DEEP LEARNING FOR RHEUMATOID ARTHRITIS: JOINT DETECTION AND DAMAGE SCORING IN X-RAYS

Krzysztof Maziarz[*]
Microsoft Research

Anna Krason
University College London

Zbigniew Wojna
Tensorflight

ABSTRACT

Recent advancements in computer vision promise to automate medical image analysis. Rheumatoid arthritis is an autoimmune disease that would profit from computer-based diagnosis, as there are no direct markers known, and doctors have to rely on manual inspection of X-ray images. In this work, we present a multi-task deep learning model that simultaneously learns to localize joints on X-ray images and diagnose two kinds of joint damage: narrowing and erosion. Additionally, we propose a modification of label smoothing, which combines classification and regression cues into a single loss and achieves 5% relative error reduction compared to standard loss functions. Our final model obtained 4th place in joint space narrowing and 5th place in joint erosion in the global RA2 DREAM challenge.

1 INTRODUCTION

Rheumatoid arthritis (RA) is an autoimmune disease which commonly affects joints in hands, wrists and feet, and can cause chronic pain (CDC, 2020). One of the impediments to its efficient diagnosis is the lack of direct markers for RA; rheumatologists have to rely on various clinical clues in order to determine whether a patient should start antirheumatic therapy (Fukae et al., 2020).

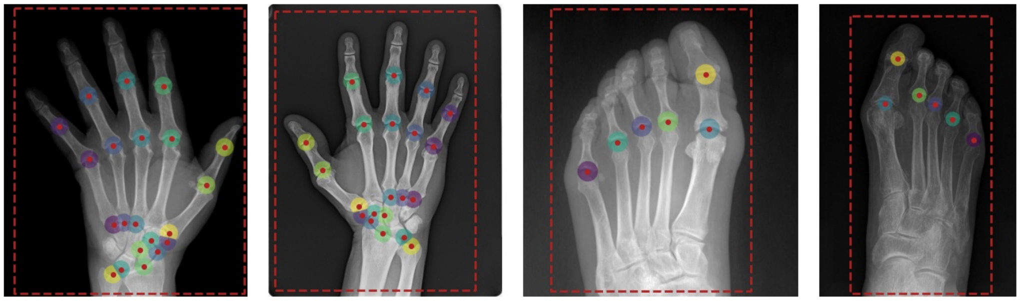

Figure 1: Four example images from the training dataset. The dashed box shows the region of interest computed during preprocessing, dots mark the ground-truth center of each joint, and colored regions correspond to the ground-truth segmentation mask for joint detection.

Joint damage scoring is typically done by a highly-trained medical professional, who has to meticulously review several radiographic images of hands and feet. This procedure is not only costly and time-consuming, but also gives subjective results. The need for medical expertise may delay access to diagnosis; this is especially true in developing countries, where hospitals are often understaffed.

Recent advancements in deep learning promise to address these challenges, by *learning* to assess joint damage directly from data. Early attempts include determining whether the patient has RA using various clinical data (Fukae et al., 2020), and diagnosing RA by training a convolutional neural network (CNN) on a radiographs dataset (Üreten et al., 2020).

One established metric to quantify joint damage is the Sharp/van der Heijde (SvH) method (Van der Heijde et al., 1995). This scoring system separately evaluates two kinds of damage: joint space

[*]Correspondence to krzysztof.maziarz@microsoft.com

narrowing and joint erosion[1]. For narrowing, 15 joints in each hand and 6 joints in each foot are given scores from 0 to 4. For erosion, 16 joints in each hand are given scores from 0 to 5, while 6 joints in each foot are scored from 0 to 10. Finally, the overall SvH score is the sum of all narrowing and erosion scores, which is an integer ranging from 0 to 448.

In this paper, we propose a deep multi-task (Caruana, 1997) neural architecture that predicts joint narrowing and erosion scores following the SvH method. We make the following contributions:

- We design a deep CNN architecture to estimate SvH scores for RA damage. It simultaneously performs *joint localization*, *joint erosion assessment* and *joint narrowing assessment*.
- We propose *local label smoothing*, which includes class order in the cross entropy classification loss. With local label smoothing, we achieve a 5% relative error reduction.

2 Related work

There have been many approaches to automated RA diagnosis, both using deep CNNs, and classical machine learning techniques such as support vector machines (SVM). These works can be divided into those exploring a coarse-grained task of classifying patients as either normal or suffering from RA, and those that produce fine-grained SvH scores for all joints.

Detecting RA as binary classification Several works tried to predict whether a patient has RA by training a classification network on some patient data. Fukae et al. (2020) used various clinical information converted into an image, which was then processed by an AlexNet (Krizhevsky et al., 2012). In contrast, other approaches used imaging data directly, either X-ray (Üreten et al., 2020) or ultrasound (Andersen et al., 2019). However, modelling RA diagnosis with a simple binary variable lacks explainability, which is essential for clinical adoption (Amann et al., 2020).

Detecting RA using the SvH method Building models that predict fine-grained SvH scores can lead to a much richer interplay between machine learning models and medical practitioners. Early works have explored this direction using simple machine learning models, such as SVMs (Tashita et al., 2017) and shallow CNNs (Hirano et al., 2019). Recently, modern CNN architectures (Li & Guan, 2020; Israel, 2020; Pataki & Olar, 2020; Dimitrovsky & Ericson, 2020; Stadler & Shi, 2020; Tran & Nguyen, 2020) have been used to predict SvH scores as part of the RA2 Dream Challenge (Bionetworks, 2021). As our method has also been submitted to this challenge, these results can be directly compared to ours, and they have been found to achieve similar results on the challenge test data. We differentiate ourselves from these methods in more detail in Section 4.

3 Our method

We approach SvH scoring of hand and foot images as simultaneous segmentation and classification. For training, we utilize four images per patient, with all narrowing and erosion scores annotated for each image. In theory, we could train a model to predict all scores from raw pixel data without localization cues. However, data in the medical domain is typically scarce, making this approach unfeasible, as it is not sample efficient. Therefore, we also assume access to annotations of center positions of all joints, and use that as an additional training signal.

Segmentation masks During data preprocessing, we convert joint center annotations into a segmentation mask, associating some pixels with the class tied to the corresponding joint, and some with an additional background class. As the optimal size of the relevant joint region is unclear, during the labelling stage we only annotate joint centers, and convert them to segmentation masks during preprocessing. In this way, we decouple generating the ground truth segmentation mask from annotating the data.

We parametrize the conversion of joint centers into a mask with two hyperparameters r and R, where $r \leq R$. First, for each pixel we find the closest joint center. If that closest center is at distance at most r, we use it as the ground truth class for that pixel. If this distance is more than R, we associate

[1] For more background information about the two kinds of joint damage see Appendix A.

the given pixel with the background class. Pixels where the distance to the closest center falls into the $\langle r, R\rangle$ range we consider to lie on a boundary; we therefore *do not assign them to any class*, and ignore them in the segmentation loss.

Training objective We train our network to minimize a simple weighted average of three per-pixel objectives. First, for each pixel, we predict the corresponding joint class. In hands, 15 joints are relevant for narrowing and 16 joints for erosion; these two sets largely overlap apart from some joints in the wrist, yielding a total of 21 points of interest (Figure 1, left). As there is a natural mapping between toes and fingers, we map the 6 feet joints (Figure 1, right) into the shared set of 21 joint types. Together with the background class, the total number of classes for the joint localization head is 22. Next, we classify each positive (joint) pixel to predict a narrowing score (5 possible values) and an erosion score (6 values). Intuitively, this multi-task formulation decouples joint localization from damage assessment, as narrowing and erosion is detected with the same network head, irrespective of the underlying joint. All three prediction tasks are modelled as classification.

As erosion in feet is scored on a 0-10 scale, we train the model to match *half* of the actual score, and then multiply the predictions by 2 during inference. In that way, the range of values for erosion scores is aligned across all joints; this is crucial as damage prediction is joint-agnostic.

Local label smoothing As explained in the previous section, we model the prediction of narrowing and erosion scores as a classification task. Casting the prediction of a small integer as classification instead of (constrained) regression is common practice in many machine learning models, as such approach is typically robust and easy to train (Kozakowski et al., 2019; Wojna et al., 2020). In preliminary experiments, we found that training for classification worked significantly better than regression. However, using classification ignores the inherent ordering of the classes: predicting a narrowing score of 1 is a much better answer if the ground truth is 0, than if the ground truth is 4.

In this work, since SvH scores are small integers, we propose to model the task as classification, but inject a small regression-inspired bias into the ground truth label. Similarly to label smoothing (Müller et al., 2019), we use a smoothed one-hot vector as ground truth, placing most of the probability mass on the target class, and a small amount of mass on the other classes. However, in contrast to classical label smoothing, we only move probability mass to neighbouring classes: if the ground truth label is x, we place extra probability mass on classes $x - 1$ and $x + 1$ ($\frac{p}{2}$ each), and the remaining $1 - p$ on x. Note that x can be one of the boundary classes (either 0 or the maximum possible score), in which case $1 - \frac{p}{2}$ of probability mass is placed on the ground truth class x.

Model We utilise the U-Net architecture, which was shown to achieve strong performance on biomedical image segmentation (Ronneberger et al., 2015; Çiçek et al., 2016; Schlemper et al., 2019). As the encoder, we use the EfficientNet B5 network (Tan & Le, 2019); for the decoder we use traditional upsampling convolutional layers (Zeiler et al., 2011). Following common practice for U-Nets, we add skip connections from encoder layers into decoder layers at a matching resolution. As the pixel location within the image can aid the classification into joint types, we use the CoordConv technique (Liu et al., 2018), and concatenate the input image with two additional channels, corresponding to x and y pixel coordinates.

The final feature map from the last decoder layer is used as input to three separate per-pixel classifiers, which give the output for each of the three tasks specified earlier. Therefore, our architecture supports multiple tasks via the *shared bottom* paradigm, as most of the network is shared. This works well since the tasks are highly related; however, more complex methods could explicitly trade off positive and negative transfer (Ma et al., 2019; Maziarz et al., 2019; Zhang et al., 2020).

During inference we produce the final narrowing and erosion scores as a weighted average, using the softmax predictions produced by the model as weights.

4 EXPERIMENTS

Training data We obtained the training data from the RA2 DREAM Challenge, which contains data from 367 patients, with four images per patient (both hands and feet). We split the provided training data into train and validation sets by using the first out of eight folds as the validation set

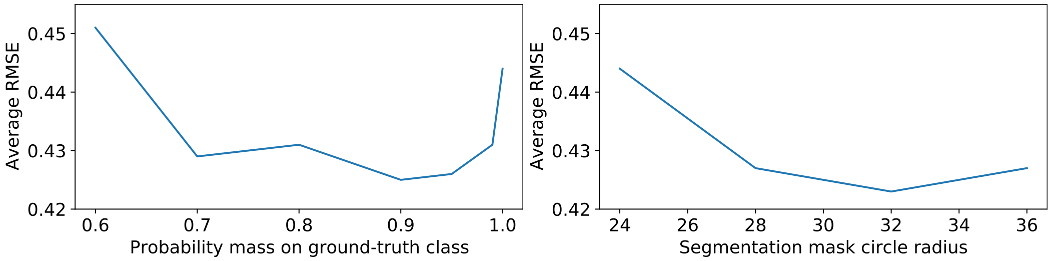

Figure 2: The effect of varying local label smoothing (left) and segmentation circle size (right) on downstream model performance.

and the rest of the data as the training set. We show four example images from the training set in Figure 1.

The input images come with a variable amount of irrelevant black pixels at the borders, while only the pixels that correspond to the hand or foot are informative. To further normalize the data, we used classical computer vision techniques such as the canny algorithm to detect the bounding box of the hand or foot visible in the image. We then cropped each image to the relevant bounding box, and scaled the resulting image to a resolution of 864 x 928. In Figure 1, we mark the bounding box computed for the example images with dashed rectangles. As the amount of data was limited, we also employed several data augmentation techniques: rotation, scaling and horizontal flips.

Results We trained the network using the AdamW optimizer (Loshchilov & Hutter, 2017) and OneCycle learning rate scheduling method (Smith, 2017). We tuned all hyperparameters by performing small-scale experiments on our local validation set. We compared experiments by computing the root mean squared error (RMSE) of SvH score predictions for all joints, averaged over all examples in the validation set.

After hyperparameter tuning, we used a well performing setting to train our model 8 times on all available data. We used an ensemble of these 8 models to predict SvH scores on unseen test data, resulting in mean RMSE of 0.4075 for narrowing, and 0.4607 for erosion. On the final leaderboard, these values fall within 10% relative difference of the top result, obtaining 4th and 5th place for narrowing and erosion, respectively. Small gains that other models submitted to the challenge achieved on top of our results (Li & Guan, 2020; Israel, 2020; Pataki & Olar, 2020) can be attributed to using much more complex pipelines (with up to three separate steps, each utilizing a different ML model), and sometimes performing extensive human labour (by authors learning the basics of RA scoring, and then manually re-annotating the data to obtain more labels). In contrast, we show that a simple architecture trained in a multi-task setting can obtain results competitive with the state-of-the-art. See Appendix B for a detailed comparison of top performing models in the challenge.

Ablation study To better understand which hyperparameters are important to achieve good performance, we performed an ablation study on the validation set. While many hyperparameters had limited impact on downstream performance or were easy to tune, we found two that showed interesting trends: the degree of local label smoothing, and the radius r used to compute joint segmentation masks.

We show the ablation results in Figure 2. We see that shifting 0.1 of the probability mass to classes adjacent to the ground truth improves results, yielding a relative error rate reduction of approximately 5%. Moreover, we see that to achieve optimal performance r has to be tuned rather carefully, with the optimal value around $r = 32$ (given a fixed value of $R = 40$).

5 Conclusion

In this work, we have shown a deep neural model which simultaneously localizes joints and assesses the severity of rheumatoid arthritis. This was enabled by posing both objectives as pixel-level classification tasks, and training the model in a multi-task setting. Moreover, we proposed local label smoothing, which allows to smoothly interpolate between a classification and a regression objective.

References

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):1–9, 2020.

Jakob Kristian Holm Andersen, Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Kathrine Holtz, Jakob Grauslund, Thiusius Rajeeth Savarimuthu, and Søren Andreas Just. Neural networks for automatic scoring of arthritis disease activity on ultrasound images. *RMD open*, 5(1):e000891, 2019.

Sage Bionetworks. Ra2 dream challenge, 2021. URL `https://www.synapse.org/#!Synapse:syn20545111/wiki/594083`.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

CDC. Rheumatoid arthritis — arthritis, cdc. `https://www.cdc.gov/arthritis/basics/rheumatoid-arthritis.html`, 2020.

Neelambuj Chaturvedi. Deepra: Predicting joint damage from radiographs using cnn with attention. *arXiv preprint arXiv:2102.06982*, 2021.

Srinivas Chilukri. Automatic joint damage measurement using computer vision on radiographic data, 2020.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pp. 424–432. Springer, 2016.

Isaac Dimitrovsky and Lars Ericson. A multistage deep learning method for scoring radiographic hand and foot joint damage in rheumatoid arthritis. 08 2020. doi: 10.7303/syn21478998.

Jun Fukae, Masato Isobe, Toshiyuki Hattori, Yuichiro Fujieda, Michihiro Kono, Nobuya Abe, Akemi Kitano, Akihiro Narita, Mihoko Henmi, Fumihiko Sakamoto, et al. convolutional neural network for classification of two-dimensional array images generated from clinical information may support diagnosis of rheumatoid arthritis. *Scientific reports*, 10(1):1–7, 2020.

Toru Hirano, Masayuki Nishide, Naoki Nonaka, Jun Seita, Kosuke Ebina, Kazuhiro Sakurada, and Atsushi Kumanogoh. Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. *Rheumatology advances in practice*, 3(2): rkz047, 2019.

Ariel Yehuda Israel. Prediction of rheumatoid arthritis scores ariel's method (prasam), 2020. URL `https://www.synapse.org/#!Synapse:syn21499370/wiki/604451`.

Piotr Kozakowski, Łukasz Kaiser, and Afroz Mohiuddin. Forecasting deep learning dynamics with applications to hyperparameter tuning. 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Hongyang Li and Yuanfang Guan. Ra2 dream challenge solution, 2020. URL `https://www.synapse.org/#!Synapse:syn21680642/wiki/604549`.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in neural information processing systems*, pp. 9605–9616, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiaqi Ma, Zhe Zhao, Jilin Chen, Ang Li, Lichan Hong, and Ed H Chi. Snr: Sub-network routing for flexible parameter sharing in multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 216–223, 2019.

Krzysztof Maziarz, Efi Kokiopoulou, Andrea Gesmundo, Luciano Sbaiz, Gabor Bartok, and Jesse Berent. Flexible multi-task networks by learning parameter allocation. *Workshop on Neural Architecture Search at ICLR*, 2019.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pp. 4694–4703, 2019.

Balint Armin Pataki and Alex Olar. Ra2 dream challenge solution by team csabaibio, 2020. URL `https://www.synapse.org/#!Synapse:syn21614092`.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472. IEEE, 2017.

Michael Stadler and Chenfu Shi. A two-stage model to classify joint damage in radiographs, 2020. URL `https://www.synapse.org/#!Synapse:syn21610007/wiki/604496`.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

Atsuki Tashita, Kento Morita, Manabu Nii, Natsuko Nakagawa, and Syoji Kobashi. Automated estimation of mts score in hand joint x-ray image using machine learning. In *2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT)*, pp. 1–5. IEEE, 2017.

Duc Tran and Tin Nguyen. Ra2 dream challenge solution by team nad, 2020. URL `https://www.synapse.org/#!Synapse:syn21587234/wiki/604514`.

Kemal Üreten, Hasan Erbay, and Hadi Hakan Maraş. Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network. *Clinical rheumatology*, 39(4):969–974, 2020.

DMFM Van der Heijde, MA Van Leeuwen, PLCM Van Riel, and LBA Van de Putte. Radiographic progression on radiographs of hands and feet during the first 3 years of rheumatoid arthritis measured according to sharp's method (van der heijde modification). 1995.

Zbigniew Wojna, Krzysztof Maziarz, Łukasz Jocz, Robert Pałuba, Robert Kozikowski, and Iasonas Kokkinos. Holistic multi-view building analysis in the wild with projection pooling. *arXiv preprint arXiv:2008.10041*, 2020.

Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*, pp. 2018–2025. IEEE, 2011.

Wen Zhang, Lingfei Deng, and Dongrui Wu. Overcoming negative transfer: A survey. *arXiv preprint arXiv:2009.00909*, 2020.

A Background on joint damage assessment

Rheumatoid arthritis is typically associated with two kinds of joint damage: *joint space narrowing* and *bone erosion*. Narrowing affects the joint cartilage, which ensures that the distance between interacting bones allows for a good range of motion. Narrowing means that the cartilage can no longer maintain a healthy distance between the bones, which leads to increased pressure and friction. Erosion is related to *bone resorption*, which is a process in which the body breaks down bone tissue, releasing the minerals into the bloodstream. While normally this is a part of a healthy life cycle of bone tissue, it occurs excessively in patients suffering from rheumatoid arthritis, leading to irreversible damage.

B Comparison of top submissions to the RA2 DREAM challenge

In Table 1 we analyse the best performing submissions to the RA2 DREAM challenge, comparing them both in terms of performance and complexity. We see that all competing methods use pipelines containing at least two steps, which typically are joint detection followed by extracting image patches and separately scoring them. Some approaches also include a third stage, which is either post-processing to combine all results for a single patient, or pre-processing to normalize the orientation of the input images. Additionally, Li & Guan (2020) ”served as an extra newbie radiologist to manually score all the training data”, therefore making use of more labels than other approaches. In contrast, our method only needs a single model trained in a multi-task setting, and did not use extra labels, while still attaining performance close to the state-of-the-art.

Table 1: Comparison of top results in the RA2 DREAM challenge. We report the number of pipeline steps (each corresponding to a separately trained ML model or a group of models), narrowing and erosion results (both as RMSE and absolute rank on the final leaderboard), and whether the method relied on additional training labels.

Method	Narrowing	Erosion	Pipeline steps	Extra labels
(Li & Guan, 2020)	**0.3758 (1)**	0.4464 (3)	3	Yes
(Israel, 2020)	0.3924 (2)	**0.4303 (1)**	2	
(Pataki & Olar, 2020)	0.3991 (3)	0.4355 (2)	3	
(Dimitrovsky & Ericson, 2020)	0.4313 (6)	0.4540 (4)	3	
(Stadler & Shi, 2020)	0.4132 (5)	0.4660 (6)	2	
(Tran & Nguyen, 2020)	0.4409 (7)	0.4679 (7)	2	
(Chaturvedi, 2021)	0.4424 (8)	0.4900 (8)	2	
(Chilukri, 2020)	0.4813 (9)	0.5096 (9)	2	
Ours	0.4075 (4)	0.4607 (5)	**1**	