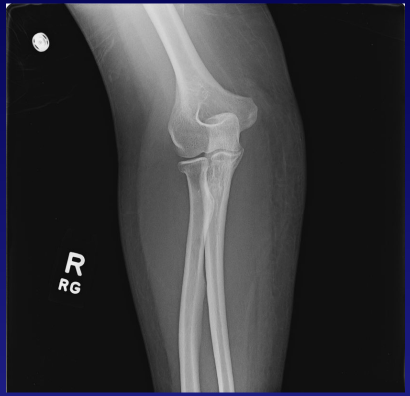
Medical Knowledge-Guided Deep Curriculum Learning for Elbow Fracture Diagnosis from X-Ray Images

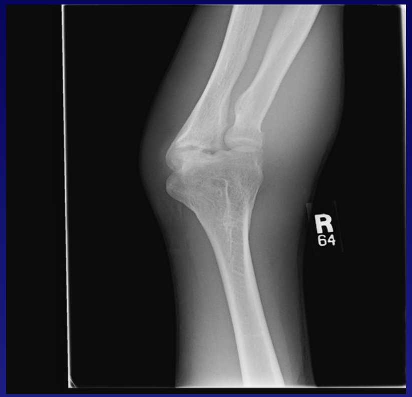
Jun Luo, MS[1]
Gene Kitamura, MD[2]
Emine Doganay, PhD[2]
Dooman Arefan, PhD[2]
Shandong Wu, PhD[1,2,3,4]

[1]Intelligent Systems Program
Department of [2] Radiology / [3]Biomedical Informatics / [4]Bioengineering
University of Pittsburgh

Background

- Elbow fracture is one of the fracture types that happens most frequently among people across all ages
  - Needs timely diagnosis and treatment since it could cause neurovascular damage [1]
  - X-ray helps assessment by visualization
  - A physician needs years of training to read and understand elbow X-ray





[1] Saeed, Wajeeha, and Muhammad Waseem. "Elbow Fractures Overview." (2017).

Background

- Elbow fracture is one of the fracture types that happens most frequently among people across all ages
  - Needs timely diagnosis and treatment since it could cause neurovascular damage [1]
  - X-ray helps assessment by visualization
  - A physician needs years of training to read and understand elbow X-ray
- Deep learning
  - Thrives in recent years
  - Needs only hours of training on the elbow X-ray images to classify or detect
  - Most methods are purely data-driven, not leveraging medical knowledge from physicians
- Some recent works investigate incorporating medical knowledge into deep learning.

[1] Saeed, Wajeeha, and Muhammad Waseem. "Elbow Fractures Overview." (2017).

Purpose

- Investigate how to incorporate clinical knowledge (medical expertise) to data-driven deep learning for elbow fracture classification

- Develop a curriculum learning based methodology for training

- Evaluate our method on whether it can improve from method without knowledge

Curriculum Learning

- Curriculum learning [2]
  - Let the machine mimic human learning by “first easy then hard”
  - Has been applied in many areas
    - image classification [3]
    - object detection [4]
    - semantic segmentation
    - self/semi supervised learning [3]
    - multi-task learning
    - multi-modal learning [3]
  - Appropriate to incorporate outside knowledge in nature (into definitions of “easy” and “hard”), but few works have done so especially in medical domain

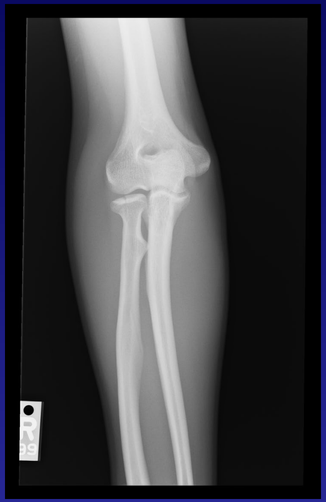
[2] Bengio, Yoshua, et al. "Curriculum learning." Proceedings of the 26th annual international conference on machine learning. (2009).
[3] Gong, Chen, et al. "Multi-modal curriculum learning for semi-supervised image classification." IEEE Transactions on Image Processing 25.7 (2016): 3249-3260.
[4] Zhang, Dingwen, et al. "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework." International Journal of Computer Vision 127.4 (2019): 363-380.

Methodology

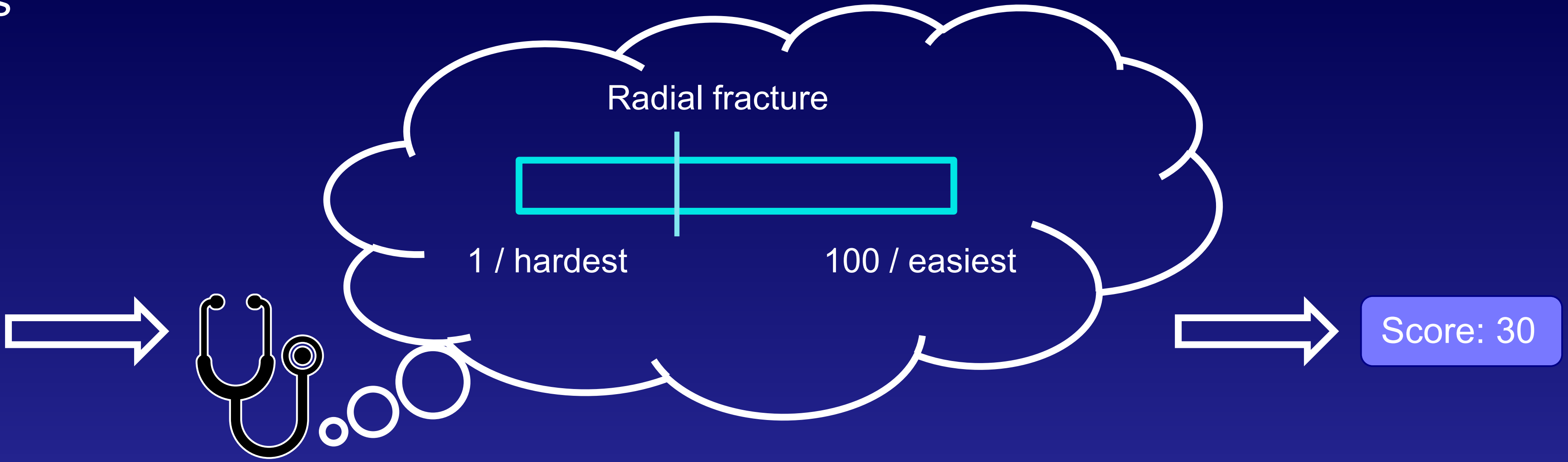
- Propose a deep curriculum learning framework
  - Incorporates medical knowledge from domain experts (radiologists) through a curriculum learning method on a binary (fracture Vs. normal) classification task of elbow fractures
  - Knowledge represented as a quantification of a radiologist's clinical experience.
  - Based on the knowledge, design a scoring criterion that scores each training image
  - Create the curriculum that is guided by the scores
  - Train the model according to the curriculum.

Method

- Scoring criterion
  - A quantifiable criterion that reflects how hard it is to classify a certain subtype of elbow fracture in clinical practice
  - Designed based on medical knowledge
  - Score of an elbow X-ray image indicates the difficulty of diagnosing its fracture subtypes



Radial fracture



Method

- Scoring of different fracture subtypes
  - Fracture images: 6 subtypes

Figure 1. Six Subtypes of elbow fractures: (a) Ulnar fracture; (b) Radial fracture; (c) Humeral fracture; (d) Dislocation; (e) Complex fracture/multi-type fracture; (f) Coronoid process fracture.

Method

- Scoring of different fracture subtypes
  - Fracture images: 6 subtypes
  - Assign scores from human expert's knowledge

Table 2. Difficultness scoring of the normal cases and six subtype fractures of the elbow (1 – hardest; 100 – easiest).

	(normal)	*(a)*	*(b)*	*(c)*	*(d)*	*(e)*	*(f)*
Score	30	30	30	70	40	90	10

(a) (b) (c) (d) (e) (f)

Figure 1. Six Subtypes of elbow fractures: (a) Ulnar fracture; (b) Radial fracture; (c) Humeral fracture; (d) Dislocation; (e) Complex fracture/multi-type fracture; (f) Coronoid process fracture.

Method

- Curriculum
  - Permute the data at each epoch by sampling without replacement
  - Sampling probability for each image is guided by the score

Method

- Curriculum
  - Permute the data at each epoch by sampling without replacement
  - Sampling probability for each image is guided by the score
- Sampling probability (notation)
  - Consider a triplet (x_i, y_i, f_i)
  - Let $p_{i,(e)}$ be its sampling probability before the sampling at epoch e

Method

- Initialization of sampling probability
  - $p_{i,(1)}$ computed from the scores for all images
  - Let s_{f_k} be the score for image x_k
  - $p_{i,(1)} = \frac{s_{f_i}}{\sum_{j=1}^{N} s_{f_j}}$
- Update of sampling probability
  - Update the value of sampling probability at the beginning of each epoch
  - $\lambda_i = \sqrt[L]{\frac{p_{i,(final)}}{p_{i,(1)}}} = \sqrt[L]{\frac{1/N}{p_{i,(1)}}}$
  - $p_{i,(e)} = \begin{cases} p_{i,(e-1)} \cdot \lambda_i, & 2 \le e \le L \\ \frac{1}{N}, & L < e \le E \end{cases}$
  - Applicable to other curriculum learning framework with sampling without replacement strategy (later denoted as probability update algorithm)

Study cohort

- Experiment of 1,865 elbow trauma patients, binary (665 fracture Vs. 1,200 normal) classification
- Testing
  - sample 100 out of 400 test set normal cases, classify against 73 test set fracture cases for more balanced ratio
  - Repeat the test 5 times, results reported are mean $\pm$ standard deviation

Table 1. Number of images of the normal cases and six subtypes of the elbow fractures.

Type	*(normal)*	*(a)*	*(b)*	*(c)*	*(d)*	*(e)*	*(f)*	Total of *(a)*-*(f)* *(fracture)*	Total
Train	800	88	340	84	11	42	27	592	1392
Test	400	10	44	9	2	4	4	73	473
Total	1200	98	384	93	13	46	31	665	1865

Evaluation

- Compared methods
  - Baseline
    - Backbone: VGG16 [2]
    - Generally used random order (shuffle)
  - A previous method, here refer to as MBDCL [3]
    - Backbone: VGG16
    - Also uses knowledge incorporated curriculum learning
    - Uses a different way of incorporating knowledge into the curriculum learning framework
      - Constant sampling probability
      - Initialization of sampling probability as ranking of difficultness

[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
[3] Jiménez-Sánchez, Amelia, et al. "Medical-based deep curriculum learning for improved fracture classification." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, (2019).

Evaluation

- Anti-curriculum settings
  - Opposite strategy of our curriculum
  - Re-assign the difficultness scores by using 100 minus the original scores
  - “Easy” in the original curriculum is considered “hard” in anti-curriculum and vis versa
- Plug in our proposed update algorithm to improve an existing method
  - Algorithm is applicable to sampling-based curriculum settings framework
  - MBDCL + Update
  - Anti-curriculum + Update

Results

	Average on 5 Different Test Subsets (100 normal + 73 fracture)			
	Accuracy	AUC	Average Precision	F1 score
Baseline	0.776±0.026	0.834±0.025	0.788±0.043	0.716±0.032
MBDCL[5]	0.797±0.025	0.865±0.019	0.831±0.029	0.763±0.022
MBDCL[5] + proposed update algorithm	0.806±0.027	0.878±0.022	0.847±0.035	0.762±0.029
Ours	**0.809±0.023**	**0.882±0.020**	**0.852±0.036**	**0.778±0.024**
Anti-curriculum	0.765±0.032	0.863±0.020	0.821±0.036	0.757±0.025
Anti-curriculum + proposed update algorithm	0.803±0.036	0.871±0.025	0.838±0.045	0.773±0.031

mean ± standard deviation

Results

Comparison 1	Average on 5 Different Test Subsets (100 normal + 73 fracture)			
	Accuracy	AUC	Average Precision	F1 score
Baseline	0.776±0.026	0.834±0.025	0.788±0.043	0.716±0.032
MBDCL[5]	0.797±0.025	0.865±0.019	0.831±0.029	0.763±0.022
MBDCL[5] + proposed update algorithm	0.806±0.027	0.878±0.022	0.847±0.035	0.762±0.029
Ours	**0.809±0.023**	**0.882±0.020**	**0.852±0.036**	**0.778±0.024**
Anti-curriculum	0.765±0.032	0.863±0.020	0.821±0.036	0.757±0.025
Anti-curriculum + proposed update algorithm	0.803±0.036	0.871±0.025	0.838±0.045	0.773±0.031

mean ± standard deviation

Results

Comparison 2	Average on 5 Different Test Subsets (100 normal + 73 fracture)			
	Accuracy	AUC	Average Precision	F1 score
Baseline	0.776±0.026	0.834±0.025	0.788±0.043	0.716±0.032
MBDCL[5]	0.797±0.025	0.865±0.019	0.831±0.029	0.763±0.022
MBDCL[5] + proposed update algorithm	0.806±0.027	0.878±0.022	0.847±0.035	0.762±0.029
Ours	**0.809±0.023**	**0.882±0.020**	**0.852±0.036**	**0.778±0.024**
Anti-curriculum	0.765±0.032	0.863±0.020	0.821±0.036	0.757±0.025
Anti-curriculum + proposed update algorithm	0.803±0.036	0.871±0.025	0.838±0.045	0.773±0.031

mean ± standard deviation

Results

Comparison 3	Average on 5 Different Test Subsets (100 normal + 73 fracture)			
	Accuracy	AUC	Average Precision	F1 score
Baseline	0.776±0.026	0.834±0.025	0.788±0.043	0.716±0.032
MBDCL[5]	0.797±0.025	0.865±0.019	0.831±0.029	0.763±0.022
MBDCL[5] + proposed update algorithm	0.806±0.027	0.878±0.022	0.847±0.035	0.762±0.029
Ours	**0.809±0.023**	**0.882±0.020**	**0.852±0.036**	**0.778±0.024**
Anti-curriculum	0.765±0.032	0.863±0.020	0.821±0.036	0.757±0.025
Anti-curriculum + proposed update algorithm	0.803±0.036	0.871±0.025	0.838±0.045	0.773±0.031

mean ± standard deviation

Results

Comparison 4	Average on 5 Different Test Subsets (100 normal + 73 fracture)			
	Accuracy	AUC	Average Precision	F1 score
Baseline	0.776±0.026	0.834±0.025	0.788±0.043	0.716±0.032
MBDCL[5]	0.797±0.025	0.865±0.019	0.831±0.029	0.763±0.022
MBDCL[5] + proposed update algorithm	0.806±0.027	0.878±0.022	0.847±0.035	0.762±0.029
Ours	**0.809±0.023**	**0.882±0.020**	**0.852±0.036**	**0.778±0.024**
Anti-curriculum	0.765±0.032	0.863±0.020	0.821±0.036	0.757±0.025
Anti-curriculum + proposed update algorithm	0.803±0.036	0.871±0.025	0.838±0.045	0.773±0.031

mean ± standard deviation

Discussion

- We designed a novel medical knowledge-guided deep curriculum learning method for elbow fracture diagnoses from X-ray images.
  - The knowledge is a pre-defined quantitative scoring criterion
    - based on classification difficulty of different elbow fracture subtypes
    - incorporation of radiologists' diagnosis knowledge

Discussion

- Our results showed that by incorporating medical knowledge:
  - Our method outperforms all other compared methods
  - The anti-curriculum settings demonstrate inferior results as expected
  - The proposed probability update algorithm can further enhance other curriculum learning methods.
- It will be a more effective way to augment the pure data-driven deep learning by leveraging the medical knowledge to the models.

Discussion

- Limitations and future works
  - The dataset:
    - Single-center study; we will need a larger dataset for further evaluation of our method.
  - The classification is based on a single perspective (frontal view)
    - Reality: X-ray images often taken from multiple perspectives (frontal and lateral views)
    - Incorporating domain knowledge with different views might further improve the performance
  - The scoring criterion on classification difficulty is based on single reader’s knowledge
    - Different forms of the knowledge from multiple readers may be more helpful

Conclusion

- Our study is a novel strategy for incorporating medical knowledge to guide and enhance data-driven deep learning for medical applications especially for elbow fracture diagnosis.

- Our method demonstrates a new mechanism of defining and incorporating existing clinical experience and knowledge into artificial intelligence (AI) tools for clinical applications.

- Knowledge-guided AI is an attractive direction of the future research for AI in medical applications.

Acknowledgement:
Intelligent Computing for Clinical Imaging (ICCI) Lab, University of Pittsburgh

- NIH/NCI R01 (#CA193603)
- NIH/NCI R01 Supplement (#CA193603-S)
- NIH/NCI R01 (#CA218405)
- PHDA / UPMC Enterprise (Early Commercialization Development)
- Amazon Machine Learning Award
- RSNA Research Scholar Grant (#RSCH1530)
- UPCI-IPM Pilot Award (#MR2014-77613)
- Pitt CTSI Biomedical Modeling Pilot Award

Thank you!

Questions?

jul117@pitt.edu