Annotation-Efficient Deep Semi-Supervised Learning for Automatic Knee Osteoarthritis Severity Diagnosis from Plain Radiographs

Huy Hoang Nguyen[*]
University of Oulu

Simo Saarakkala
University of Oulu

Matthew B. Blaschko
KU Leuven

Aleksei Tiulpin
KU Leuven, Ailean Technologies Oy &
University of Oulu

1 Introduction

Osteoarthritis (OA) is a worldwide disease that occurs in joints causing irreversible damage to cartilage and other joint tissues. The knee is particularly vulnerable to OA, and millions of people, regardless of gender, geographical location, and race, suffer from knee OA. When the disease reaches the late stages, patients have to undergo a total knee replacement (TKR) surgery to avoid disability. For society, direct and indirect costs of OA are notably high, and for instance, OA is one of the five most expensive healthcare expenditures in Europe [1]. In the United States, the burden of knee OA is also huge, and TKR surgeries annually cost over 10 billion dollars [2]. If knee OA could be detected at an early stage, its progression might be slowed down, thereby yielding significant benefits at personal and societal levels [3].

Radiographs, affordable and widely available in primary care, are sufficiently informative for knee OA severity diagnosis. However, the process of visual assessment of radiographs is rather tedious, and as a result, various Deep Learning (DL) based methods for automatic diagnosis of knee OA severity have recently been developed [4–6]. The primary drawback of these methods is their dependency on large amounts of annotations, which are expensive in terms of cost and time to collect.

In this paper, we introduce *Semixup*, a novel Semi-Supervised Learning (SSL) method, which we apply for automatic diagnosis of the knee OA severity in an annotation-efficient manner. This extended abstract is a short version of [7].

2 Method

2.1 Data

Our datasets were from the Osteoarthritis Initiative (OAI) and the Multicenter Osteoarthritis Study (MOST) patient cohorts. While OAI data were used for training and validation, MOST data were used for an independent testing. Bilateral posterior-anterior knee radiographs in those datasets were acquired in fixed flexion using Synaflexer® positioning frame and 10° beam angle. All available radiographs from all the follow-ups in the OAI were utilized. In the MOST dataset, we applied similar strategy, but excluded the images acquired with 5° or 15° beam angle as well as the ones without the Osteoarthritis Research Society International (OARSI) scores. We used the Kellgren Lawrence (KL) grading [8] as a reference standard to assess the severity of knee OA. The KL system classifies OA in 5 levels, where 0 indicates no signs of OA, 1 – doubtful OA, 2 – early OA, 3 – moderate OA, and 4 –

[*] Corresponding author: huy.nguyen@oulu.fi

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

end stage OA. After the data selection phase, we obtained 39,902 and 3,445 knee images for training and independent evaluation phases, respectively.

For training and model selection, we formed 4 settings of annotation data: 50, 100, 500, and 1000 samples with annotation per KL grade, each of which has 6 unannotated data settings: $1\times$, $2\times$, $\dots$, $6\times$ more samples without annotation. Thus, in total, we had 24 data settings for SSL methods. In addition, we gathered all annotated data in the OAI to form a large data setting for training a SL model.

2.2 Siamese Architecture

We designed a Siamese DL architecture for all SL and SSL methods [7]. The main component in the model is 3×3 convolutional blocks of a 3×3 convolutional layer, an instance normalization, and a LeakyRELU activation. The network has 4 subsampling layers separating groups of 3×3 convolutional blocks, and in the bottleneck, it produces a feature map of 16×16 for a given 128×128 input image in each branch of the model. Each of the feature maps is compressed into 1×1 feature by a pooling layer. Finally, we concatenate the features from each of the branches and use a fully connected layer to predict the KL grades.

To extract the left and right knee sides for our Siamese network, we closely followed [6], except that we performed a tighter crop with the size $110mm \times 110mm$ instead of $130mm \times 130mm$ and used a mean of 0.5 and a standard deviation of 0.5 instead of extracting them from OAI data. As a result, we obtained pairs of 128×128-pixel images ($0.37mm$ pixel spacing).

2.3 Semixup

The core idea of our method is to enforce continuous predictions within and out of the data manifold $\mathcal{M}$. As such, explicitly enforce the smoothness of predictions within the data manifold $\mathbb{E}_x \|J_{\mathcal{M}}\|_F^2$, where $J_{\mathcal{M}}$ denotes the Jacobian along the data manifold $\mathcal{M}$ and $\|\cdot\|_F$ is the Frobenius norm. To achieve this, we use *consistency regularization*, which widely used in the SSL domain [9], and thus minimize:

$$\mathbb{E}_{x \sim p(x)} \mathbb{E}_{T, T' \sim p(\tau)} \left\| f_\theta(Tx) - f_\theta(T'x) \right\|_2^2, \tag{1}$$

where f_θ is our model with learnable parameters θ, $p(x)$ is the distribution of annotated and unannotated data, and $p(\tau)$ is the distribution of stochastic transformations.

To generate the out-of-manifold samples, we utilize *mixup* [10], which generates new samples by creating a convex combination of two arbitrary samples x_i, x_j with a coefficient $\lambda \sim \text{Beta}(\alpha, \alpha)$:

$$x_{mix} = \text{Mix}_\lambda(x_i, x_j) = \lambda x_i + (1 - \lambda) x_j. \tag{2}$$

Here, we also use the consistency regularization, but enforce consistency between the predictions on the augmented version of the sample, and the one, which is outside the manifold, but in a close proximity to it. This is achieved by making sure that the original sample's visual appearance is still dominant (i.e., $\lambda > 0.5$). Formally, we target to minimize:

$$\mathbb{E}_{\lambda \sim \text{Beta}(\alpha,\alpha)} \mathbb{E}_{x, x' \sim p(x)} \mathbb{E}_{T \sim p(\tau)} \left\| f_\theta(\text{Mix}_\lambda(x, x')) - f_\theta(Tx) \right\|_2^2. \tag{3}$$

Our final regularization term is the interpolation consistency (IC) [11], which also relies on *mixup*, and it encourages a linear behavior of the model's predictions with linear change of the input. We note that from the other regularizers, we obtain almost all the ingredients needed to compute IC term:

$$\mathbb{E}_{\lambda \sim \text{Beta}(\alpha,\alpha)} \mathbb{E}_{x_i, x_j \sim p(x)} \left\| \text{Mix}_\lambda(f_\theta(x_i), f_\theta(x_j)) - f_\theta(x_{mix}) \right\|_2^2. \tag{4}$$

In *Semixup*, for any pair of samples with or without label x_i and x_j, we apply 2 random augmentations T and T' to x_i, and then compute $x_{mix} = \text{Mix}_\lambda(Tx_i, x_j)$ and $p_{mix} = \text{Mix}_\lambda(f_\theta(Tx_i), f_\theta(x_j))$ with $\lambda > 0.5$. To this end, we minimize the unsupervised loss that is a linear combination of the in-manifold term $\|f_\theta(Tx_i) - f_\theta(T'x_i)\|_2^2$, the out-of-manifold terms $\|f_\theta(x_{mix}) - f_\theta(Tx_i)\|_2^2$, $\|f_\theta(x_{mix}) - f_\theta(T'x_i)\|_2^2$, and the interpolation term $\|f_\theta(x_{mix}) - p_{mix}\|_2^2$. With respect to annotated data, we simply applied *mixup* on them, and used cross-entropy as described in [10].

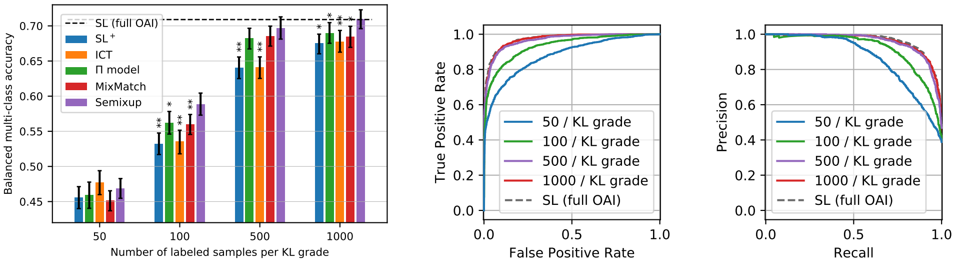
2.4 Statistical Analyses

To measure the performance of predicting KL grading, we utilized balanced multi-class accuracy (BA), quadratic Cohen kappa score (KC), and mean square error (MSE). Besides, we investigated the ability of early knee OA detection (KL$\geq$ 2) by receiver operating characteristic (ROC) curves, area under the ROC curve (AUC), precision-recall (PR) curves, and average precision (AP). For statistical analyses, we used the Wilcoxon signed-rank test [12].

3 Experiments

We compared *Semixup* to interpolation consistency training (ICT) [11], Π model [9], and MixMatch [13]. All the SL and SSL methods were trained using the same Siamese network presented in 2.2 and augmented by an identical set of transformation settings. The training was done on OAI data, and testing on MOST data.

Results in Figure 1 shows that *Semixup* significantly outperformed the SSL baselines throughout the annotated data settings, expect for the case of 50 annotated samples per KL grade. The fully supervised model achieved a BA of 70.9 ± 0.8, which is a state-of-the-art (SOTA) performance. However, it required the large data setting with $31,902$ annotated samples. Semixup reached BAs of 69.7 ± 0.8 and 71 ± 0.8 with 500 and 1000 annotated images per KL grade, respectively. Statistical analyses showed that the fully supervised model is insignificantly different from the 2 SSL models, having $p = 0.054$ and $p = 0.368$, respectively. Finally, Figures 1 and Table 1 consistently show that Semixup performed closely to the SL model in various metrics on the KL grade prediction as well as the early knee OA detection task while requiring over 6 times less annotated samples.



(a) Balanced accuracy (95% CI) (b) ROC curve (c) PR curve

Figure 1: Performance comparisons between *Semixup* and the baseline methods on the MOST dataset. SL (full OAI) indicates the performance of the SL model trained on the full OAI dataset. (a) SL$^+$ indicates the SL model, * and ** indicate the statistically significant difference ($p < 0.05$ and $p < 0.001$, respectively). (b) + (c) Early knee OA detection (KL$\geq$ 2).

Table 1: Comparison of our best models trained with *Semixup* against our well-tuned SL model trained on full OAI data.

Method	# annotations	KC	MSE	AUC (KL $\geq$ 2)	AP (KL $\geq$ 2)
Semixup	250	0.708 [0.692, 0.724]	1.080 [1.000, 1.022]	0.880 [0.869, 0.891]	0.861 [0.849, 0.872]
	500	0.789 [0.776, 0.801]	0.810 [0.764, 0.860]	0.933 [0.926, 0.940]	0.916 [0.906, 0.925]
	2,500	0.877 [0.870, 0.884]	0.442 [0.416, 0.466]	0.967 [0.962, 0.972]	0.956 [0.950, 0.962]
	5,000	0.878 [0.870, 0.885]	0.458 [0.430, 0.487]	0.972 [0.968, 0.976]	0.959 [0.953, 0.965]
SL (full OAI)	31,922	0.881 [0.873, 0.889]	0.440 [0.414, 0.471]	0.974 [0.969, 0.978]	0.963 [0.958, 0.969]

4 Conclusion

We present a novel SSL method – *Semixup*, which achieves high performance in the task of automatic knee osteoarthritis grading from plain radiographs while being still annotation-efficient. As such, we empirically show that our method requires 6 times less annotations compared to a well-tuned SL model.

5 Broader Impact

DL-based methods are criticized for being data-greedy, which makes small- and medium-scale companies being less competitive compared to the large ones. Our work shows that one can use novel techniques and decrease the requirements for annotated data by a large margin. We hope that the results presented in this paper will be of value to practitioners who aim to develop DL systems for medical imaging applications and beyond.

Acknowledgments

The OAI is a public-private partnership comprised of five contracts (N01- AR-2-2258; N01-AR-2-2259; N01-AR-2- 2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health.

The MOST is comprised of four cooperative grants (Felson - AG18820; Torner - AG18832; Lewis - AG18947; and Nevitt - AG19069) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by MOST study investigators. This manuscript was prepared using MOST data and does not necessarily reflect the opinions or views of MOST investigators.

We would like to acknowledge the strategic funding of the University of Oulu, KAUTE foundation and Sigrid Juselius Foundation, Finland.

Dr. Claudia Lindner is acknowledged for providing BoneFinder. Phuoc Dat Nguyen is acknowledged for discussions about *mixup*.

References

[1] Ali Mobasheri and Mark Batt. An update on the pathophysiology of osteoarthritis. *Annals of physical and rehabilitation medicine*, 59(5-6):333–339, 2016.

[2] Bart S Ferket, Zachary Feldman, Jing Zhou, Edwin H Oei, Sita MA Bierma-Zeinstra, and Madhu Mazumdar. Impact of total knee replacement practice: cost effectiveness analysis of data from the osteoarthritis initiative. *bmj*, 356, 2017.

[3] Aleksei Tiulpin, Stefan Klein, Sita MA Bierma-Zeinstra, Jérôme Thevenot, Esa Rahtu, Joyce van Meurs, Edwin HG Oei, and Simo Saarakkala. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific reports*, 9(1):1–11, 2019.

[4] Joseph Antony, Kevin McGuinness, Noel E O'Connor, and Kieran Moran. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1195–1200. IEEE, 2016.

[5] Joseph Antony, Kevin McGuinness, Kieran Moran, and Noel E O'Connor. Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *International conference on machine learning and data mining in pattern recognition*, pages 376–390. Springer, 2017.

[6] Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports*, 8(1):1–10, 2018.

[7] Huy Hoang Nguyen, Simo Saarakkala, Matthew Blaschko, and Aleksei Tiulpin. Semixup: In-and out-of-manifold regularization for deep semi-supervised knee osteoarthritis severity grading from plain radiographs. *IEEE Transactions on Medical Imaging*, 2020.

[8] JH Kellgren and Frida Bier. Radiological signs of rheumatoid arthritis: a study of observer differences in the reading of hand films. *Annals of the rheumatic diseases*, 15(1):55, 1956.

[9] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[11] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *arXiv preprint arXiv:1903.03825*, 2019.

[12] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[13] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.