Predicting Future Knee Osteoarthritis Using Baseline Knee Radiographs

Matthew R. Titchenal, Nishant Pandit , Stephanie R. Young
mtitch@stanford.edu, nish1519@stanford.edu, sryoung@stanford.edu

Predicting Knee Osteoarthritis

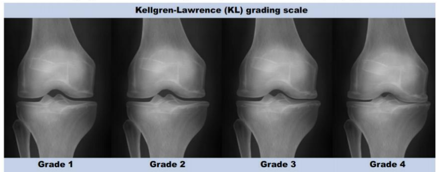
Background and Motivation

- Osteoarthritis (OA) is a leading cause of disability worldwide[1]

Figure 1. OA description

- Need for prognostic techniques to detect early OA

OA severity quantified using Kellgren Lawrence (KL) grades



CLASSIFICATION	Normal	Doubtful	Mild	Moderate	Severe
DESCRIPTION	No features of OA	Minute osteophyte: doubtful significance	Definite osteophyte: normal joint space	Moderate joint space reduction	Joint space greatly reduced: subchondral sclerosis

Figure 2. KL Grading Source: Antony et al, 2016[2]

The Project

- Build a 2-layer neural net with binary classifier to identify patients with normal or doubtful grades (KL= 0 or 1) at baseline as progressors (KL ≥ 2 at follow-up) or non-progressors (KL ≤ 1 at follow-up)

Inputs		**Output**
Baseline Knee Radiographs (KL 0 or 1)	→	Progressor (y=1) or Non-Progressor (y=0)

Results

- Achieved 70% precision/recall: proof-of-concept

Transfer Learning & Feature Extraction

Transfer Learning

- Very deep pre-trained CNN on ImageNet Data (VGG-16)
  - 3x3 convolutional filter (small)
  - Generalize well to other data sets

Feature Extraction

- VGG-16 & Maxpool-5

Final Model

- Two layer neural net, Dropout with 50% probability, Batch Normalization, Mini-Batch Size = 20, Learning Rate = 0.01, Mini-Batch Gradient Descent, Binary Cross-Entropy Loss

Dataset, Labeling, and Pre-processing

Osteoarthritis Initiative (OAI) Dataset

- Patient data (4794 patients) at 0, 12, 24, 36, 48, 72, and 96 month time points
  - Bilateral knee radiographs (DICOM images)
  - Bilateral Radiologist KL Grades

Data Pre-Processing and Labeling

- Excluded all knees with OA at baseline (KL ≥ 2)
- Classified knees as progressors (y=1) if KL ≥ 2 at later time point.
- Total number of progressors = 1586 knees
- Total number of non-progressors = 7540 knees
- Converted DICOM images to PNG
- Split images into right and left knees
- Mirrored left knees and normalize images
- Pair images with associated labels
- Data resizing / cropping to identify knee region

Data Augmentation

- Brought in data from multiple time points to augment data set
- Sampled X-ray images from all KL ≤ 1 knees

Figure 3. Pre-processing of knee radiographs

Model Results & Software Flow

Binary Cross Entropy Loss: $\mathcal{L}(y, \hat{y}) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$

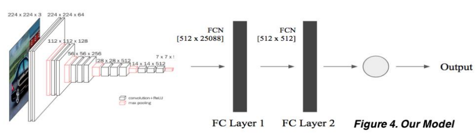


Figure 4. Our Model

Final Models	Parameters	Train	Test
Model #1: 1500 progressors / 1500 non-progressors (90% / 10%)	Batch Normalization Learning Rate = 0.01 **2 Layer Network** **No Dropout**	P: 99% R: 99%	P: 67% R: 67%
Model #2: same as #1	Batch Normalization Learning Rate = 0.01 **2 Layer Network** **Dropout = 0.9**	P: 99% R: 99%	P: 66% R: 66%
Model #3: same as #1	Batch Normalization Learning Rate = 0.01 **1 Layer Network** **Dropout = 0.9**	P: 99% R: 99%	P: 68% R: 68%
Model #4: same as #1	Batch Normalization Learning Rate = 0.01 **2 Layer Network** **Dropout = 0.5**	P: 99% R: 99%	P: 70% R: 70%

Figure 6. Model results / parameters. Best model is #4 highlighted in green

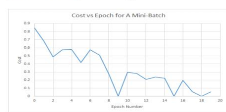


Figure 5. Cost vs. Epoch

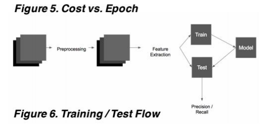


Figure 6. Training / Test Flow

Experiments

Experiment	Parameters	Train	Test
#1: 700 progressors / 700 non-progressors (90% / 10%)	No Batch Normalization Learning Rate = 0.01 2 Layer	P: 58% R: 58%	P: 47% R: 47%
#2: Same as #1	No Batch Normalization Learning Rate = 0.001 2 Layer	P: 75% R: 71%	P: 53% R: 53%
#3: Same as #1	Batch Normalization Learning Rate = 0.01 2 Layer	P: 99% R: 99%	P: 61% R: 61%
#4: Same as #1	Batch Normalization Learning Rate = 0.01 2 Layer Adam Optimizer	P: 97% R: 98%	P: 60% R: 60%
#5: Same as #1	Batch Normalization Learning Rate = 0.01 2 Layer L2 Regularization	P: 99% R: 99%	P: 60% R: 60%
#6: Same as #1	Batch Normalization Learning Rate = 0.01 2 Layer Dropout = 0.9	P: 99% R: 99%	P: 59% R: 59%

- Batch Normalization very important to model
- SGD better performance than Adam optimizer
- Lower alpha without BN reduce overfitting
- Regularization had little effect
- Dropout at 90% had negative effect but helped at 50%
- Data augmentation using multiple time points
  - Increased precision and recall

Discussion

- This study showed proof of concept for the use of deep learning to detect features of "healthy" knee radiographs that are predictive of OA that current medical techniques have failed to identify
  - Using only KL <= 1 knee radiographs
  - 70% precision / recall on the test set
  - Deep learning detecting features predictive of OA that may be undetectable to the human eye
- Increasing dataset size improved results
  - Still overfitting on our training data set
  - More data will be useful for mitigating the high variance
- It may not be possible to achieve accuracy much higher than this using only knee radiographs
  - Radiographs contain only information on bony structures
  - Soft tissue information from cartilage and other structures in the knee may be important in predicting knee OA

Future Work

- Increasing the size of the data set would improve results
- Visualizing the features learned by these networks could help identify clinically correctable problems
- Include patient demographic data to give model more features
- Performance could be improved by using magnetic resonance images (MRI)
  - MRI contains more information on various features important to OA including soft tissues like cartilage, synovium, and ligaments

References

[1] J. Antony, K. McGuinness, N. E. O'Connor, K. Moran, N. E. O. Connor, and K. Moran, "Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks," *ArXiv e-prints*, vol. 1609, no. ICPR 2016 proceedings, pp. 1195–1200, 2016.
[2] "United States Bone and Joint Initiative: The Burden of Musculoskeletal Diseases in the United States (BMUS)," *Third Edition*, 2014. [Online]. Available: http://www.boneandjointburden.org/
[3] S. Suresha, L. Kidziński, E. Halilaj, G. E. Gold, and S. L. Delp, "Automated staging of knee osteoarthritis severity using deep neural networks," *Osteoarthr. Cartil.*, vol. 26, p. S441, Apr. 2018.
[4] S. Suresha, A. Mahajan, N. Dalal, "Automatically Quantifying Radiographic Knee Osteoarthritis Severity," *CS229* Final Report
[5] https://github.com/ayush1997/Xvision