(x,y) * g_t. \quad (4)$$

$$\hat{\mathbf{X}}_{SW-GAP}(c) = \sum_{(x,y)} \mathbf{a}_{ws}(x,y) * \mathbf{X}(x,y,c), \quad (5)$$

where x, y ∈ {1...D} and XSW –GAP is a 1-by-C vector representing the global visual information, guided by both text- and visual-based attention



1. TieNet- Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays On the Automatic Generation of Medical Imaging Reports (CVPR 18)

Quantitative Analysis

Disease	ChestX-ray14					Hand-labeled					OpenI				
	R /	I+R /	I [35] /	I+GR /	#	R /	I+R /	I [35] /	I+GR /	#	R /	I+R /	I [35] /	I+GR /	#
Atelectasis	.983 /	.993 /	.700 /	.732 /	3255	.886 /	.919 /	.680 /	.715 /	261	.981 /	.976 /	.702 /	.774 /	293
Cardiome															
Effusion															
Infiltratio															
Mass															
Nodule															
Pneumon															
Pneumotl															
Consolida															
Edema															
Emphyse															
Fibrosis															
PT															
Hernia															
NoFindin															
AVG															
#wAVG															

Table 1. Eva
four method

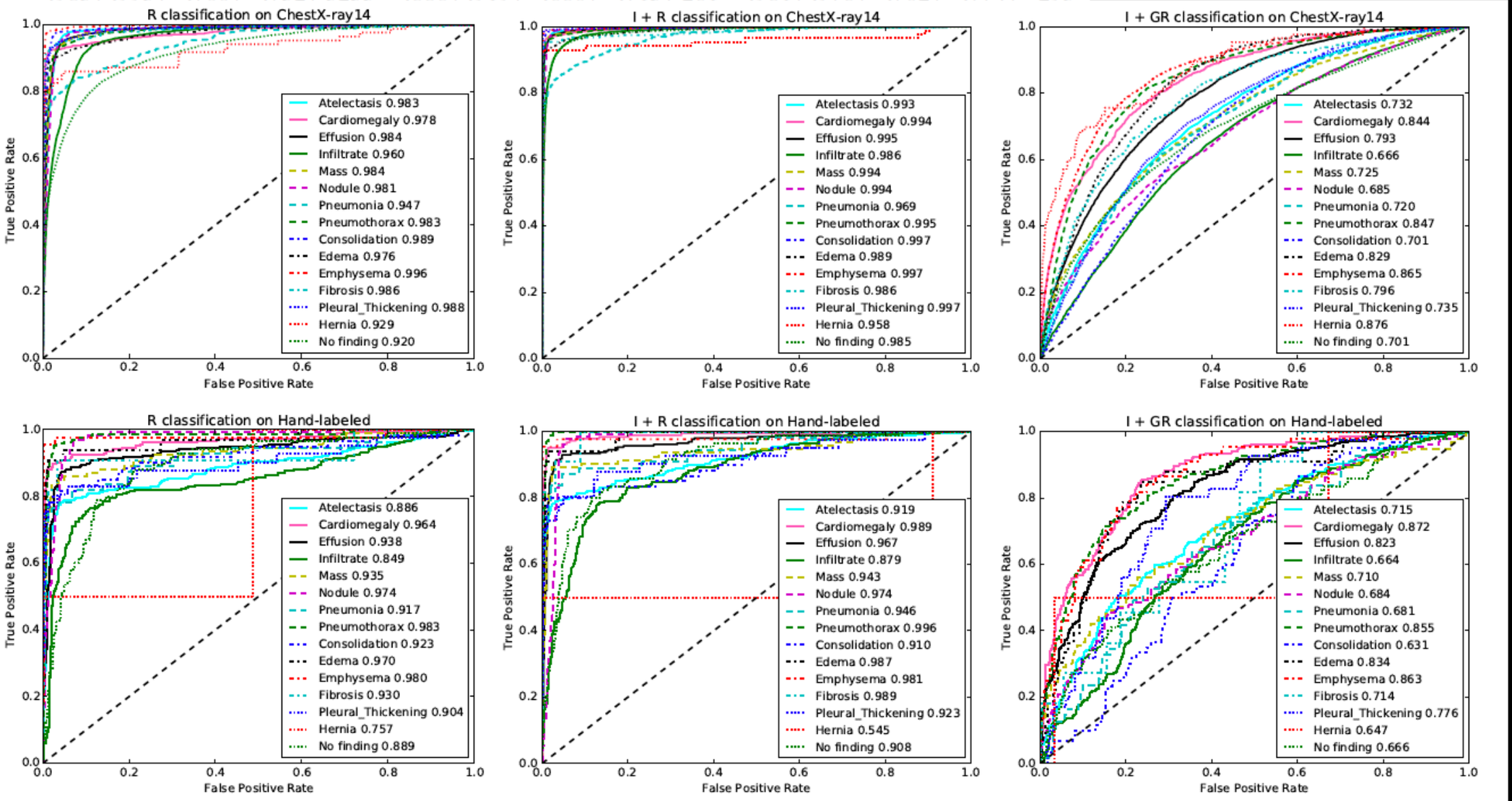


Figure 3. A comparison of classification performance using ROC curves with different testing inputs, *i.e.* Report (R), Image+Report (I+R), and Image+Generative Report(I+GR).

1. TieNet- Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays On the Automatic Generation of Medical Imaging Reports (CVPR 18)

Qualitative Analysis

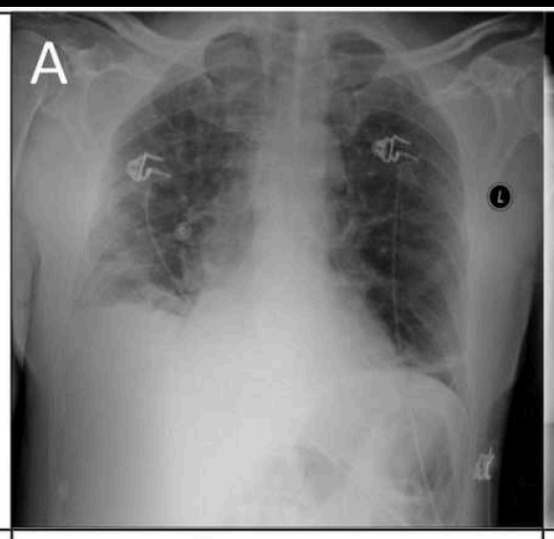
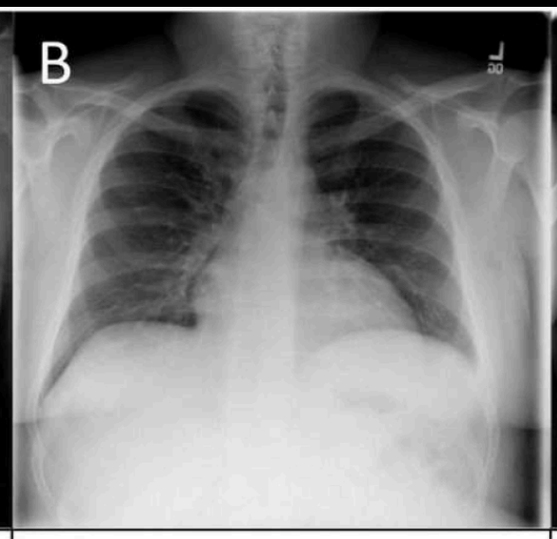
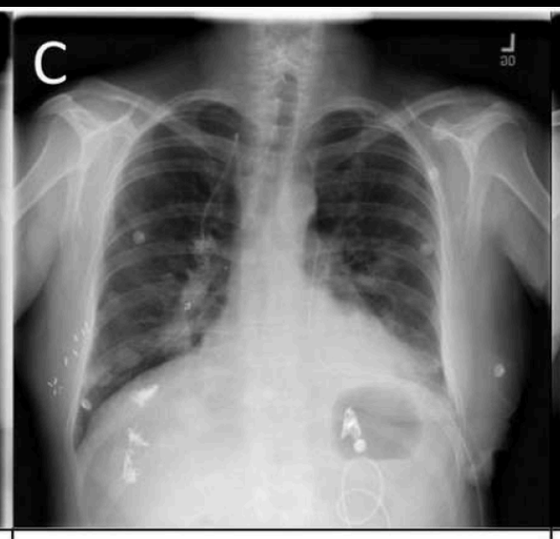
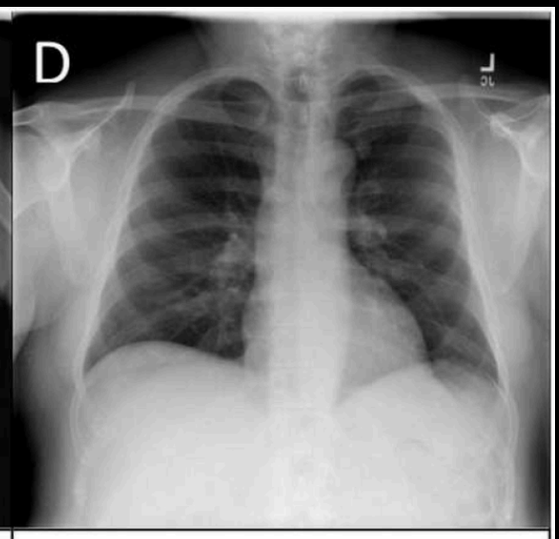
Image Sample cases				
P	Atelectasis Effusion	No finding	Nodule Pneumothorax Mass Consolidation	Mass
Original report	findings : a single ap view of the chest demonstrates increasing bibasilar interstitial opacities with decreased overall aeration . increasing blunting of right costophrenic angle. … impression : increasing bibasilar atelectasis with possible development of right pleural effusion .	Normal no evidence of lung infiltrate .	findings : heart and mediastinum unchanged . multiple lung nodules . evidence of recent left chest surgery with left chest tube in place . very small left apical pneumothorax . lungs unchanged , no evidence of acute infiltrates . impression : stable chest .	findings : large left suprahilar and infrahilar masses as well as the well circumscribed nodule the level of the aortic knob . the right infrahilar mass as well . no effusion . impression : metastatic lung disease .
Generated Report	findings : a single ap view of the chest demonstrates unchanged bilateral reticular opacities , consider atelectasis . continued left basilar atelectasis . no evidence of developing infiltrate . the cardiac and mediastinal contours are stable . impression : no evidence of developing infiltrate .	findings : pa and lateral views of the chest demonstrate lungs that are clear without focal mass , infiltrate or effusion . cardiomediastinal silhouette is normal size and contour . pulmonary vascularity is normal in caliber and distribution . impression : no evidence of acute pulmonary pathology	findings : pa and lateral views of the chest demonstrate unchanged bilateral chest tubes . again pulmonary nodules are seen on the right and cardiac silhouette unchanged . the cardiac and mediastinal contours are stable . impression : 1. bilateral masses and left lower lung field consolidation . 2.new bilateral lung masses .	comparison is to previous upright study of no significant interval change is seen in the appearance of the chest . the mediastinal soft tissue and pulmonary vascularity are stable . there are blastic bone lesions in the chest . bones , soft tissues are normal . the lung fields are clear . there are calcified lymph nodes in the left lower lung . impression : . sclerotic lesions in the left humeral , consistent with metastasis.

Figure 4. 4 sample image Classification Predictions (P) along with original and generated reports. Text attentions are highlighted over the generated text. Correct predication is marked in green, false prediction in red and missing prediction in blue.

2. Baselines for Chest X-Ray Report Generation <NIPS 19>

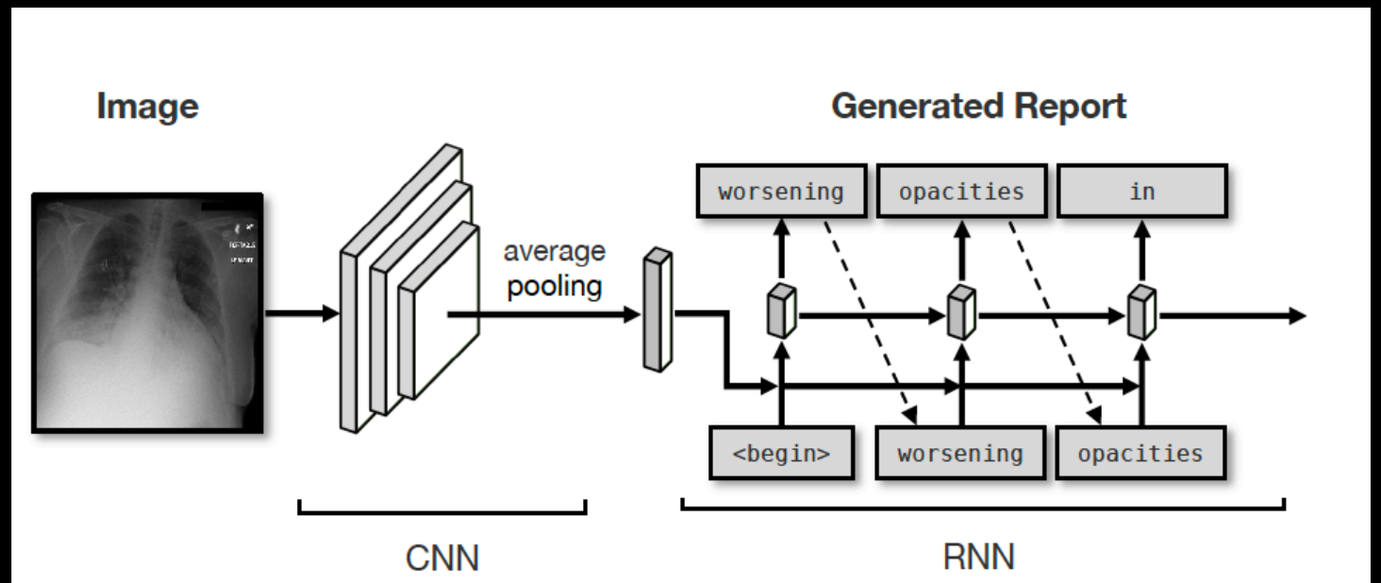


Figure 2: Full CNN-RNN pipeline.

Dataset : MIMIC-CXR

Motivation : Focus on the Natural Language Generation

Objective : Automatic Generation of Medical Reports

Main Idea : 4 different types of language generation model

1> Random Retrieval Baseline

2> Conditional n-gram Language Model

3> Nearest Neighbor

4> CNN+RNN (+ Beam)

Image Embedding : DenseNet121model 8x8x1024 to 1024-D (by global average-pooling)

2. Baselines for Chest X-Ray Report Generation <NIPS 19>

4 types of language generation model

1> Random Retrieval Baseline : It is unconditioned upon the query image, and instead merely draws a random report from the training set. These reports will be readable, but not relevant to the query image at all

2> Conditional n-gram Language Model : These models are learned from simply tallying how often a given word actually follows a given phrase in the training set. We make these models conditional on a query image by learning a per-instance language model for each image based on the reports corresponding to the closest 100 train images

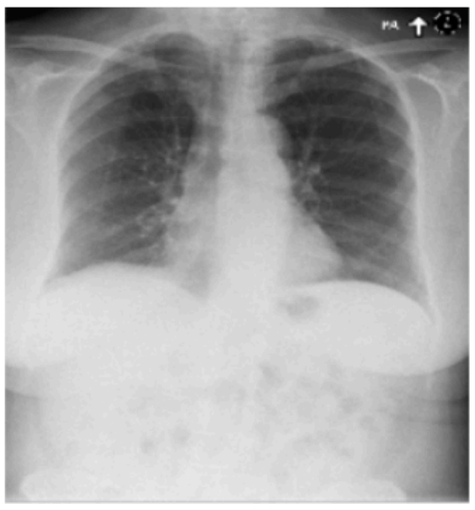
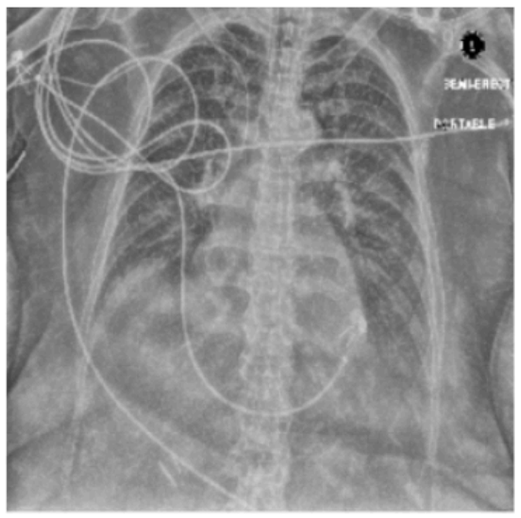
3> Nearest Neighbor : we generate our text by returning the caption of the training image with the largest cosine similarity (in the DenseNet-induced space) to the test query image. Here they should also be clinical relevant.

✓ 4> CNN+RNN (Greedy decoding / Beam search decoding) : Beam search decoding mechanism will always track the subsequent top-K most likely next tokens at each step, exploring a broader set of possible sentences. Final candidates are reweighted by a scoring function which encourages longer, more complex sentences, before finally returning and output sequence.

2. Baselines for Chest X-Ray Report Generation <NIPS 19>

Evalu

We as … e
CheX… ee
with t…
precis…

Image		
Reference	pa and lateral views of the chest demonstrate the lungs are well-expanded and clear. the cardiomediastinal silhouette is normal. there is no pleural effusion or pneumothorax.	in comparison with the study of DATE, the monitoring and support devices are in essentially unchanged position. there is again mild enlargement of the cardiac silhouette with pulmonary edema and bilateral layering pleural effusions.
3-gram	pa and lateral views of the chest . there is no pleural effusion , or pleural effusion or pneumothorax . the cardiomediastinal silhouette is within normal limits . lungs are essentially clear . no acute osseous abnormality . levoconvex scoliosis of the chest were obtained . low lung volumes . there are no pleural effusion or pneumothorax is seen . the mediastinal and hilar contours are normal .	et tube , enteric tube tip is difficult to assess the status of the intra-aortic balloon pump , with mild increase in pulmonary outflow tract remain unchanged . at the level of the exam is a moderate left-sided pleural effusion with bibasilar pelural fluid and atelectasis . there are no acute bony abnormality .
KNN	left pleural tube is in stable position. there has been a slight increase in the left pleural effusion with increased atelectasis at the left base. there is a stable left apical pneumothorax and atelectasis at the right base. cardiomediastinal and hilar contours are stable. there is no focal consolidation concerning for pneumonia.	on the first radiograph, obtained at 1249, there was malposition of the dobbhoff catheter in the right bronchial system. no evidence of pneumothorax or other complications. on the radiograph performed at 1255, the dobbhoff catheter follows the course of the esophagus, with the tip in the proximal parts of the stomach. again, no complication such as pneumothorax is seen.
CNN-RNN + Beam	pa and lateral views of the chest were obtained . no focal consolidation , pleural effusion , or evidence of pneumothorax . the cardiac and mediastinal silhouettes are unremarkable .	the et tube is in the stomach . there is no pneumothorax is in the tip in the svc . there is no pneumothorax . there is a NAME right pleural effusion is unchanged . no pneumothorax . the heart size is normal . the mediastinal and hilar contours are normal .

Model	… pert F1
Rando…	…148
1-gram	…174
2-gram	…193
3-gram	…185
1-NN	**…258**
CNN-…	…067
CNN-…	…186

Figure 3: Example outputs from the best versions of each model type.

3. Clinically Accurate Chest X-Ray Report Generation (CVPR 19)

Dataset : MIMIC-CXR (anteroposterior (AP) views)

Motivation : Readable reports, Importance of clinical accuracy

Objective : Automatically generate clinically accurate medical reports

Main Idea : Reinforcement learning

1> Hierarchical Generation via CNN-RNN-RNN

2> Reinforcement Learning for Readability (Natural Language Generation Reward)

3> Novel Reward for Clinically Accurate Reinforcement Learning (Clinically Coherent Reward)

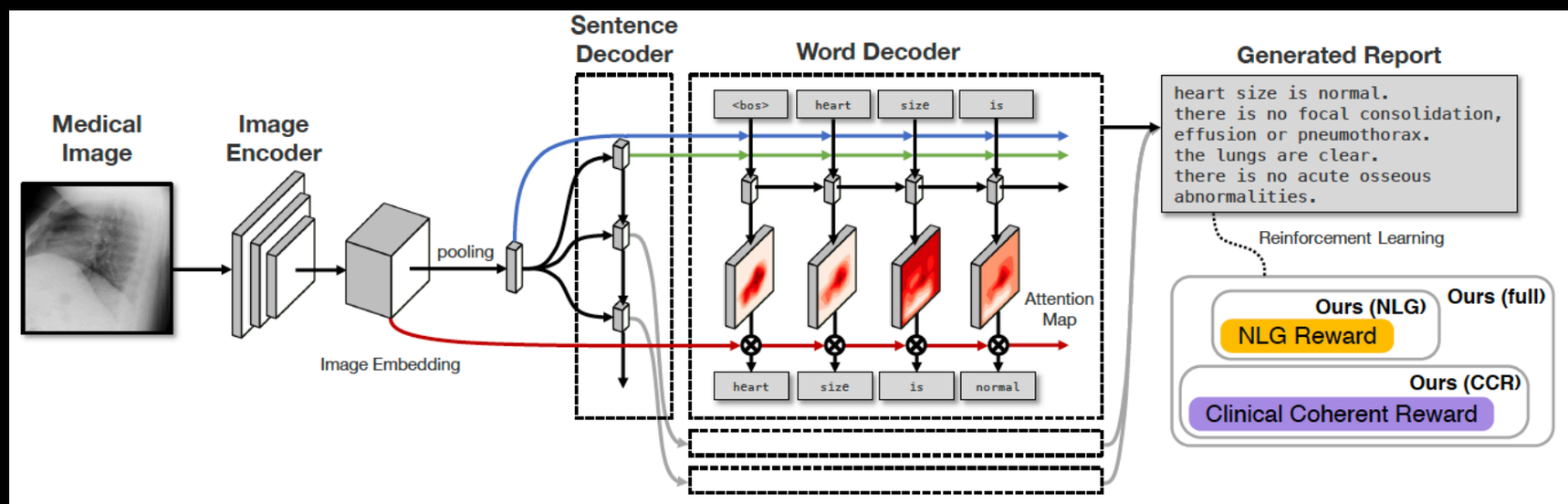


Figure 2: **The model for our proposed *Clinically Coherent Reward*.** Images are first encoded into image embedding maps, and a sentence decoder takes the pooled embedding to recurrently generate topics for sentences. The word decoder then generates the sequence from the topic with attention on the original images. NLG reward, clinically coherent reward, or combined, can then be applied as the reward for reinforcement policy learning.

3. Clinically Accurate Chest X-Ray Report Generation (CVPR 19)

Method – Text Image Embedding Network

1> Hierarchical Generation via CNN-RNN-RNN

Image encoder CNN : Input image is passed through a CNN to obtain the last layer

Sentence decoder RNN : Producing two vector (topic vector, stop vector) using LSTM

Word decoder RNN : After decode the sentence topics, we start to decode the words given the topic vector.

2> Reinforcement Learning for Readability

Reward function that takes a sampled report and a ground truth report to minimize the negative expected reward via self-critical sequence training (SCST)

3> Novel Reward for Clinically Accurate Reinforcement Learning (Clinically Coherent Reward)

Rule based disease mention annotator, Chexpert, to optimize our generated report for clinical efficacy directly.

3. Clinically Accurate Chest X-Ray Report Generation (CVPR 19)

Result

	Model	Natural Language						Clinical
		CIDEr	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Accuracy
MIMIC-CXR	*Major Class*	-	-	-	-	-	-	0.828
	Noise-RNN	0.716	0.272	0.269	0.172	0.113	0.074	0.803
	1-NN	0.755	0.244	0.305	0.171	0.098	0.057	0.818
	S&T	0.886	0.300	0.307	0.201	0.137	0.093	0.837
	SA&T	0.967	0.288	0.318	0.205	0.137	0.093	0.849
	TieNet	1.004	0.296	0.332	0.212	0.142	0.095	0.848
	Ours (NLG)	**1.153**	**0.307**	**0.352**	**0.223**	**0.153**	**0.104**	0.834
	Ours (CCR)	0.956	0.284	0.294	0.190	0.134	0.094	**0.868**
	Ours (full)	1.046	**0.306**	0.313	0.206	0.146	**0.103**	**0.867**
Open-I	*Major Class*	-	-	-	-	-	-	0.911
	Noise-RNN	0.747	0.291	0.233	0.130	0.087	0.061	0.914
	1-NN	0.728	0.201	0.232	0.116	0.051	0.018	0.911
	S&T	0.926	0.306	0.265	0.157	0.105	0.073	0.915
	SA&T	1.276	0.313	0.328	0.195	0.123	0.080	0.908
	TieNet	1.334	0.311	0.330	0.194	0.124	0.081	0.902
	Ours (NLG)	**1.490**	**0.359**	**0.369**	**0.246**	**0.171**	**0.115**	0.916
	Ours (CCR)	0.707	0.244	0.162	0.084	0.055	0.036	**0.917**
	Ours (full)	1.424	0.354	0.359	0.237	0.164	**0.113**	**0.918**

Table 2: **Automatic Evaluation Scores.** The table is divided into natural language metrics and clinical finding accuracy scores. BLEU-n counts up n-gram for evaluation, and accuracy is the averaged macro accuracy across all clinical findings. *Major class* always predicts negative findings.

3. Clinically Accurate Chest X-Ray Report Generation (CVPR 19)

Result

Ground Truth	TieNet	Ours (full)
cardiomegaly is moderate. bibasilar atelectasis is mild. there is no pneumothorax. a lower cervical spinal fusion is partially visualized. healed right rib fractures are incidentally noted.	ap portable upright view of the chest. there is no focal consolidation, effusion, or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact.	pa and lateral views of the chest. there is mild enlargement of the cardiac silhouette. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.
ap upright and lateral views of the chest. there is moderate cardiomegaly. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.	as compared to the previous radiograph, there is no relevant change. tracheostomy tube is in place. there is a layering pleural effusions. NAME bilateral pleural effusion and compressive atelectasis at the right base. there is no pneumothorax.	
(a)	(b)	

Figure 3: **Visualization of the generated report and image attention maps.** Different words are underlined with its corresponding attention map shown in the same color. Best viewed in color.

as at the right base favoring resolving atelectasis. no pneumothorax is appreciated on this semi upright study. heart remains stably enlarged. mediastinal contours are stably widened, although this NAME be related to portable technique and positioning. this can be better evaluated on followup imaging. no pulmonary edema.	nasogastric tube is in unchanged position. the lung volumes remain low. moderate cardiomegaly with minimal fluid overload but no overt pulmonary edema. no larger pleural effusions. no pneumonia.	chest. there is moderate cardiomegaly. there is no pleural effusion or pneumothorax. there is no acute osseous abnormalities.

Table 4: **Sample images along with ground truth and generated reports.** Note that upper case tokens are results of anonymization.

4. Unsupervised Multimodal Representation Learning across Medical Images and Reports (NIPS 18)

Dataset : MIMIC-CXR (anteroposterior (AP) views)

Motivation : Parallel image/report pairs are not always feasible

Objective : Multimodal Representation Learning

Main Idea : Establish baseline results and evaluation methods for jointly embedding radiological images and reports via retrieval and distance metrics.

- We profile the impact of supervision level on the quality of representation learning in joint embedding spaces.

- We characterize the influence of using different sections from the report on representation learning.

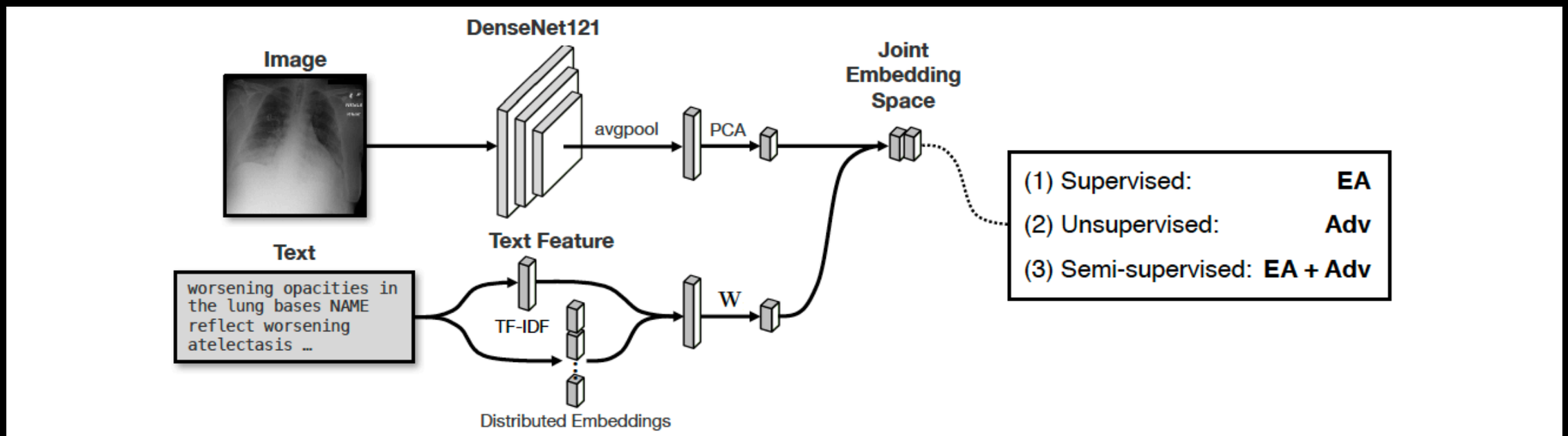


Figure 1: The overall experimental pipeline. EA: embedding alignment; Adv: adversarial training.

4. Unsupervised Multimodal Representation Learning across Medical Images and Reports (NIPS 18)

All methods are follow this process

Images : Resized to 256 x 256 -> DenseNet 121 -> PCA onto the 1024-d to obtain 64-d features.

4 types of Text embedding:

1. Term frequency-inverse document frequency (TF-IDF) over bi-grams
2. Pre-trained GloVe word embeddings
3. Sentence Embeddings or paragraph embeddings -> After sentence/paragraph splitting, fine-tune a deep averaging network (DAN) encoder with the corpus.

Embedding Alignment (EA) for Supervised learning

-> linear transformation between two sets of matched points

Adversarial Domain Adaption (Adv) for Unsupervised learning

-> Discriminator D implemented as a 2 additional layer using scaled exponential linear unit (SELUs) against a projection matrix W, as the generator. D is trained in the joint space, W is trained adversarially to fool D.

EA + Adv for Semi-supervised learning

4. Unsupervised Multimodal Representation Learning across Medical Images and Reports (NIPS 18)

Result

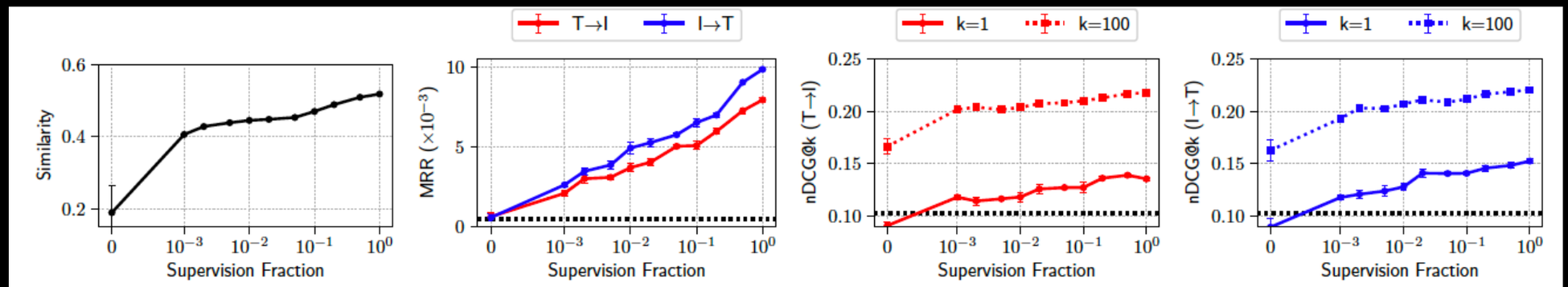


Figure 2: Performance measures of retrieval tasks at k retrieved items as a function of the supervision fraction. Higher is better. Note the x-axis is in log scale. Unsupervised is on the left, increasingly supervised to the right. Dashed lines indicate the performance by chance. Vertical bars indicate the 95% confidence interval, and some are too narrow to be visible.

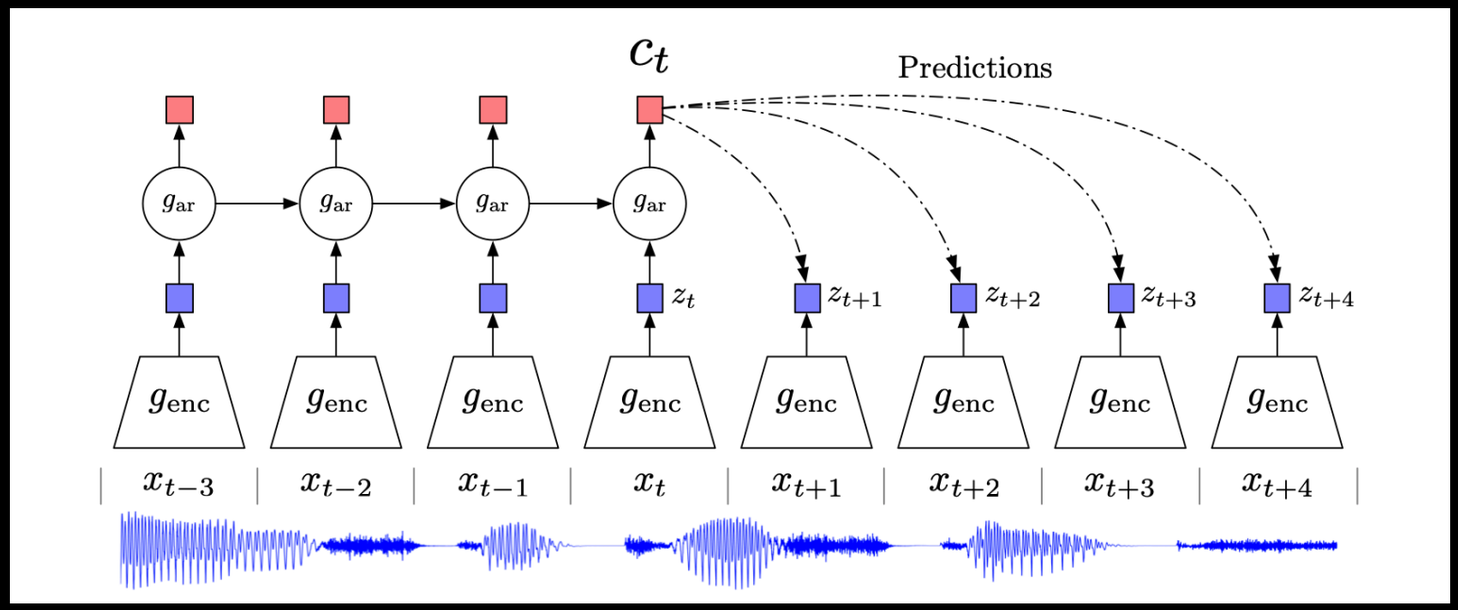
Text Feature	Method	Similarity	MRR($\times 10^{-3}$)		nDCG@1		nDCG@10		nDCG@100	
			T → I	I → T	T → I	I → T	T → I	I → T	T → I	I → T
	chance		0.50	0.50	0.103	0.103	0.103	0.103	0.103	0.103
bi-gram	EA	$\mathbf{0.613}_{.000}$	$\mathbf{7.33}_{.04}$	$\mathbf{11.65}_{.07}$	$\mathbf{0.147}_{.001}$	$\mathbf{0.162}_{.001}$	$\mathbf{0.148}_{.000}$	$\mathbf{0.159}_{.000}$	$\mathbf{0.225}_{.000}$	$\mathbf{0.231}_{.000}$
word	EA	$0.542_{.000}$	$2.00_{.01}$	$4.52_{.02}$	$0.096_{.002}$	$0.128_{.001}$	$0.116_{.000}$	$0.130_{.000}$	$0.202_{.000}$	$0.205_{.000}$
sentence	EA	$0.465_{.000}$	$1.08_{.00}$	$2.74_{.02}$	$0.073_{.001}$	$0.101_{.000}$	$0.100_{.000}$	$0.111_{.000}$	$0.189_{.000}$	$0.177_{.000}$
paragraph	EA	$0.505_{.000}$	$1.57_{.01}$	$2.53_{.01}$	$0.082_{.001}$	$0.134_{.000}$	$0.107_{.000}$	$0.124_{.000}$	$0.195_{.000}$	$0.196_{.000}$
bi-gram	Adv	$0.218_{.073}$	$0.77_{.23}$	$0.85_{.33}$	$0.095_{.006}$	$0.090_{.003}$	$0.101_{.004}$	$0.098_{.003}$	$0.171_{.005}$	$0.166_{.004}$
bi-gram	Adv + Proc	$0.221_{.074}$	$0.77_{.24}$	$0.87_{.32}$	$0.094_{.006}$	$0.091_{.004}$	$0.102_{.004}$	$0.099_{.002}$	$0.171_{.005}$	$0.166_{.004}$
word	Adv	$0.268_{.016}$	$0.65_{.12}$	$0.54_{.12}$	$0.096_{.006}$	$0.091_{.003}$	$0.105_{.004}$	$0.099_{.003}$	$0.176_{.003}$	$0.165_{.004}$
word	Adv + Proc	$\mathbf{0.269}_{.013}$	$0.64_{.11}$	$0.57_{.07}$	$\mathbf{0.098}_{.006}$	$0.092_{.002}$	$\mathbf{0.107}_{.005}$	$0.099_{.003}$	$\mathbf{0.179}_{.003}$	$0.165_{.004}$
sentence	Adv	$0.265_{.010}$	$0.64_{.08}$	$\mathbf{1.07}_{.24}$	$0.095_{.007}$	$0.094_{.002}$	$0.103_{.006}$	$0.100_{.001}$	$0.176_{.006}$	$0.167_{.001}$
sentence	Adv + Proc	$0.266_{.012}$	$0.68_{.10}$	$1.07_{.21}$	$0.096_{.005}$	$0.094_{.004}$	$0.105_{.006}$	$0.100_{.002}$	$0.178_{.005}$	$0.166_{.002}$
paragraph	Adv	$0.045_{.136}$	$0.69_{.03}$	$0.70_{.04}$	$0.062_{.025}$	$\mathbf{0.123}_{.029}$	$0.082_{.015}$	$\mathbf{0.118}_{.017}$	$0.163_{.013}$	$\mathbf{0.169}_{.003}$
paragraph	Adv + Proc	$0.225_{.061}$	$\mathbf{1.15}_{.60}$	$0.77_{.21}$	$0.093_{.057}$	$0.092_{.011}$	$0.090_{.034}$	$0.103_{.008}$	$0.163_{.023}$	$0.166_{.005}$

Table 1: Comparison among supervised (upper) and unsupervised (lower) methods. Subscripts show the half width of 95% confidence intervals. **Bold** denotes the best performance in each group. *Chance* is the expected value if we randomly yield retrievals. Higher is better for all metrics.

Methods for self-supervised Visual representation learning

Papers	Researchers	Conference
CPC-v2: Data-efficient image recognition with contrastive predictive coding	DeepMind	ICML2020
Learning Representations by Maximizing Mutual Information Across Views	Microsoft Research	NeurIPS2019
MoCo: Momentum contrast for unsupervised visual representation learning	Facebook AI Research	CVPR2020
PIRL: Self-supervised learning of pretext-invariant representations	Facebook AI Research	CVPR2020
SimCLR: A Simple Framework for Contrastive Learning of Visual Representations	Google Research	ICML2020
Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning	Deepmind	Under review
Unsupervised Learning of Visual Features by Contrasting Cluster Assignments	Facebook AI Research	Under review

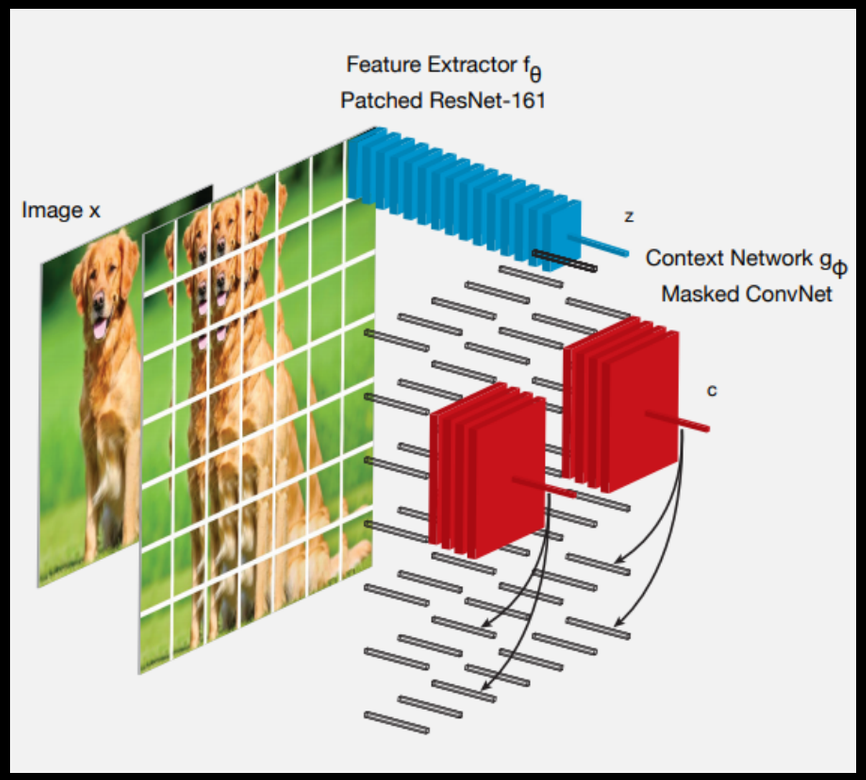
- CPC-v2: Data-efficient image recognition with contrastive predictive coding



Overview of Contrastive Predictive Coding

Contrastive Predictive Coding(CPC)

- CPC is a general technique that observations be ordered along e.g. temporal or spatial dimensions
    - Applied to a variety of different modalities (e.g. speech, natural language and images)
- CPC learns representations by training neural network to predict the representations of future observations from those of past ones
- These predictions are evaluated using a contrastive loss, in which the network must correctly classify ‘future’ representations among a set of unrelated ‘negative’ representations



Overview of unsupervised pre-training with the spatial prediction task

- First, Each input image is divided into a grid of overlapping patches $x_{i,j}$,

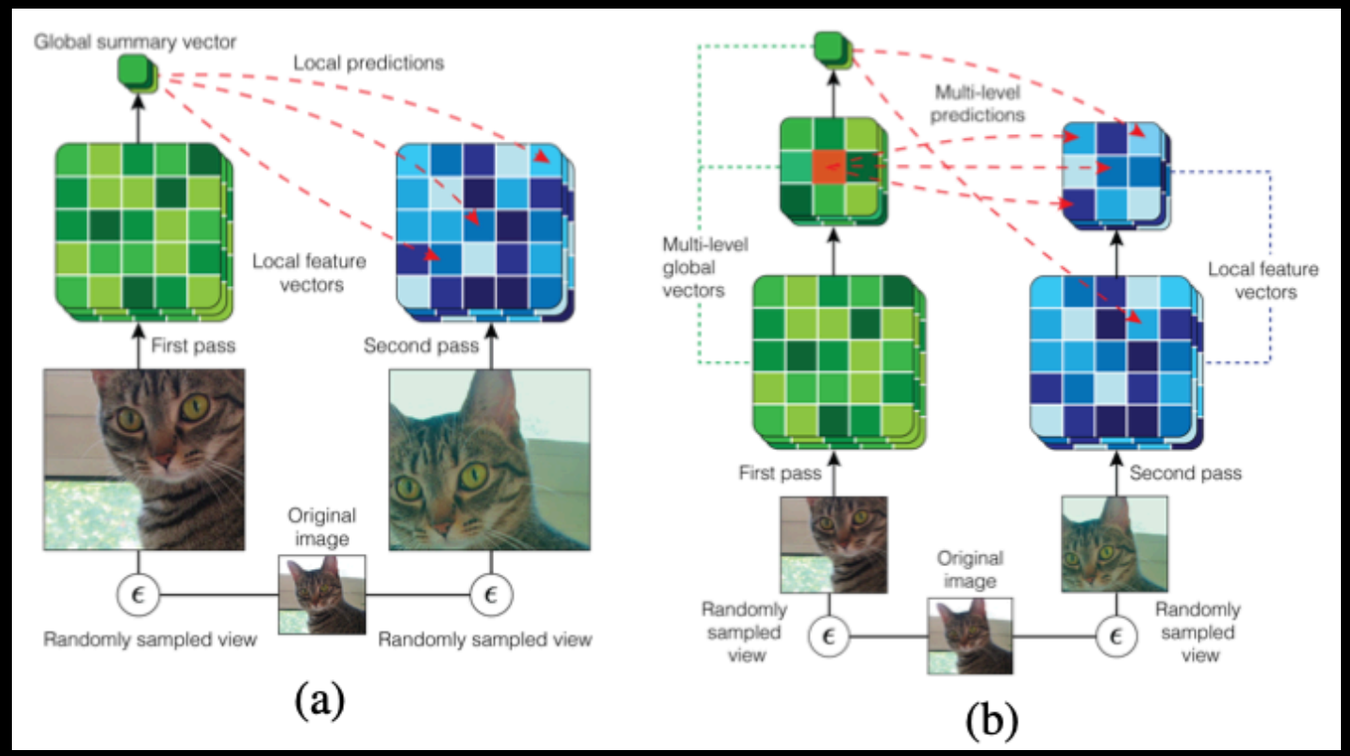
  where i, j denote the location of the patch

- Each patch is encoded with a neural network f_θ into a single vector $z_{i,j} = f_\theta(x_{i,j})$
- A masked convolutional network g_φ is applied to the grid of feature vectors
- The masks are such that the receptive field of each resulting context vector $c_{i,j}$ only includes feature vectors that lie above it in the image,

  i.e. $c_{i,j} = g_\varphi(\,(z_{u,v})_{u,<,i,v}\,)$
- The prediction task consists of predicting 'future' feature vectors $z_{i,+,k,j}$ from current context vector $c_{i,j}$ where $k > 0$

- The predictions are made linearly:
  - Given a context vector $c_{i,j}$, a prediction length $k > 0$
  - a prediction matrix W_k
  - the prediction feature vector is $z_{i,+,k,j} = W_k c_{i,j}$
- Then, the quality of this prediction is evaluated using a contrastive loss
- The goal is correctly recognize the target $z_{i,+,k,j}$ among a set of randomly sampled feature vectors $\{z_l\}$
  - The negative samples $\{z_l\}$ are taken from other locations in the image and other images in the mini-batch

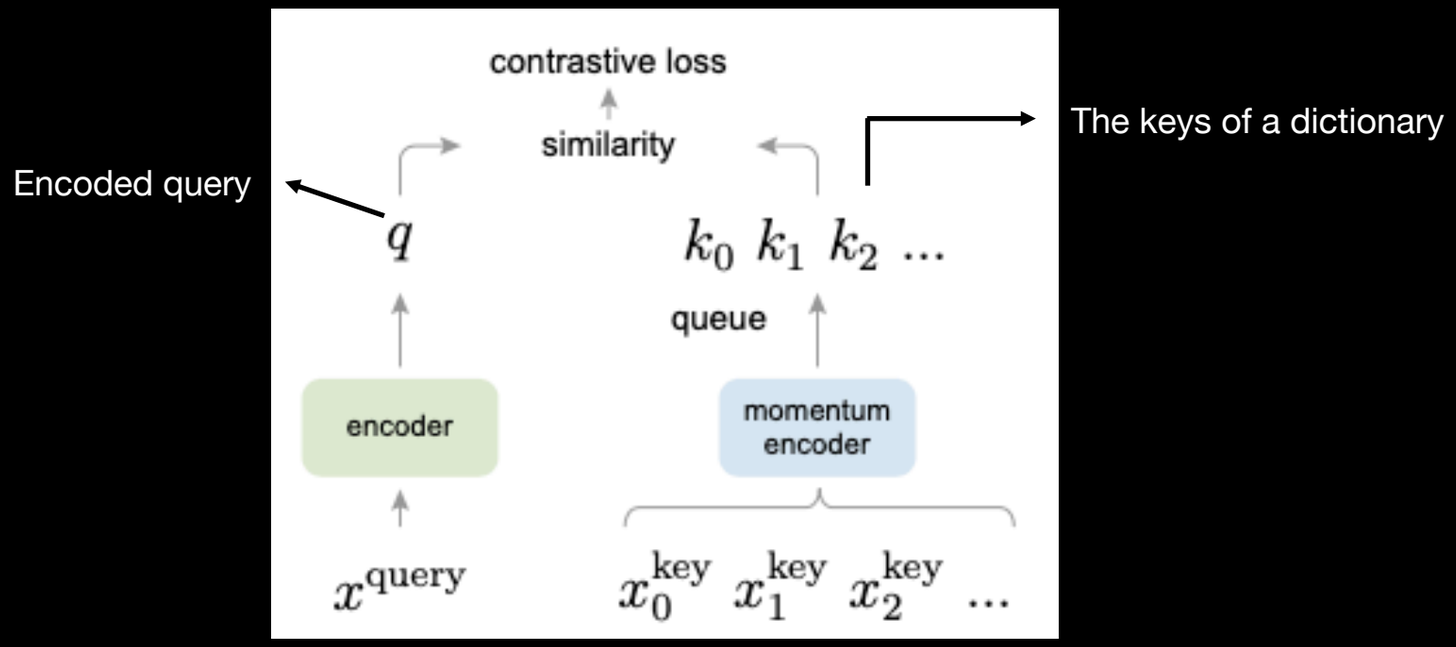
- Learning Representations by Maximizing Mutual Information Across Views



(a) Local Deep InfoMax(Local DIM) with predictions across view generated by data augmentation
(b) Augmented Multiscale DIM(AMDIM) with multi scale infomax across views generated by data augmentation

- Maximizing mutual information between features extracted from multiple views of a shared context
    - The shared context could be an image from the ImageNet training set
    - Multiple views of the context could be produced by applying data augmentation to the image
- Augmented Multi-scale Deep InfoMax (AMDIM)
    - AMDIM extends Local Deep InfoMax (Local DIM)
        - Local DIM maximizes mutual information
          between a global summary feature vector, which depends on the full input,
          and a collection of local feature vectors pulled from an intermediate layer in the encoder
    - First, we maximize mutual information between features extracted
      from independently-augmented copies of each image
    - Second, maximize mutual information between multiple feature scales simultaneously
    - Third, use a more powerful encoder architecture

- MoCo: Momentum contrast for unsupervised visual representation learning



Overview of Momentum contrast (MoCo)

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)}$$

- Momentum Contrast (MoCo) trains a visual representation encoder

  by matching an encoded query q to a dictionary of encoded keys k_i using a contrastive loss

    - The dictionary keys $\{k_0, k_1, k_2, \ldots\}$ are defined on-the-fly by a set of data samples
- Momentum encoder
    - Dynamic dictionary with a queue and a moving-averaged encoder
        - Large and consistent dictionaries

- Build dictionaries that are (1) **large** and (2) **consistent** as they evolve during training
    (1) **Large**
        - Maintain the dictionary as a queue of data samples:
        - The queue decouples the dictionary size from the mini-batch size, allowing it to be large
    (2) **Consistent**
        - As the dictionary keys come from the preceding several mini-batches,
        - A slowly progressing key encoder is proposed to maintain consistency
        - It is implemented as a momentum-based moving average of the query encoder

Momentum update. Using a queue can make the dictionary large, but it also makes it intractable to update the key encoder by back-propagation (the gradient should propagate to all samples in the queue). A naïve solution is to copy the key encoder f_{k} from the query encoder f_{q}, ignoring this gradient. But this solution yields poor results in experiments (Sec. 4.1). We hypothesize that such failure is caused by the rapidly changing encoder that reduces the key representations' consistency. We propose a momentum update to address this issue.

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}$$

- θ_k: the parameters of f_k and those of f_q as θ_q
- $m \in [0,1)$: a momentum coefficient

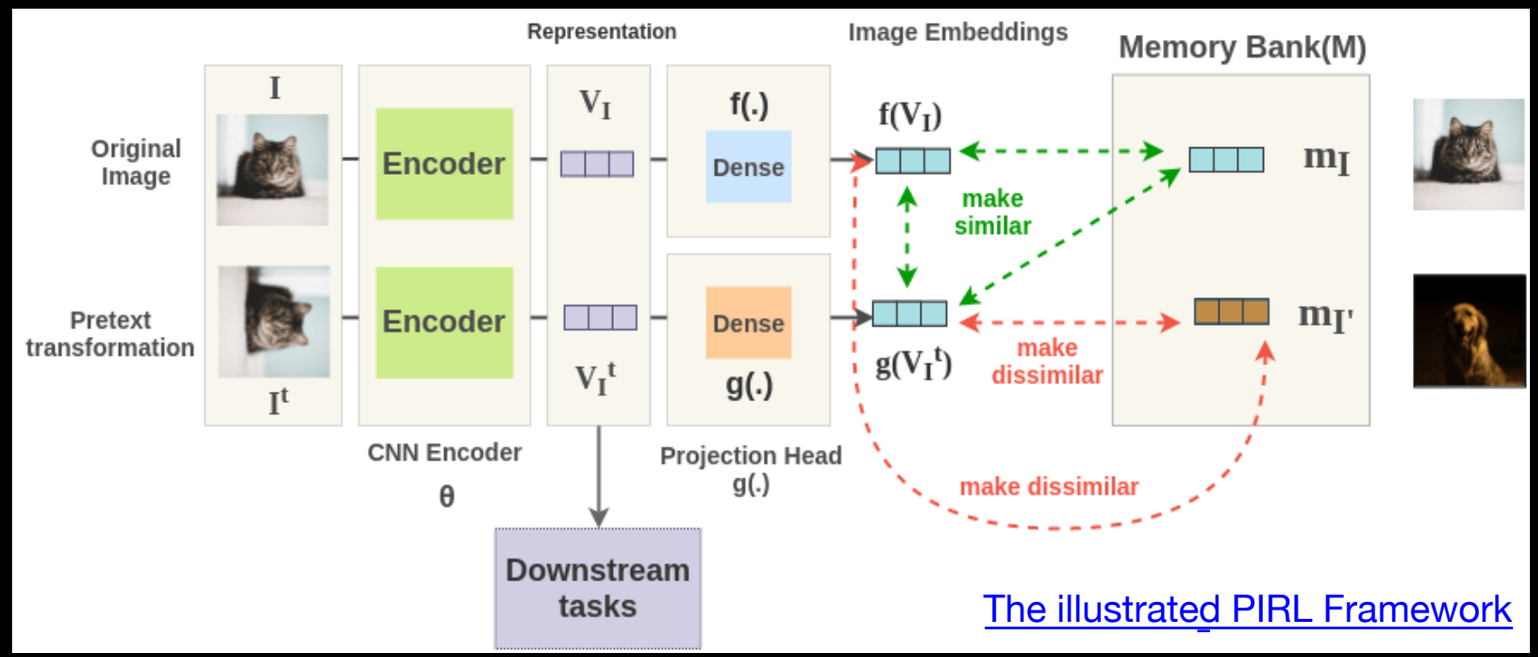
- Only the parameters θ_q are updated by back-propagation
- The momentum update makes θ_k evolve more smoothly

Ablation: momentum. The table below shows ResNet-50 accuracy with different MoCo momentum values (m in Eqn.(2)) used in pre-training ($K = 4096$ here) :

momentum m	0	0.9	0.99	0.999	0.9999
accuracy (%)	*fail*	55.2	57.8	59.0	58.9

It performs reasonably well when m is in 0.99 ~ 0.9999, showing that a slowly progressing (*i.e.*, relatively large momentum) key encoder is beneficial. When m is too small (*e.g.*, 0.9), the accuracy drops considerably; at the extreme of *no momentum* (m is 0), the training loss oscillates and fails to converge. These results support our motivation of building a consistent dictionary.

- PIRL: Self-supervised learning of pretext-invariant representations

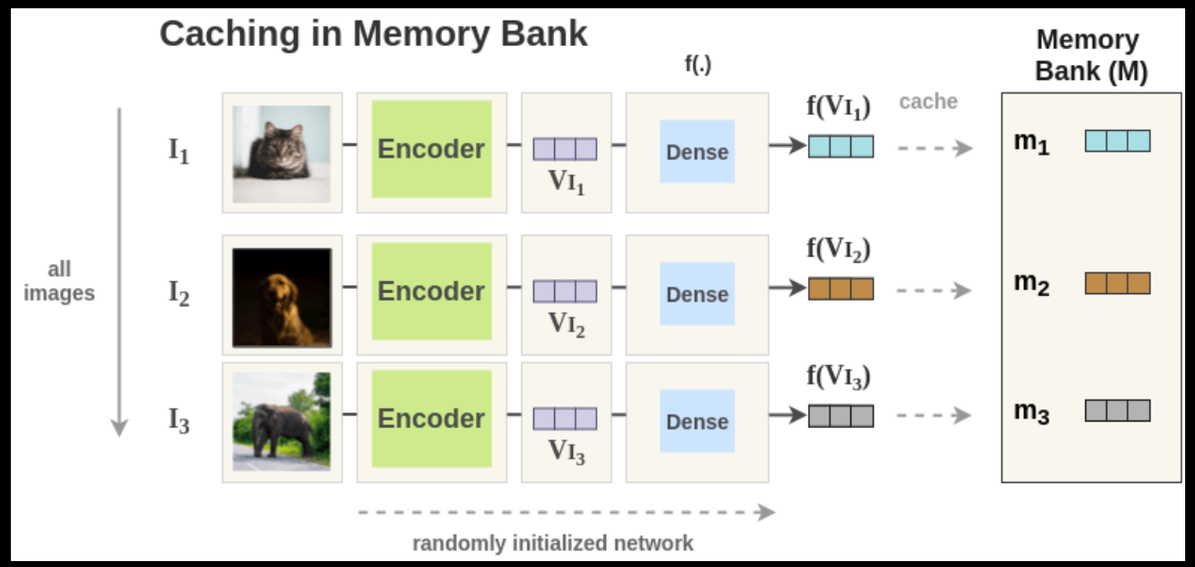


The illustrated PIRL Framework

Overview of PIRL

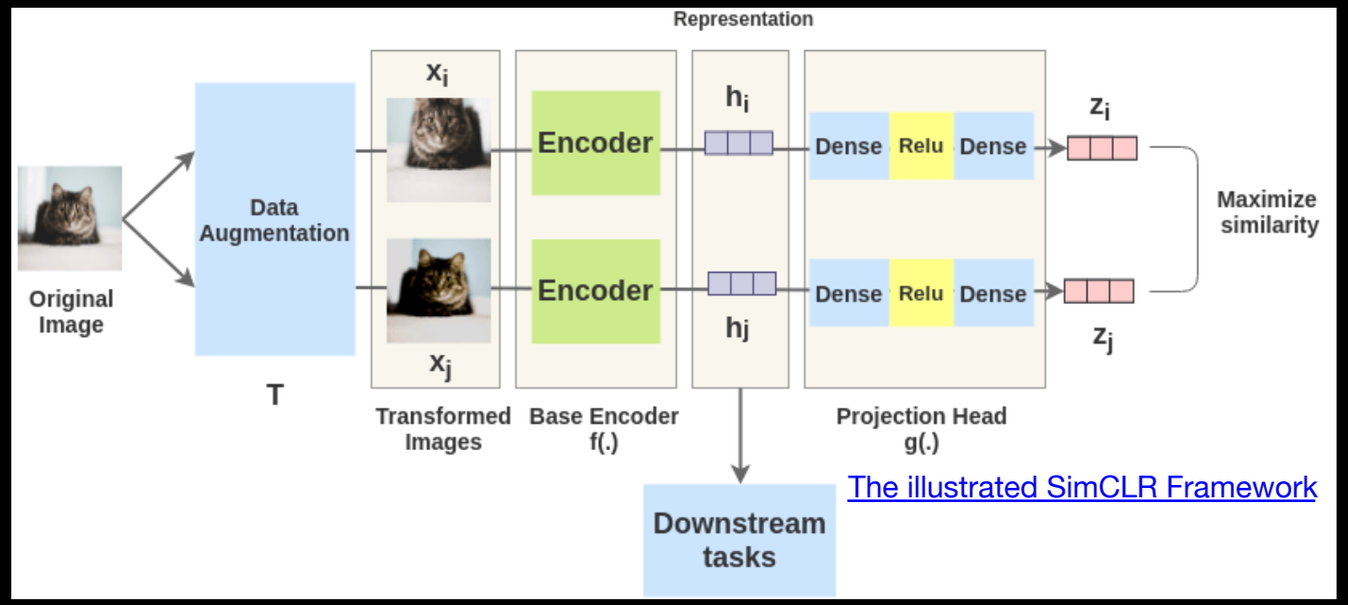
- Pretext-Invariant Representation Learning(PIRL) that learns invariant representations based on pretext tasks
    - With memory bank, M
- Make the transformation of an image *similar* to the original image
- Make representations of the original and transformed image *different* from other random images in the dataset

Using a Memory Bank of Negative Samples



- Using a memory bank M which caches representations m_I of all images and use that during training
  - The representation m_I is an exponential moving average of feature representations $f(v_I)$ that were computed in prior epochs
  - This allows us to use a large number of negative pairs without increasing batch size

- SimCLR: A Simple Framework for Contrastive Learning of Visual Representations



The illustrated SimCLR Framework

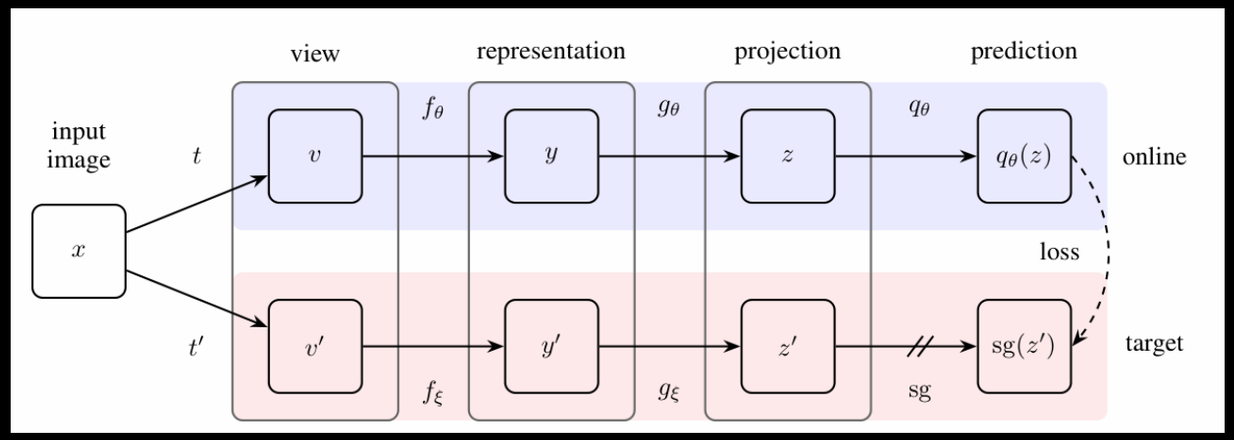
Overview of SimCLR

- Without requiring specialized architectures or a memory bank

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of $f, g, \mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}, t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i, j)$ **as** $\ell(i, j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

- Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning



BYOL's architecture

- BYOL uses two neural networks, referred to as online and target networks,
  that interact and learn from each other
- Trains its online network to predict the target network's prediction
- Target network provides the regression targets(feature vector) to train the online network
  - Its parameters ξ are an exponential moving average of the online parameters θ

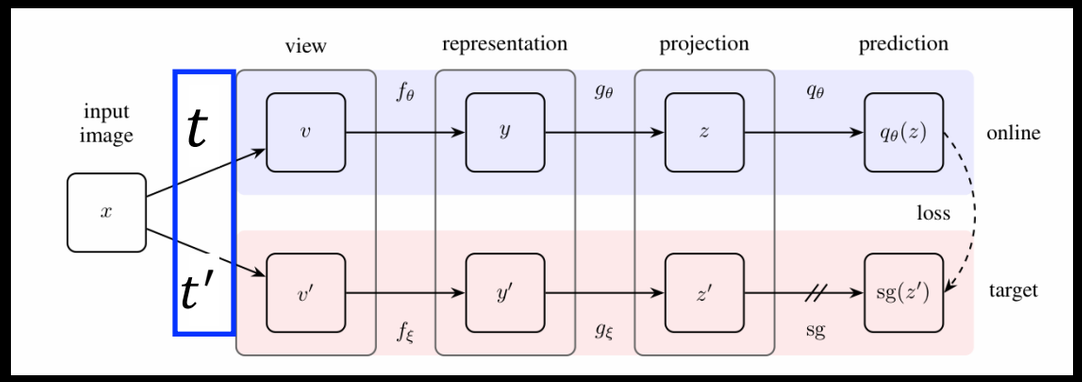
$$\xi \leftarrow \tau\xi + (1-\tau)\theta.$$

$$\tau \triangleq 1 - (1-\tau_{\text{base}}) \cdot (\cos(\pi k/K)+1)/2$$

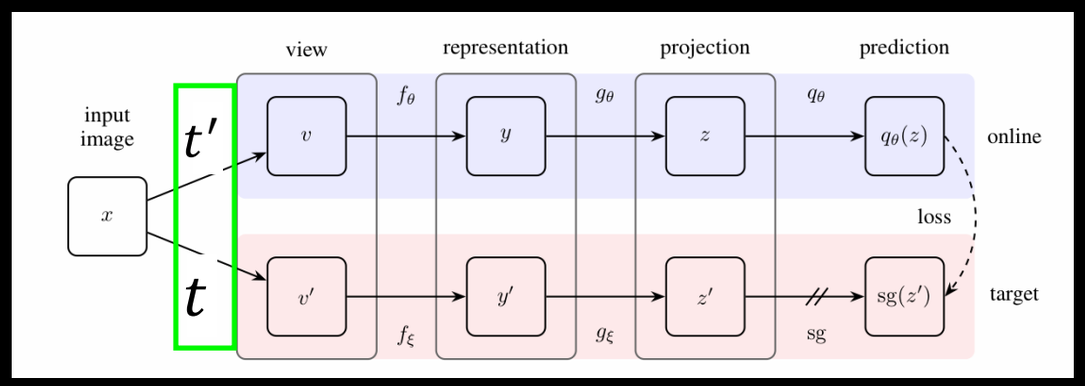
$$\tau_{\text{base}} = 0.996$$

k: the current training step
K: the maximum number of training steps

Loss function



$$\mathcal{L}_\theta^{\mathtt{BYOL}}$$



$$\widetilde{\mathcal{L}}_\theta^{\mathtt{BYOL}}$$

Final loss

$$\mathcal{L}_\theta^{\mathtt{BYOL}} + \widetilde{\mathcal{L}}_\theta^{\mathtt{BYOL}}$$

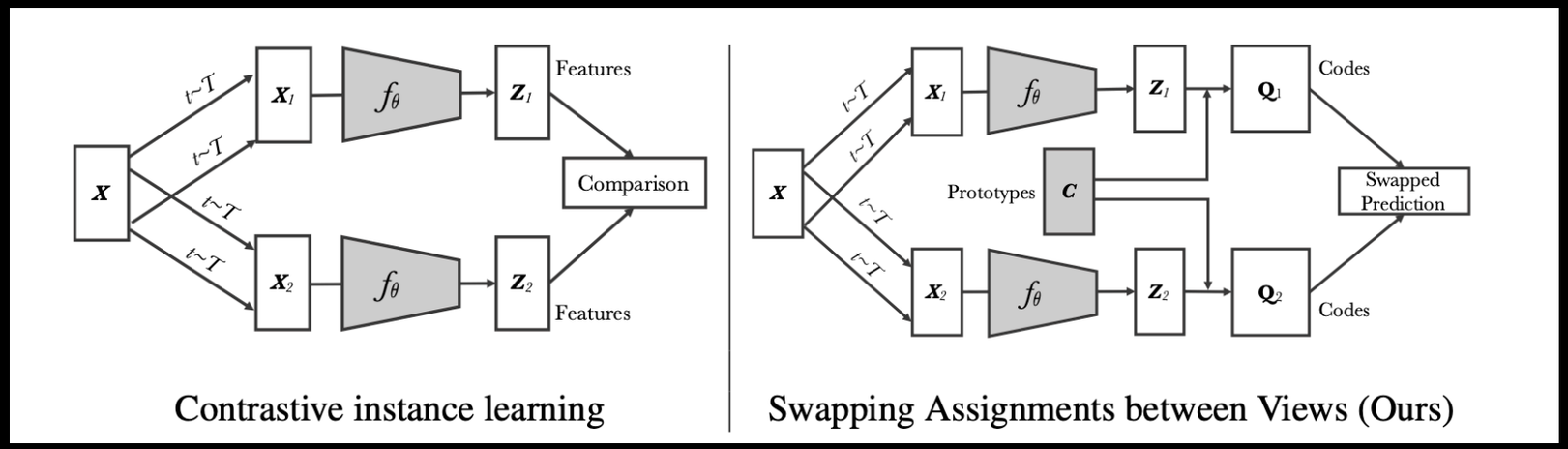
$$\overline{q_\theta}(z_\theta) \triangleq q_\theta(z_\theta)/\|q_\theta(z_\theta)\|_2$$

$$\overline{z}'_\xi \triangleq z'_\xi/\|z'_\xi\|_2$$

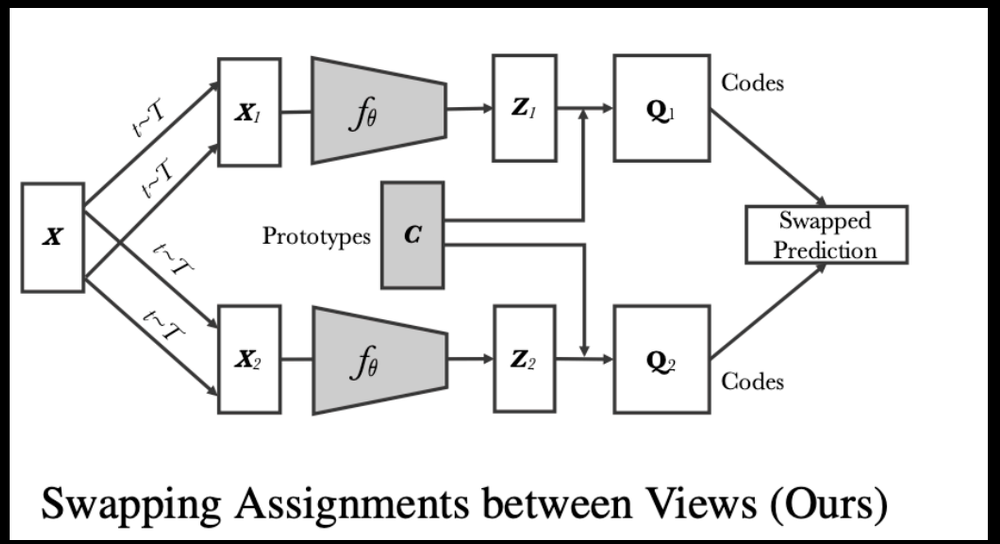
$$\mathcal{L}_\theta^{\mathtt{BYOL}} \triangleq \left\|\overline{q_\theta}(z_\theta) - \overline{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi\rangle}{\left\|q_\theta(z_\theta)\right\|_2 \cdot \left\|z'_\xi\right\|_2}.$$

Mean squared error with l_2 normalize

- Unsupervised Learning of Visual Features by Contrasting Cluster Assignments



- In contrastive learning methods
  - The features from different transformations of the same images are compared directly to each other
  - Contrastive methods work online and rely on a large number of explicit pairwise feature comparison, which is computationally challenging
- In Swapping Assignments between multiple Views of the same image (SwAV)
  - Propose an online cluster-based self-supervised method
    - “Swapped” prediction mechanism
    - Multi-crop
      - Uses a mix of views with different resolutions in place of two full-resolution views



Swapping Assignments between Views (Ours)

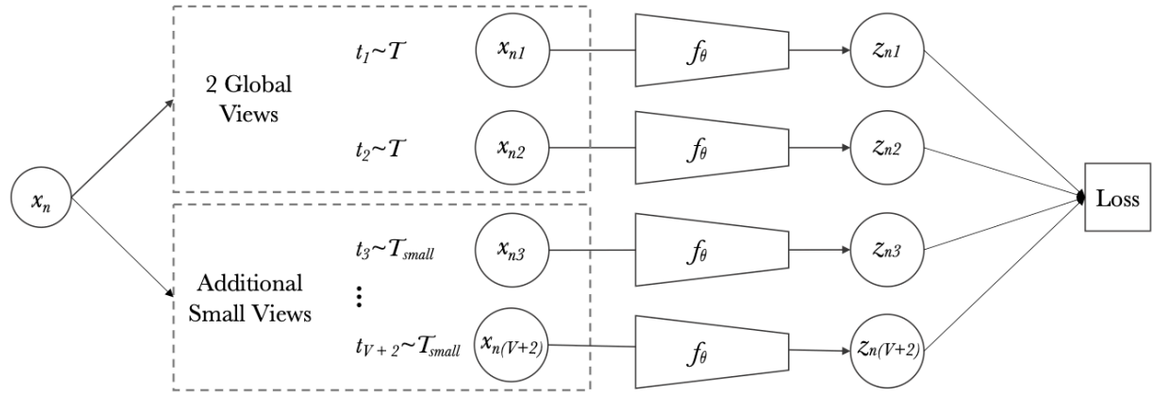
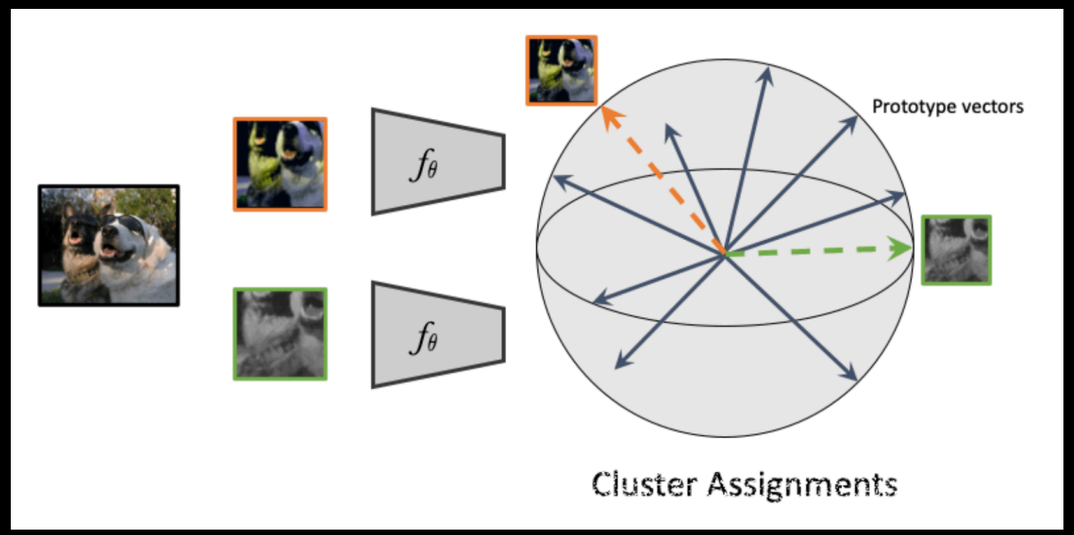
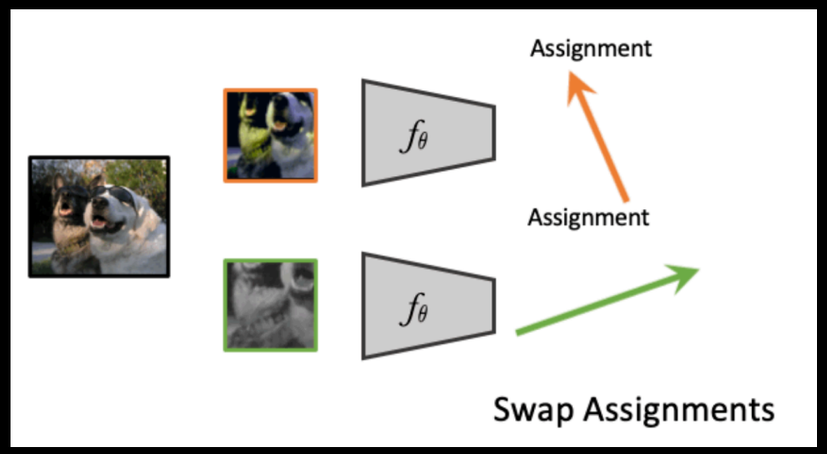


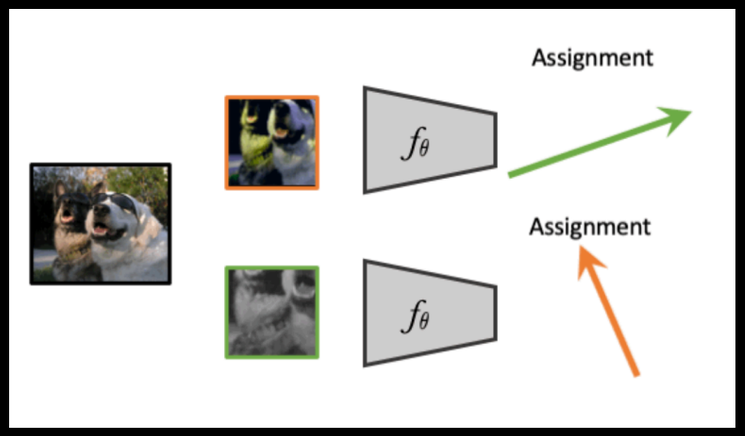
Figure 5: **Multi-crop**: the image x_n is transformed into $V + 2$ views: two global views and V small resolution zoomed views.

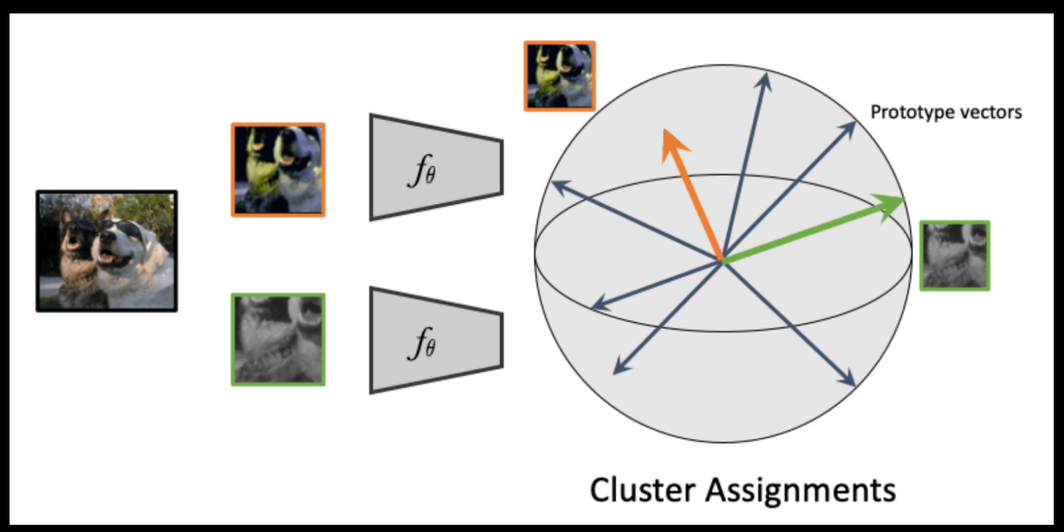
- Given two image features z_t, z_s from two different augmentation of the same image
  - We compute their codes q_t, q_s by matching these features to a set of K prototypes $\{c_1, \ldots, c_k\}$
  - Predict this code from other augmented versions of the same image
- Thus, SwAV does not directly compare image features
- Prototype vectors are learned along with the ConvNet parameters by back propagation

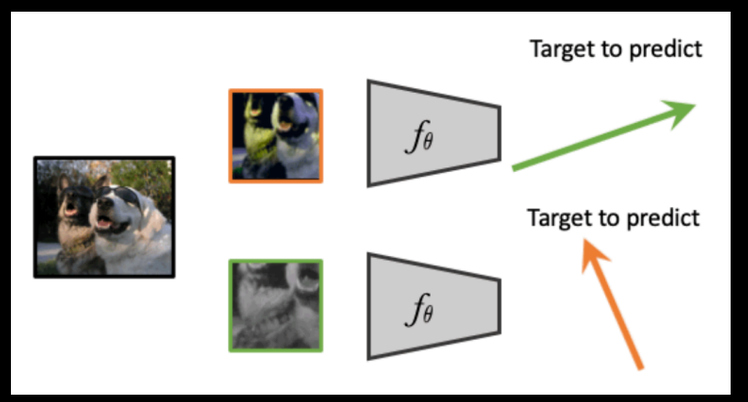
Process of SwAV











Our Strategy - Objective

1. Self-supervised image-text embedding model.

2. Image classification and generating radiology text.

3. Classification result AUC.

4. Highlight the ROI by heatmap.

Our Strategy - Method

Multi-modality (Free text & Image)

1> Radiology report generation (Image captioning)

- CNN + RNN model
- Automatic Generation of Medical Reports
- Evaluate NLG result with other NLP Tool to measure accuracy, precision, F1 score.

2> Lesion detection and annotation

- CNN + RNN model
- Lesion detection with highlight the ROI by heatmap
- AUC result and IOU result with NIH-dataset

3> 1 + 2

- CNN + RNN model
- Report Generation
- Lesion and radiology activation map visualization
- AUC result and Model evaluation with labeler NLP tools

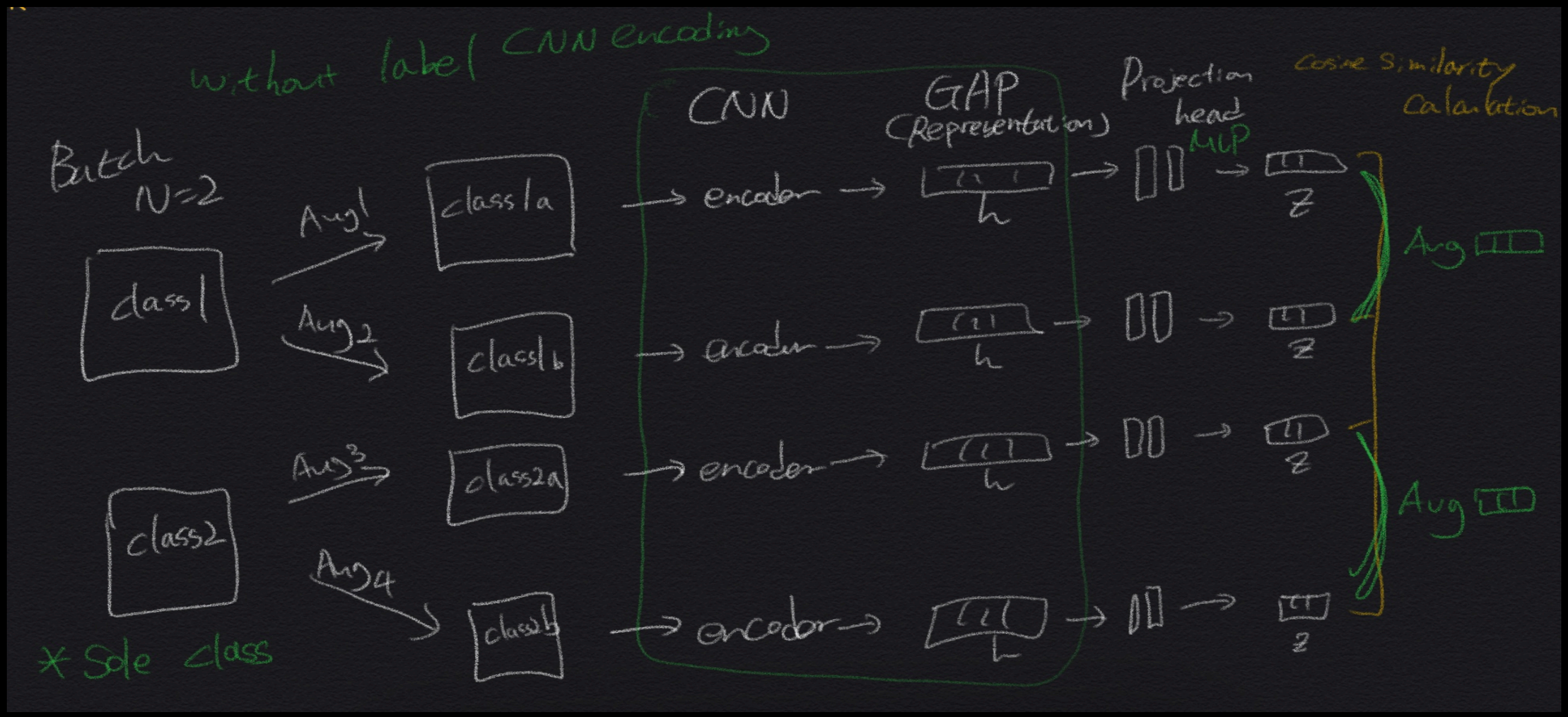
Self-supervision

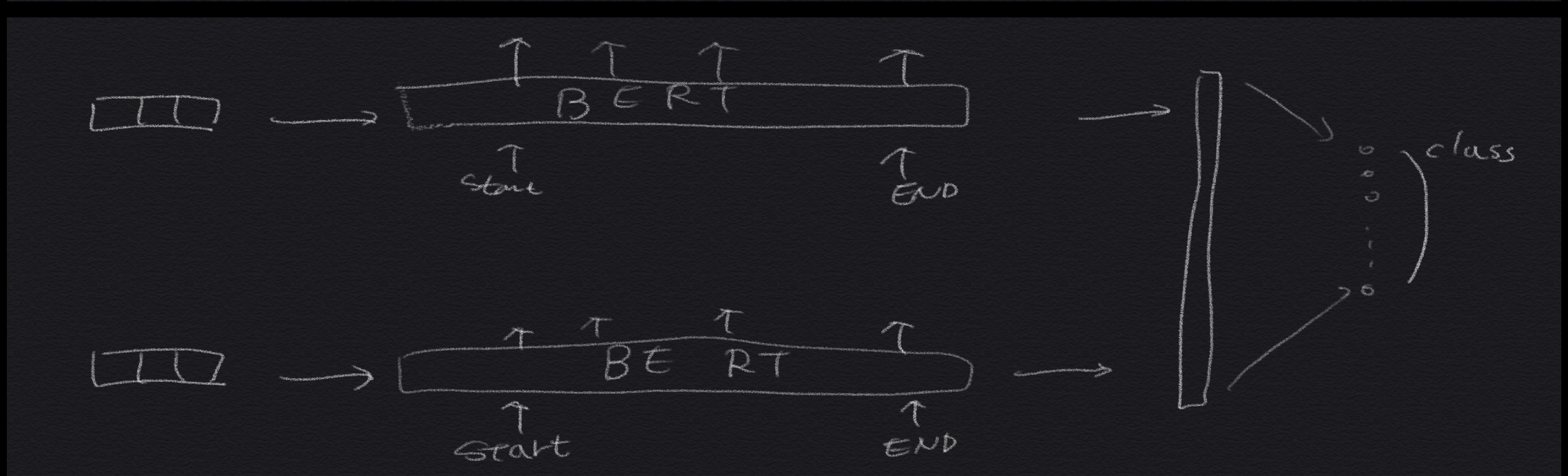
1> Contrastive learning for Image representation

& BERT for Natural Language

2> BERT based self-supervision (Image & Natural Language)

Our Strategy – scenario 1





Plan for CVPR 2021.

Paper submission due date : Nov 16 (15 weeks left)

1. Related work survey & Detailed Novel method suggestion (2-3 weeks)

2. Setup the Development Environment & EDA (1-2 weeks)

4. Modelling (6-8 weeks)

5. Paper (4-6 weeks)

Plan for next meeting

1. Detailed Method Suggestion

 How attention?

 How self-supervised for NLP? -> paper review

2. Development Environment Setting (Kubernetes)