Competence-based Multimodal Curriculum Learning for Medical Report Generation

Fenglin Liu[1*], Shen Ge[2], Xian Wu[2]

[1] Peking University

[2] Tencent Medical AI Lab, Beijing, China

北京大学 PEKING UNIVERSITY

Contents

1. Introduction

Medical Report Generation

- **Task Definition**: It aims to generate a long paragraph describing both the normal and abnormal regions, which can assist radiologists in clinical decision-making.
- **Task Objectives**:
  - a long and coherent report.
  - cover key medical findings:
    - ✓ e.g., heart size and lung opacity.
  - correctly describe any abnormalities and its details:
    - ✓ e.g., the location and shape of the abnormality.
  - correctly describe potential diseases:
    - ✓ e.g., effusion and consolidation.

Indication: No acute cardiopulmonary abnormality.

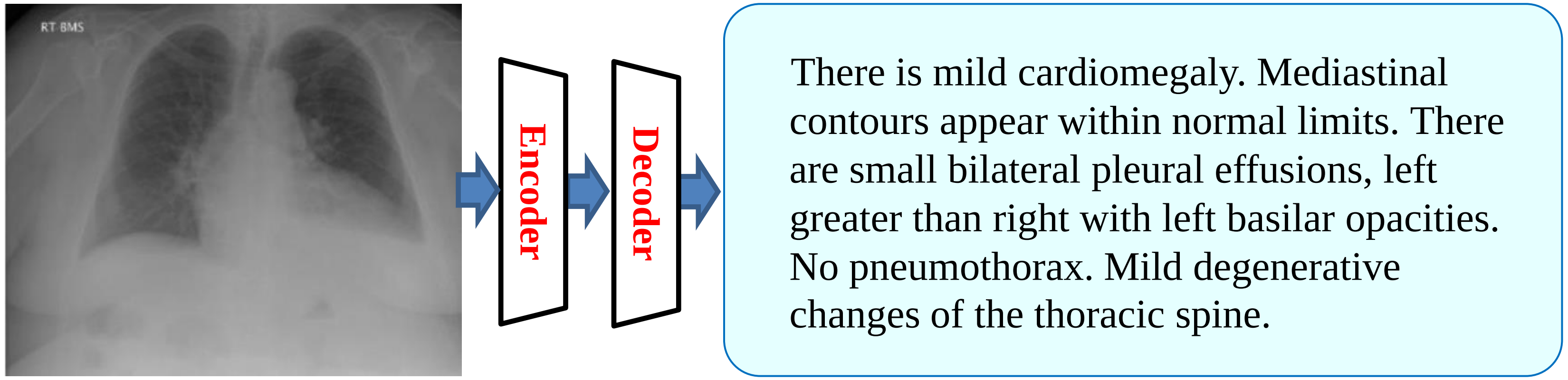
Findings: Lungs are clear without focal infiltrates. Calcified right upper lobe granuloma unchanged from prior. No pneumothorax or pleural effusion. Normal heart size. Normal pulmonary vascularity. Bony thorax intact.

Impression: No acute cardiopulmonary abnormality.

Tags: Calcified Granuloma

Medical Report Generation

- **Urgent goal and core value:** correctly **capturing** and **depicting** the abnormalities.
- **Dataset**: (I, S), where I and $S = \{s_1, s_2, \ldots, s_T\}$ represent the input medical image and the target report, respectively.
- **Encoder-Decoder Framework**: In the encoding stage, the visual representation V are extracted by an image encoder; In the decoding stage, the report is generated using RNN/Transformer.
- **Training Objective**: The widely-used training objective is to minimize the cross entropy loss.

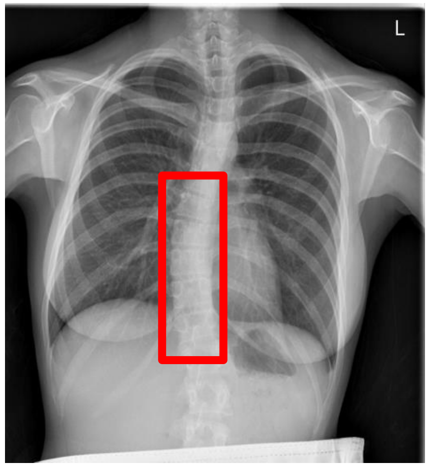


Visual Encoder: $I \rightarrow V$; Target Decoder: $V \rightarrow S$.

$$L_{CE}(\theta) = -\sum_{t=1}^{T} \log\left(p_\theta\left(s_t^* | s_{1:t-1}^*; I\right)\right)$$

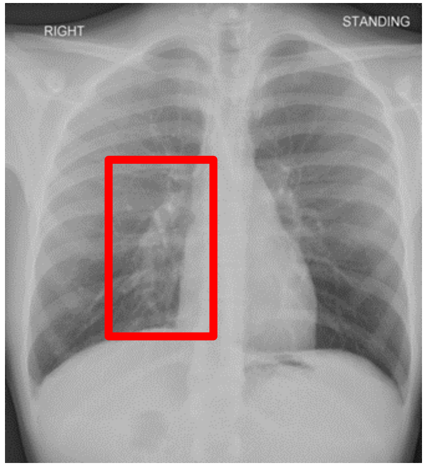
Motivations

- **Urgent goal and core value**: correctly **capturing** and **depicting** the abnormalities.
- **Problems**:
  - **Visual Data Bias**:
    - ✓ The normal images dominate the dataset over the abnormal ones, especially for the rare diseases [1].
    - ✓ The abnormal regions (red bounding box) only occupy a small part of the entire image.
  - **Textual Data Bias**:
    - ✓ The abnormal description (red colored text) only occupy a small part of the entire report.
    - ✓ There are many similar sentences (blue colored text) used in each report to describe the normal regions
  - **Limited Medical Data:**
    - 4K samples (IU X-ray) << 14M samples (ImageNet) / 3.3M samples (Captioning)



Medical Report:

Lungs are clear. No pleural effusions or pneumothoraces. Heart and mediastinum of normal size and contour. [1]scoliosis.

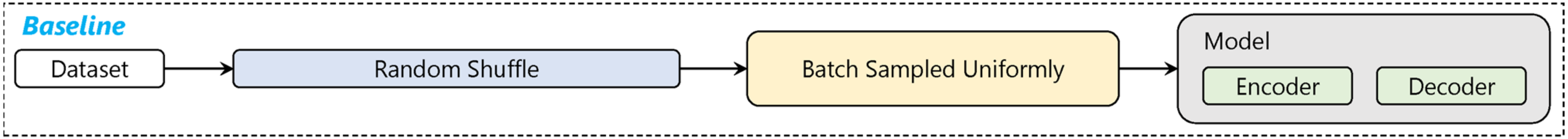


Medical Report:

The heart and mediastinum are normal. The lungs are clear. [1]There is mild blunting of the right costophrenic XXXX. There is no infiltrate, mass or pneumothorax.

[1] Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In CVPR, 2016.

Motivations



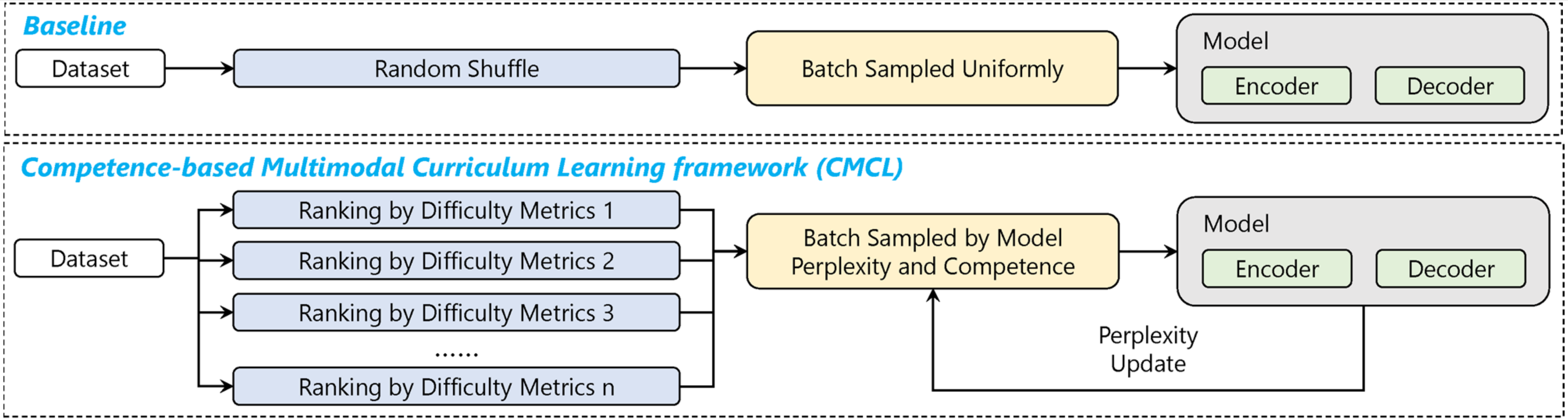
- During training, most existing works treat all the training samples equally without considering their difficulties:
  - All training samples from the limited medical data are randomly shuffled and grouped into batches for training.

- As a result, due to the visual and textual data biases could mislead the model training, existing data-driven neural models are biased towards generating plausible but general reports without prominent abnormal narratives.

2. Competence-based Multimodal Curriculum Learning (CMCL)

北京大學
PEKING UNIVERSITY

Framework

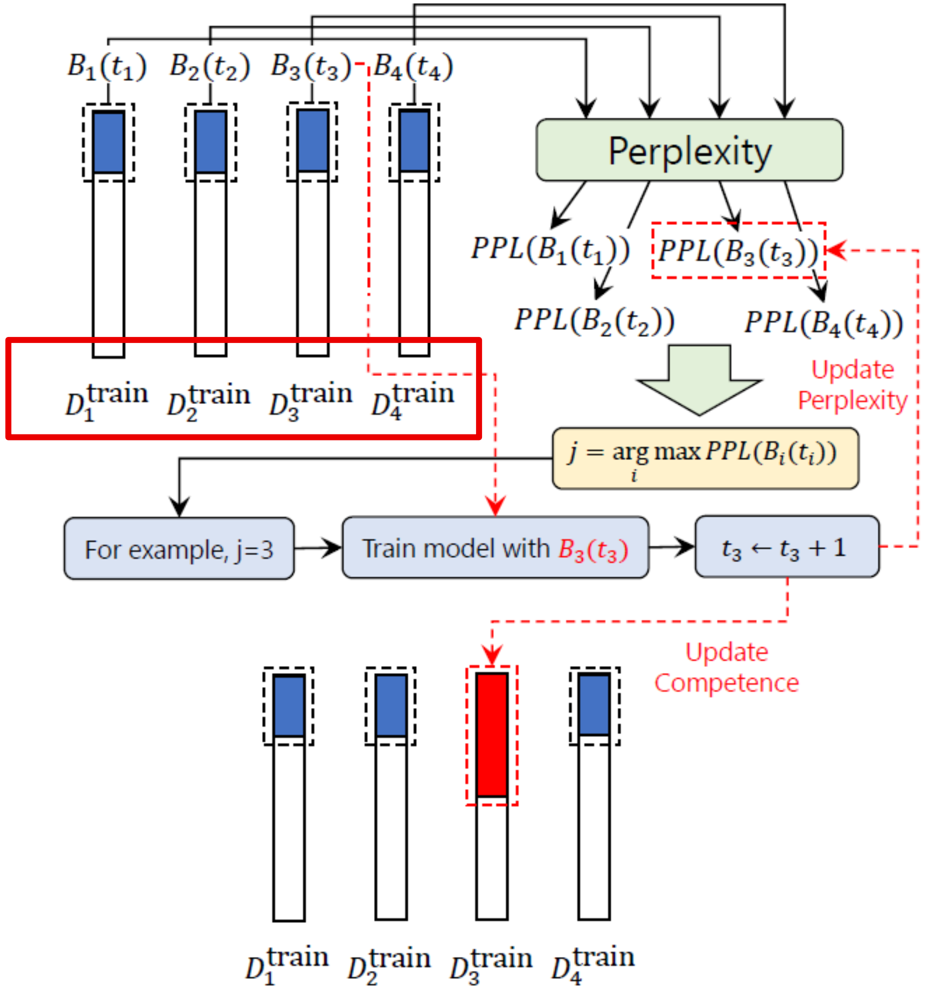


- CMCL progressively learns medical reports following an easy-to-hard fashion, helping existing models better utilize the limited medical data to alleviate the visual and textual data biases.
- Such process is similar to the learning curve of radiologists:
  - (1) first start from simple and easy-written reports;
  - (2) then attempt to consume harder reports, which include rare and diverse abnormalities.

PEKING UNIVERSITY

Algorithm

CMCL first assesses the difficulty of each training sample from multiple perspectives (i.e., the visual complexity -> visual bias; textual complexity -> textual bias), which include four difficulty metrics.

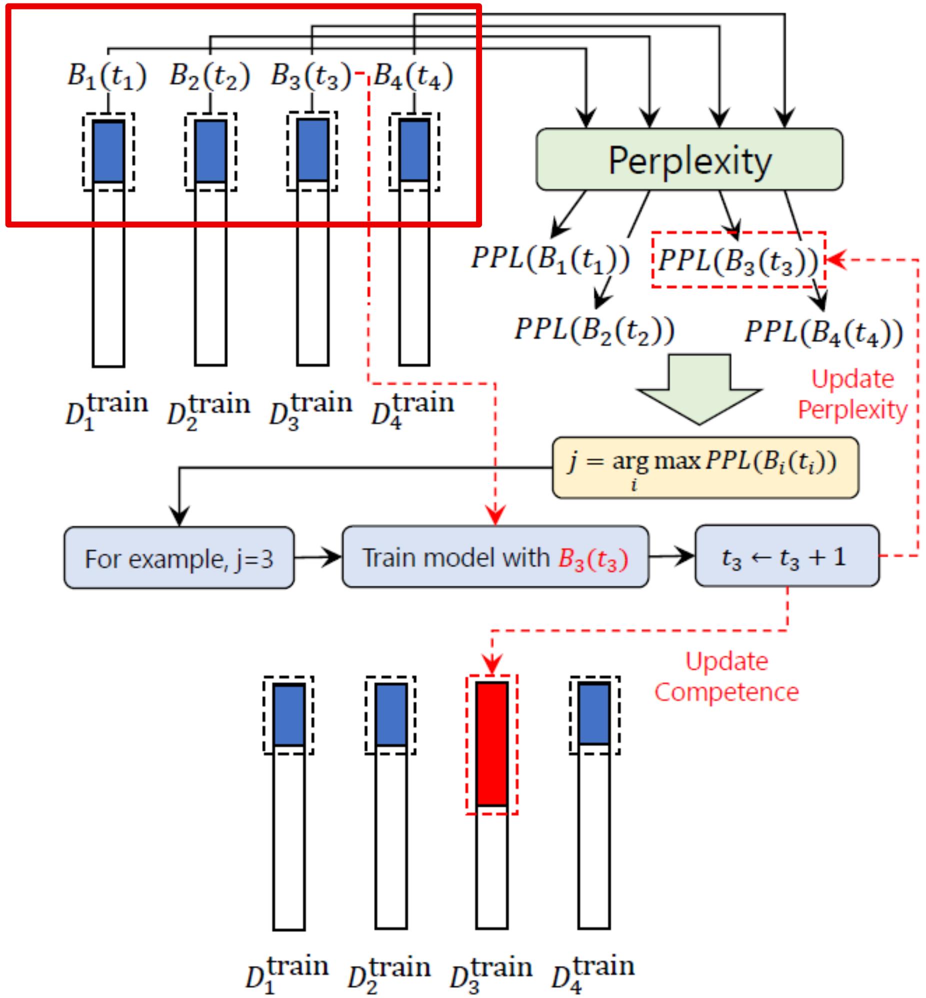


Input: The training set D^{train}, $i \in \{1, 2, 3, 4\}$.
Output: A model with multiple difficulty-based curriculum learning.

Compute four difficulties, d_i, for each training sample in D^{train};
Sort D^{train} based each difficulty of every sample, resulting in D_i^{train} (i.e., D_1^{train}, D_2^{train}, D_3^{train}, D_4^{train});
for $i = 1, 2, 3, 4$ **do**
    $t_i = 0$; Initialize the model competence from i^{th} perspective, $c_i(0)$, by Eq. (2); Uniformly sample a data batch, $B_i(0)$, from the top $c_i(0)$ portions of D_i^{train};
    Compute the perplexity (PPL) on $B_i(0)$, PPL$(B_i(0))$;
end for
repeat
    $j = \arg\max_i(\text{PPL}(B_i(t_i)))$;
    Train the model with the $B_j(t_j)$;
    $t_j \leftarrow t_j + 1$;
    Estimate the model competence from j^{th} perspective, $c_j(t_j)$, by Eq. (2); Uniformly sample a data batch, $B_j(t_j)$, from the top $c_j(t_j)$ portions of D_j^{train};
    Compute the perplexity (PPL) of model on $B_j(t_j)$, PPL$(B_j(t_j))$;
until Model converge.

Algorithm

Then, CMCL generates one batch for each difficulty metric, resulting in four different curricula.

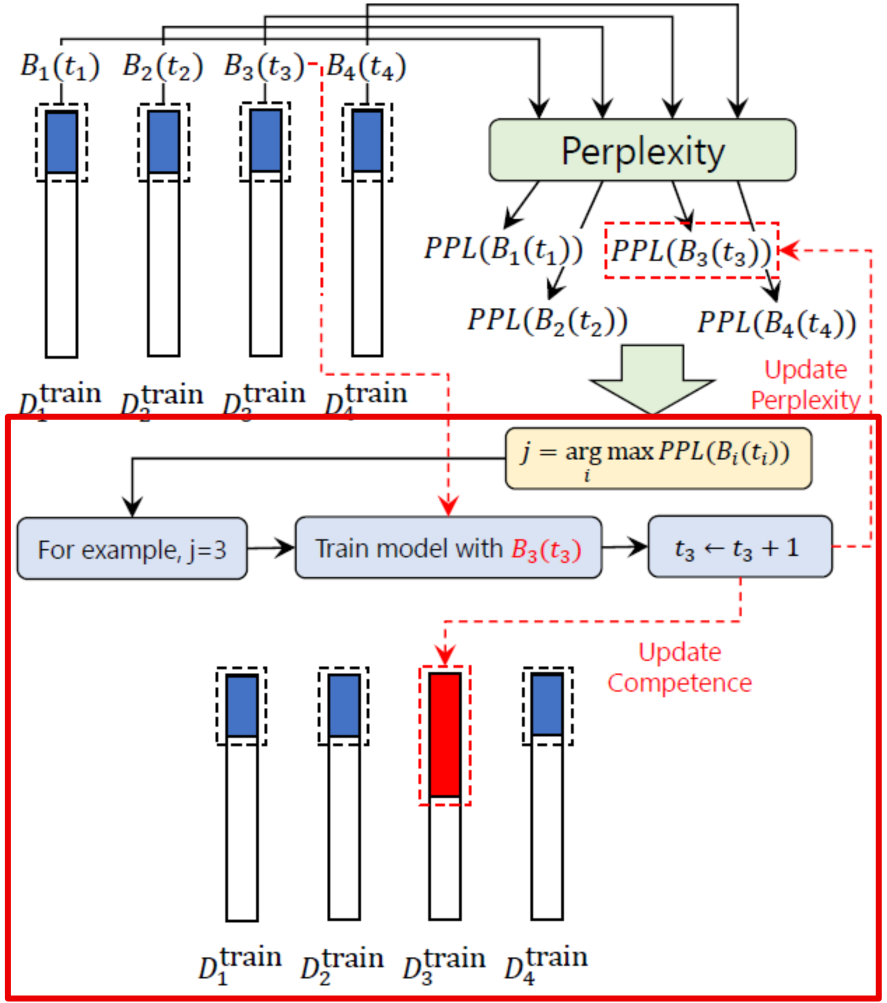


Input: The training set D^{train}, $i \in \{1, 2, 3, 4\}$.
Output: A model with multiple difficulty-based curriculum learning.

Compute four difficulties, d_i, for each training sample in D^{train};
Sort D^{train} based each difficulty of every sample, resulting in D_i^{train} (i.e., D_1^{train}, D_2^{train}, D_3^{train}, D_4^{train});
for $i = 1, 2, 3, 4$ **do**
    $t_i = 0$; Initialize the model competence from i^{th} perspective, $c_i(0)$, by Eq. (2); Uniformly sample a data batch, $B_i(0)$, from the top $c_i(0)$ portions of D_i^{train};
    Compute the perplexity (PPL) on $B_i(0)$, $\text{PPL}(B_i(0))$;
end for
repeat
    $j = \arg\max_i(\text{PPL}(B_i(t_i)))$;
    Train the model with the $B_j(t_j)$;
    $t_j \leftarrow t_j + 1$;
    Estimate the model competence from j^{th} perspective, $c_j(t_j)$, by Eq. (2); Uniformly sample a data batch, $B_j(t_j)$, from the top $c_j(t_j)$ portions of D_j^{train};
    Compute the perplexity (PPL) of model on $B_j(t_j)$, $\text{PPL}(B_j(t_j))$;
until Model converge.

Algorithm

At last, CMCL adaptively selects the most appropriate curricula for each training step according to the current competence of the model.

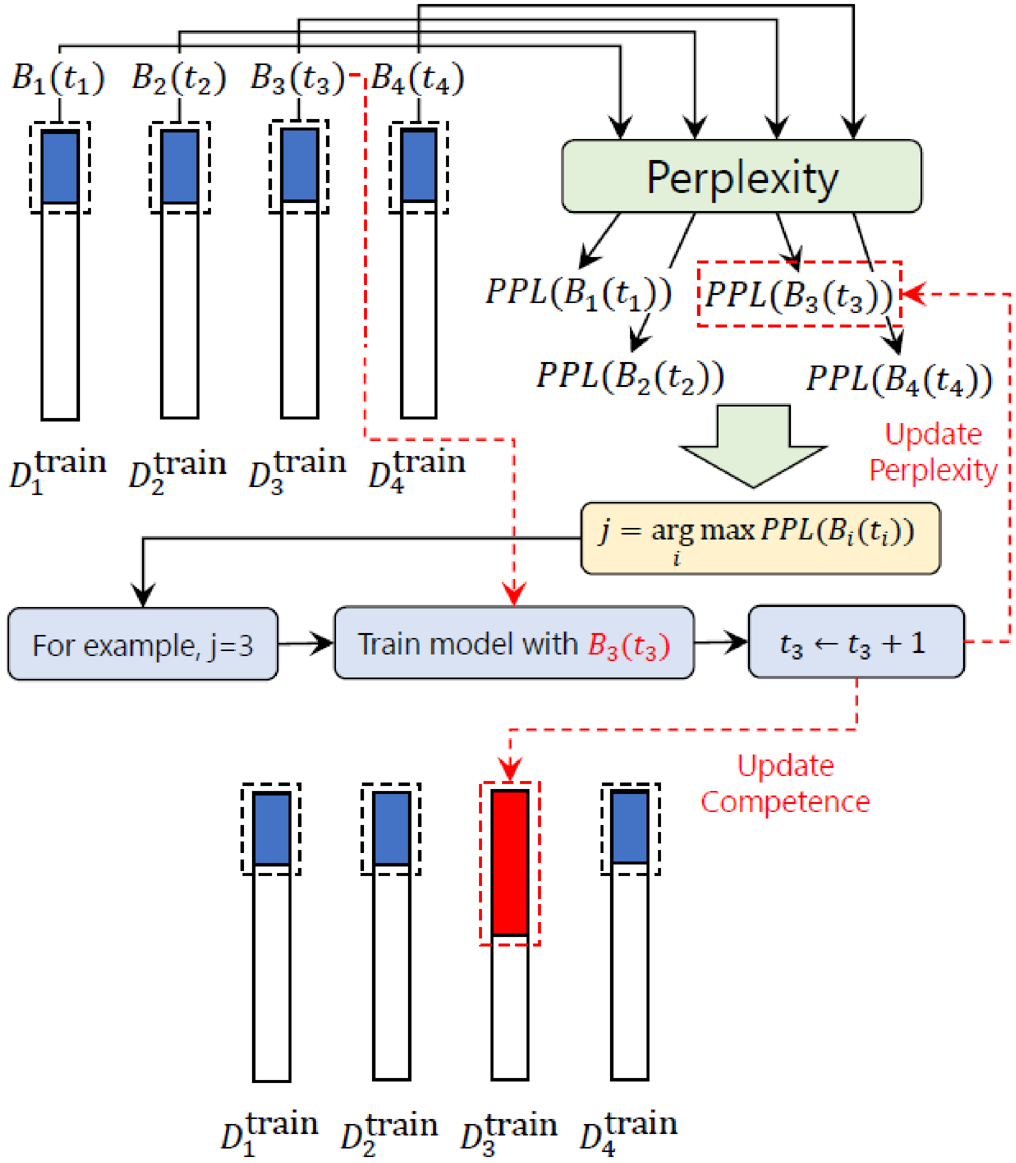


Input: The training set D^{train}, $i \in \{1, 2, 3, 4\}$.
Output: A model with multiple difficulty-based curriculum learning.

Compute four difficulties, d_i, for each training sample in D^{train};
Sort D^{train} based each difficulty of every sample, resulting in D_i^{train} (i.e., D_1^{train}, D_2^{train}, D_3^{train}, D_4^{train});
for $i = 1, 2, 3, 4$ **do**
    $t_i = 0$; Initialize the model competence from i^{th} perspective, $c_i(0)$, by Eq. (2); Uniformly sample a data batch, $B_i(0)$, from the top $c_i(0)$ portions of D_i^{train};
    Compute the perplexity (PPL) on $B_i(0)$, PPL($B_i(0)$);
end for
repeat
    $j = \arg\max_i(\text{PPL}(B_i(t_i)))$;
    Train the model with the $B_j(t_j)$;
    $t_j \leftarrow t_j + 1$;
    Estimate the model competence from j^{th} perspective, $c_j(t_j)$, by Eq. (2); Uniformly sample a data batch, $B_j(t_j)$, from the top $c_j(t_j)$ portions of D_j^{train};
    Compute the perplexity (PPL) of model on $B_j(t_j)$, PPL($B_j(t_j)$);
until Model converge.

PEKING UNIVERSITY

Algorithm



- It can be understood that CMCL sets multiple curricula in parallel, and the model is optimized towards the one with lowest competence at each training step.

- In this way, once the easy and simple samples are well-learned, CMCL increases the chance of learning difficult and complex samples.

Difficulty Metrics

- **Urgent goal and core value**: correctly **capturing** and **depicting** the abnormalities.

- **Visual Difficulty**:
  - The difficulty of accurately capturing the abnormalities.
- **Textual Difficulty**:
  - The difficulty of accurately describing the abnormalities.

Difficulty Metrics: Visual Difficulty

- Heuristic Metric:
  - To measure the visual difficulty, we first extract the normal image embeddings of all normal training images from the ResNet-50 [1]. Then, given an input image, we again use the ResNet-50 to obtain the image embedding. At last, the average cosine similarity between the input image and normal images is adopted as the heuristic metric of visual difficulty.

- Model Confidence:
  - We adopt the ResNet-50 [1] to conduct the abnormality classification task. We acquire the classification probability distribution $P(I) = \{p_1(I), p_2(I), \dots, p_{14}(I)\}$ among the 14 diseases for each image I in the training dataset. Then, we employ the entropy value $H(I)$ of the probability distribution, defined as follows:

$$H(I) = -\sum_{n=1}^{14} \left(p_n(I) \log\left(p_n(I)\right) + \left(1 - p_n(I)\right) \log\left(1 - p_n(I)\right)\right)$$

  - We employ the entropy value $H(I)$ as the model confidence measure, indicating whether an image is easy to be classified or not.

[1] Deep residual learning for image recognition. In CVPR, 2016.

Difficulty Metrics: Textual Difficulty

- Heuristic Metric:
  - We adopt the number of abnormal sentences in a report to define the difficulty of a report. Following [1], we consider sentences which contain "no", "normal", "clear", "stable" as normal sentences, the rest sentences are consider as abnormal sentences.

- Model Confidence:
  - To this end, we employ the negative loss value of a report sample from the trained report generator as the model confidence measure to indicate whether a sampled report is easy to be generated or not.

[1] On the automatic generation of medical imaging reports. In ACL, 2018.

3. Experiments

Datasets and Metrics

Dataset

MIMIC-CXR [1] and IU-Xray [2]

Medical Report:

Lungs are clear. No pleural effusions or pneumothoraces. Heart and mediastinum of normal size and contour. [1]scoliosis.

Evaluation Metrics

- ✓ BLEU [3]
- ✓ METEOR [4]
- ✓ ROUGE [5]

[1] MIMIC-CXR: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042.
[2] Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Medical Informatics Assoc., 23(2):304–310.
[3] BLEU: a Method for automatic evaluation of machine translation. In ACL, 2002.
[4] METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In IEEvaluation@ACL, 2005
[5] ROUGE: A package for automatic evaluation of summaries. In ACL, 2004.

Quantitative Results

Methods	Dataset: MIMIC-CXR (Johnson et al., 2019)						Dataset: IU-Xray (Demner-Fushman et al., 2016)					
	B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
NIC (Vinyals et al., 2015)†	0.290	0.182	0.119	0.081	0.112	**0.249**	0.352	**0.227**	0.154	0.109	0.133	0.313
w/ CMCL	**0.301**	**0.189**	**0.123**	**0.085**	**0.119**	0.241	**0.358**	0.223	**0.160**	**0.114**	**0.137**	**0.317**
Spatial-Attention (Lu et al., 2017)†	0.302	0.189	0.122	0.082	**0.120**	**0.259**	0.374	0.235	0.158	0.120	0.146	0.322
w/ CMCL	**0.312**	**0.200**	**0.125**	**0.087**	0.118	0.258	**0.381**	**0.246**	**0.164**	**0.123**	**0.153**	**0.327**
Adaptive-Attention (Lu et al., 2017)†	**0.307**	**0.192**	0.124	0.084	0.119	0.262	0.433	**0.285**	0.194	0.137	0.166	**0.349**
w/ CMCL	0.302	**0.192**	**0.129**	**0.091**	**0.125**	**0.264**	**0.437**	0.281	**0.196**	**0.140**	**0.174**	0.338
CNN-HLSTM (Krause et al., 2017)†	0.321	0.203	0.129	0.092	0.125	0.270	0.435	0.280	0.187	0.131	0.173	0.346
w/ CMCL	**0.337**	**0.210**	**0.136**	**0.097**	**0.131**	**0.274**	**0.462**	**0.293**	**0.207**	**0.155**	**0.179**	**0.360**
HLSTM+att+Dual (Harzig et al., 2019)†	0.328	0.204	0.127	**0.090**	**0.122**	0.267	0.447	0.289	0.192	0.144	**0.175**	0.358
w/ CMCL	**0.330**	**0.206**	**0.133**	0.088	0.119	**0.272**	**0.461**	**0.298**	**0.201**	**0.150**	0.173	**0.359**
Co-Attention (Jing et al., 2018)†	0.329	0.206	0.133	0.095	0.129	0.273	0.463	0.293	0.207	0.155	0.178	0.365
w/ CMCL	**0.344**	**0.217**	**0.140**	**0.097**	**0.133**	**0.281**	**0.473**	**0.305**	**0.217**	**0.162**	**0.186**	**0.378**

Table 1. B-n, Mand R-L are short for BLEU-n, METEOR and ROUGE-L, respectively. Higher is better in all columns. As we can see, all baseline models enjoy comfortable improvements in most metrics with our CMCL, which doesn’t introduce additional parameters and only requires a small modification to the training data pipelines.

北京大学 PEKING UNIVERSITY

Quantitative Results

Methods	Dataset: MIMIC-CXR (Johnson et al., 2019)						Dataset: IU-Xray (Demner-Fushman et al., 2016)					
	B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
HRGR-Agent (Li et al., 2018)	-	-	-	-	-	-	0.438	0.298	0.208	0.151	-	0.322
CMAS-RL (Jing et al., 2019)	-	-	-	-	-	-	0.464	0.301	0.210	0.154	-	0.362
SentSAT + KG (Zhang et al., 2020a)	-	-	-	-	-	-	0.441	0.291	0.203	0.147	-	0.367
Up-Down (Anderson et al., 2018)	0.317	0.195	0.130	0.092	0.128	0.267	-	-	-	-	-	-
Transformer (Chen et al., 2020)	0.314	0.192	0.127	0.090	0.125	0.265	0.396	0.254	0.179	0.135	0.164	0.342
R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.277	0.470	0.304	0.219	0.165	0.187	0.371
CMCL (Ours)	0.344	0.217	0.140	0.097	0.133	0.281	0.473	0.305	0.217	0.162	0.186	0.378

Table 2. Comparison with existing state-of-the-art methods on the MIMIC-CXR and the IU-X-ray datasets. CMCL is taken from the "Co-Attention w/ CMCL" in Table 1. In this table, the Red and Blue colored numbers denote the best and second best results across all approaches, respectively.

Human Evaluation

vs. Models	Baseline wins	Tie	'w/ CMCL' wins
CNN-HLSTM (Jing et al., 2018)†	15	28	**57**
Co-Attention (Jing et al., 2018)†	24	35	**41**

Table 3. We invite 2 professional clinicians to conduct the human evaluation for comparing our method with baselines. All values are reported in percentage (%).

Qualitative Results

Ground Truth:
The heart and mediastinum are normal. The lungs are clear. [1]There is mild blunting of the right costophrenic XXXX. There is no infiltrate, mass or pneumothorax. The right internal jugular catheter has been removed.

Co-Attention [1]**:**
The heart is enlarged. *There is no pneumothorax.* No acute bony abnormality. There is a moderate right pleural effusion with associated atelectasis. The left lung is clear. *No pneumothorax is seen*.

Ours:
[1]Blunting of right costophrenic. Heart size is normal. No acute bony abnormality. There is no pleural effusion. No visualized pneumothorax. The lungs are clear.

Ground Truth:
Lungs are clear. No pleural effusions or pneumothoraces. Heart and mediastinum of normal size and contour. [1]Scoliosis.

Co-Attention [1]**:**
No acute bony abnormalities. *No pneumothorax* or pleural effusion. The heart is normal in size. The lungs are clear. The hilar and mediastinal contours are normal. *No evidence of pneumothorax.*

Ours:
No acute cardiopulmonary abnormality. No focal airspace consolidation. Clear lungs. There is no pneumothorax or pleural effusion. [1]Scoliosis is present.

Figure 1. Two examples of ground truth reports and reports generated by a state-of-the-art approach Co-Attention [1] and our approach. The Red colored text indicate the abnormalities. The Co-Attention [1] fails to depict some rare but important abnormalities and generates some error sentences (Underlined text) and repeated sentences (*Italic* text). Our approach generates structured and robust reports, which show significant alignment with ground truth reports (Blue colored text) and are supported by accurate abnormal descriptions (Red colored text).

[1] On the automatic generation of medical imaging reports. In ACL, 2018.

4. Conclusions

Conclusions

- In this paper, we propose the competence-based multimodal curriculum learning framework (CMCL) to alleviate the data bias by efficiently utilizing the limited medical data for medical report generation.

- To this end, considering the difficulty of accurately capturing and describing the abnormalities, we first assess four sample difficulties of training data from the visual complexity and the textual complexity, resulting in four different curricula.

- Next, CMCL enables the model to be trained with the appropriate curricula and gradually proceed from easy samples to more complex ones in training.

- Experimental results demonstrate the effectiveness and the generalization capabilities of CMCL, which consistently improves the performance of the baselines under most metrics.

☺

Thank you for your attention!

If you have any questions about our paper, you can send an email to fenglinliu98@pku.edu.cn