Optimizing labeling

The limits of weakly supervised learning in osteophyte severity grading and localisation in 4 hip quadrants

AUTHORS

Alvin Ye[1], a.d.ye@student.tudelft.nl
Supervisors: Gijs van Tulder[1], Jesse Krijthe[1]

AFFILIATIONS

Delft University of Technology[1]

Severity grade 0 through 3 of superior femoral osteophyte based on the OARSI severity scale

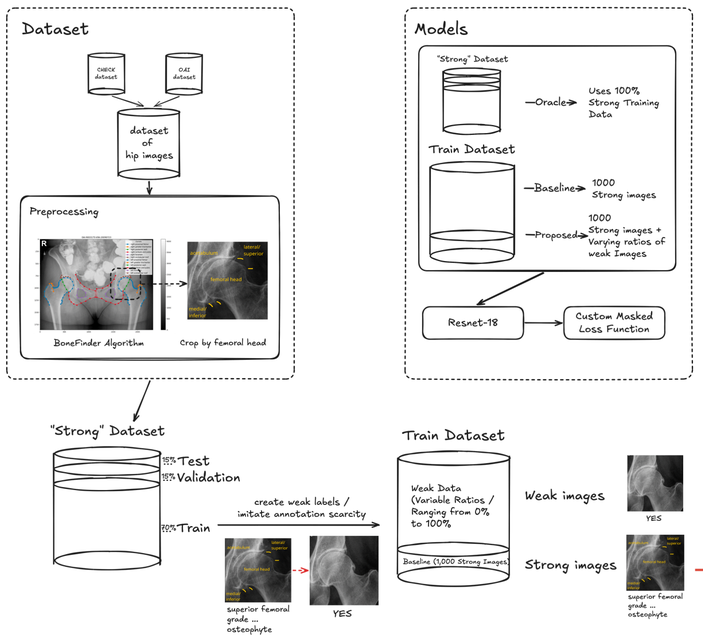
1. THE 500 MILLION PROBLEM

- Osteoarthritis (OA) is a common disease that develops osteophytes.
- In 2020, about 595 million people had a form of osteoarthritis.
- Detailed osteophyte annotation in X-Ray Images is time-consuming and costly requiring an expert.
- Consequently, clinical datasets are expensive to scale with annotated data.
- We categorize weak data with a binary indicator of osteophytes in the entire image. We categorize strong data as expert annotated severity grade in the 4 hip joint locations
- It is unknown to what extent adding weak data continues to help grading severity and localising osteophytes.

THE MAIN QUESTION

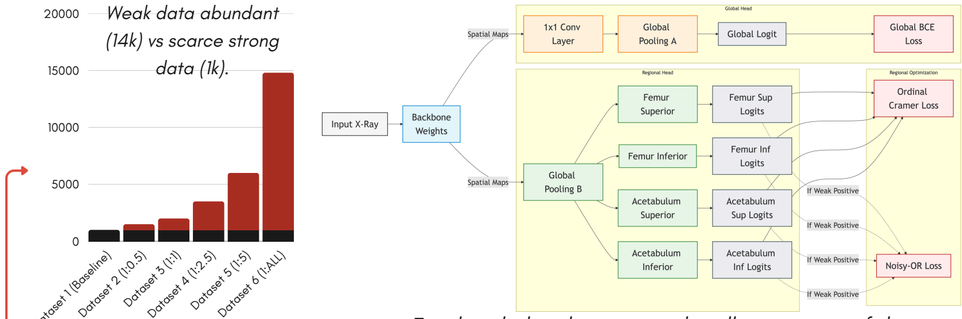
> How do different weak-to-strong data mixes change performance in a multi task machine learning pipeline and specifically when does it stop improving?

2. TRAINING DATA



Preprocessing the datasets to create a Train/Validation/Test purely on strong images. The Train split is then modified to become 1000 strong images + remaining becomes weak.

3. FOUR CONFIGURATIONS



Two headed architecture to handle two types of data

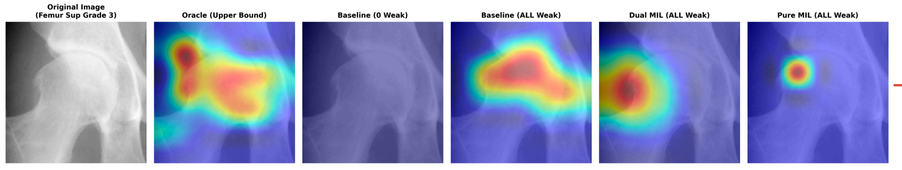
We evaluate 4 model configurations across all six datasets:

- Oracle Model (Upper Bound): Trained on 100% strong data using the dual-head pipeline.
- Masked Baseline: Dual-head pipeline, but ignores weak positives (uses [0,0,0,0] ground truth for healthy joints).
- Dual-Head MIL (Proposed): Full architecture capabilities, leveraging Noisy-OR Loss.
- Pure MIL Ablation: Regional head only (isolates and verifies the dual-head benefit).
- Quadratic Weighted Kappa as metric indicating agreement with experts (0 = random 1 = perfect agreement). QWK score is lower when predictions fall further from true grade.

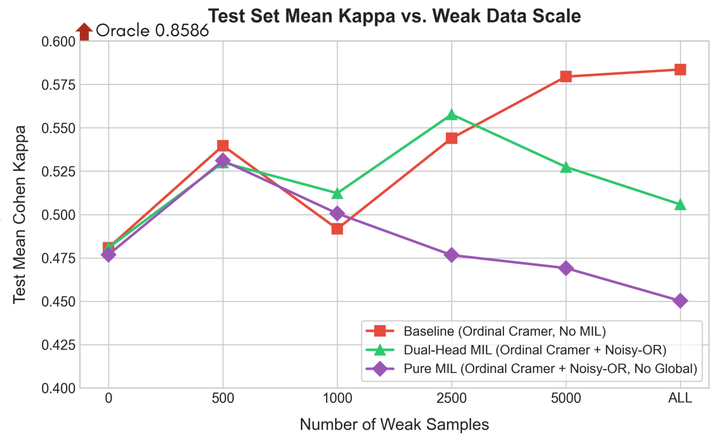
4. RESULTS & KEY FINDINGS

Model	ρ (Weak Samples)	Test Mean Kappa
Oracle	100% Strong	0.8586
Masked Baseline	0	0.4809
Masked Baseline	ALL (∼15k) (Peak)	**0.5836**
Dual-Head MIL	0	0.4809
Dual-Head MIL	1000 (Val-Selected Peak)	0.5123
Dual-Head MIL	2500 (Test-Absolute Peak)	0.5577
Dual-Head MIL	ALL (∼15k)	0.5058
Pure MIL	0	0.4769
Pure MIL	500 (Peak)	0.5311
Pure MIL	ALL (∼15k)	0.4502

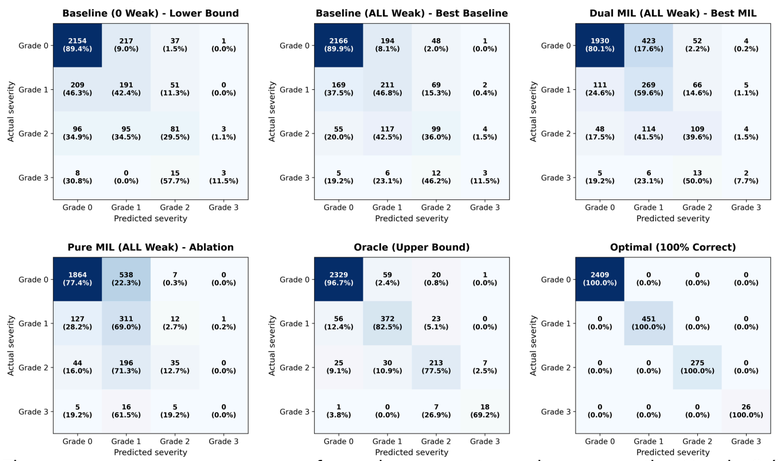
Masked baseline scales as data scales to the maximum (Kappa 0.48 → 0.58). MIL peak early then degrade. Validation selected model does not contain the highest test kappa.



Grading-vs-localisation trade-off: masked baseline grades best but shows diffusely like the Oracle, Dual-MIL localises tightly but grades worse. Pure MIL misses the correct area



Scaling data ratio shows that Pure MIL Kappa peaks early and consistently drops, whereas Dual-Head peaks later but also subsequently drops. Masked Baseline improves but sees an anomaly at 1000 with the other networks



The gain in Kappa score comes from clean negatives sharpening the Grade 0/1 boundary not from learning osteophytes (masked never sees a weak positive). MIL models struggle with predicting the correct grade due to missing severity grading information beyond Grade 1

5. LIMITATIONS

- Dataset is skewed, only few severe cases to train on
- Single seed
- Best model ignores weak positive leaving good data unused
- Qualitative GradCAM instead of quantitative

6. CONCLUSION & FUTURE WORK

- Weak labels provide better prediction, only when using healthy samples
- Best grading performance is not always best localisation performance
- Missing severity labels provide noisy data dropping performance
- Vision-language extraction of rich features (severity) may be a solution
- Future work can constrain spatial analysis using anatomical priors
- Future work should verify results statistically