Classifying Treatment Effectiveness in Chronic Recurrent Multifocal Osteomyelitis from MRIs

Anna Merkoulovitch,[1] Zach Wener-Fligner[1]
{annamerk, zbwener}@stanford.edu
[1]Stanford Center for Professional Development (SCPD), Stanford University

Stanford Computer Science

Abstract

Chronic Recurrent Multifocal Osteomyelitis (CRMO) is a rare condition mainly affecting the distal regions of long bones in the body including the femur and tibula [1]. We classify progression of CRMO by considering pairs of MRI images of the knee and long bones of the leg. In this approach, we train multiple classifiers: a logistic classifier with features extracted using a pre-trained Inception-v3 CNN and an SVM classifier on a *bag of visual words.* We use ensemble voting to combine these models and present results for both multi-class (*improved*; *persisted*; and *regressed*) and binary classes (*improved*; and *persisted*/*regressed*).

Data Preparation

Whole body MRIs of 45 CRMO patients (containing multiple scans per patient) were procured from the Bristol Royal Hospital. Original scans were visually inspected to select partial slices of the knee and leg with visible bone structure. Before and after scan pairings and corresponding class were manually curated from radiological assessments. Data was augmented by reversing image pair order and programmatically adding noise.

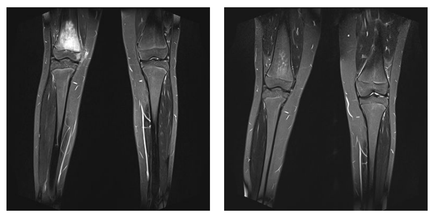

Fig 1. Before and after images of typical MRI appearances of improved CRMO condition after treatment with pamidronate therapy (7 months apart).

Models/Methods

We developed two models by extracting different feature sets. Models were trained on a 70/30 training and development split, with K-fold cross validation. A voting ensemble was used to combine the two methods as shown in figure 2.

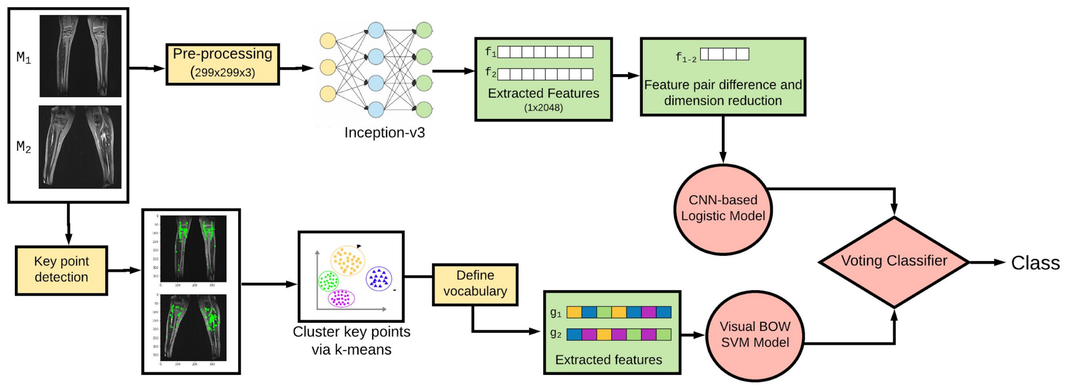

Fig 2. Architecture for selected approach to classifying pairs of patient MRIs to assess disease progression.

Transfer Learning

Deep learning is the current state of the art for image classification. Pre-trained models enable quality feature generation where training a custom model is infeasible due to small data.

CNN Feature Extraction

- Pre-processed images to have size 299x299x3.
- Features extracted from final layer of Inception-v3 convolutional neural network [2].
- Low Σ^2 feature reduction.
- Pairs of images represented as the difference between image feature vectors (399 features).

Softmax with Regularization

- L2 regularization to prevent overfitting.
- Cross-entropy cost function:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{K=1}^{3} y_K^{(i)} \log h_\theta(x^{(i)}) + \lambda||\theta||_2^2$$

Bag of Visual Words

We modify a *bag of visual words* approach [3] for use with pairs of images; we then train SVM with radial basis function and Naive Bayes classifiers on this data.

Bag of Visual Words

- Using ORB, SURF, and SIFT, extract features for each image.
- Build *visual vocabulary* by clustering on the union of features across all images:

$$V = \{\mu_1, ..., \mu_n\}$$

where the resulting *visual words* are obtained via running k-means updates to convergence:

$$c^{(i)} := \arg\min_j ||x^{(i)} = \mu_j||^2; \mu_j := \frac{\sum_{i=1}^{m} \mathbb{I}\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} \mathbb{I}\{c^{(i)} = j\}}$$

- Represent a pair of images as a vector in $\mathbb{R}^{2n}$ obtained by concatenating the *bag of words* representation for each image.

Results

We show results for a subset of models which were relatively high-performing in cross-validation. Final test error is shown for all models for completion.

Model	Training Error	Dev Error	Test error
Multi-class			
CNN Softmax	0.08	0.31	0.64
NB-BOW, SIFT, |V| = 500	0.0	0.33	0.78
Ensembled model	0.05	0.37	0.58
Binary			
CNN Logistic	0.05	0.12	0.29
SVM-BOW, ORB, |V| = 50	0.0	0.63	0.83
Ensembled Model	0.11	0.22	0.29

* Error = 1 - f1-score. |train| = 57, |dev| = 25, |held-out test set| = 7

Discussion and Future Work

Our ensemble approach produced models which were more stable and lower-variance than the constituent models in the binary case. In the multi-class case, however, our models failed to generalize past the training data. In particular, models failed to predict the *persisted* class, which may be due to lack of data for this class.

Noisy data and small dataset size made it difficult for models to learn useful information without succumbing to high variance. Visual inspection of incorrectly-classified examples also suggests potential human error in curation. We plan to increase robustness by combining multiple radiologist assessments in the future.

Developing heuristics for key point usefulness could decrease variance by minimizing the feature set. We would like to expand on our transfer learning approach by using models trained specifically on WB-MRI data as opposed to general-purpose image classifiers.

References

[1] Roderick et. al. (2016). Chronic recurrent multifocal osteomyelitis (CRMO) - advancing the diagnosis. *Pediatric Rheumatology.*, 14:47. doi: 10.1186/s12969-011-0109-1
[2] Szegedy, C., et. al. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).
[3] Sivic, J., & Zisserman, A. (2006). Video Google: Efficient visual search of videos. In *Toward category-level object recognition* (pp. 127-144). Springer, Berlin, Heidelberg.