DOI: https://doi.org/10.17816/DD430358

Моделирование индивидуального стиля разметки врача-рентгенолога для улучшения точности нейронных сетей

Е.Д. Никитин

Медицинские Скрининг Системы, Санкт-Петербург, Российская Федерация

АННОТАЦИЯ

Обоснование: одна из часто встречающихся проблем при разметке медицинских изображений — расхождения между разными врачами-разметчиками (inter-observer variability). Одно и то же изображение может быть размечено врачами по-разному. Основные причины — человеческий фактор, разница в опыте и квалификации, разные «радиологические школы», плохое качество изображения, нечёткие инструкции. Влияние некоторых факторов можно уменьшить за счёт правильной организации процесса разметки, но мнение врачей всё равно нередко различается.

Цель: проверить, может ли нейронная сеть с помощью дополнительного модуля обучиться стилю и особенностям разметки разных врачей-рентгенологов и может ли такое моделирование улучшить итоговые метрики детекции объектов на радиологических изображениях.

Методы: для обучения систем искусственного интеллекта в радиологии часто используется перекрёстная разметка, т.е. разметка одного и того же исследования несколькими врачами. Существуют разные методы работы с такой разметкой. Самый простой способ — использовать разметку от каждого врача как независимый пример при обучении модели. Есть методы, которые с помощью различных правил или алгоритмов объединяют разметку перед обучением. Наконец, M.Y. Guan и соавт. используют отдельные классификационные головы для моделирования стиля разметки разных врачей. К сожалению, этот метод не подходит для более сложных задач, например детекции объектов на изображении. Для данного анализа использовалась модель машинного обучения, предназначенная для детекции объектов различных классов на маммографических исследованиях. Эта модель является нейронной сетью, основанной на архитектуре Deformable DETR. Для обучения нейронной сети использовался набор данных, состоящий из 7756 маммографических исследований молочных желёз и 12 543 уникальных аннотации от 19 врачей; для валидации — набор данных, состоящий из 700 исследований, размеченных по шкале Bi-Rads; для тестирования — набор данных, состоящий из 300 исследований, размеченных по шкале Bi-Rads. Во всех наборах данных доля исследований с патологией находилась в диапазоне 15–20%. Каждому из 19 врачей был присвоен уникальный индекс, по которому специальный модуль на каждой итерации обучения нейронной сети находит соответствующий этому индексу вектор. Этот вектор расширялся до размера карты признаков каждого уровня пирамиды признаков, а затем присоединялся отдельными каналами к этим картам. Таким образом, энкодер и декодер детектора получали доступ к информации о том, какой врач разметил данное исследование. Векторы обновлялись с помощью метода обратного распространения ошибки. Для сравнения были выбраны три метода.

1. Базовая модель — объединение разметки методом разных врачей с помощью метода "голосования".
2. Использование нового стилистического модуля. Для предсказаний на тестовом наборе данных был использован индекс одного врача, который показал лучшие метрики на валидационном наборе данных.
3. Использование нового стилистического модуля. Для предсказаний на тестовом наборе данных были использованы индексы пяти врачей с лучшими метриками на валидационном наборе данных. Для объединения предсказаний применялся метод Weighted Boxes Fusion.

В качестве основной метрики был использован ROC-AUC на тестовом наборе данных (к патологии относились категории Bi-Rads 3, 4, и 5). В качестве вероятности злокачественности для каждого метода использовалась сумма максимальных вероятностей обнаруженных злокачественных объектов (злокачественные образования и кальцинаты) по проекциям CC и MLO.

Результаты: по результатам обучения были получены следующие метрики ROC-AUC для трёх приведённых методов — 0,82; 0,87 и 0,89.

Заключение: использование информации о враче-разметчике позволяет нейронной сети эффективнее обучаться и моделировать стиль разметки разных врачей. Данный метод может также применяться для получения оценки не-

Рукопись получена: 15.05.2023 **Рукопись одобрена:** 05.06.2023 **Опубликована Online:** 10.07.2023

уверенности сети в своём предсказании. Использование эмбеддингов разных врачей, приводящее к разным предсказаниям, может означать сложность данного исследования для обработки системой искусственного интеллекта.

Ключевые слова: машинное обучение; разметка данных.

КАК ЦИТИРОВАТЬ

Никитин Е.Д. Моделирование индивидуального стиля разметки врача-рентгенолога для улучшения точности нейронных сетей // *Digital Diagnostics*. 2023. Т. 4, № 1 Supplement. С. 99–101. DOI: https://doi.org/10.17816/DD430358

СПИСОК ЛИТЕРАТУРЫ

1. Jensen M.H., Jørgensen D.R., Jalaboi R., Hansen M.E., Olsen M.A. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement // Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Vol. 11767. Cham : Springer, 2019. P. 540–548. doi: 10.1007/978-3-030-32251-9_59

2. Jungo A., Meier R., Ermis E., et al. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation // Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Vol. 11070. Cham : Springer, 2018. P. 682–690. doi: 10.1007/978-3-030-00928-1_77

3. Guan M.Y., Gulshan V., Dai A.M., Hinton G.E. Who Said What: Modeling Individual Labelers Improves Classification // Proceedings of the AAAI Conference on Artificial Intelligence. 2017. Vol. 32, N 1. doi:10.1609/aaai.v32i1.11756

4. Zhu X., Su W., Lu L., et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection // arXiv:2010.04159. doi: 10.48550/arXiv.2010.04159

5. Solovyev R., Wang W., Gabruseva T. Weighted boxes fusion: Ensembling boxes from different object detection models // Image and Vision Computing. 2021. Vol. 107. P. 104117. doi: 10.1016/j.imavis.2021.104117

DOI: https://doi.org/10.17816/DD430358

Learning radiologists' annotation styles with multi-annotator labeling for improved neural network performance

Evgeniy D. Nikitin

Medical Screening Systems LLC, Saint Petersburg, Russian Federation

ABSTRACT

BACKGROUND: One of the common problems in labeling medical images is inter-observer variability. The same image can be labeled differently by doctors. The main reasons are the human factor, differences in experience and qualifications, different "radiology schools", poor image quality, and unclear instructions. The influence of some factors can be reduced by proper organization of the annotation; however, the opinion of doctors frequently differs.

AIM: The study aimed to test whether a neural network with an additional module can learn the style and labeling features of different radiologists and whether such modeling can improve the final metrics of object detection on radiological images.

METHODS: For training artificial intelligence systems in radiology, cross-labeling, i.e., annotation of the same image by several doctors, is frequently used. The easiest way is to use labeling from each doctor as an independent example when training the model. Some methods use different rules or algorithms to combine annotation before training. Finally, Guan et al. use separate classification heads to model the labeling style of different doctors. Unfortunately, this method is not suitable for more complex tasks, such as detecting objects on an image. For this analysis, a machine learning model designed to detect objects of different classes on mammographic scans was used. This model is a neural network based on Deformable DETR architecture. A dataset consisting of 7,756 mammographic breast scans and 12,543 unique annotations from 19 doctors was used to train the neural network. For validation and testing, a dataset consisting of 700 and 300 Bi-Rads-labeled scans, respectively, was taken. In all data sets, the proportion of images with pathology was in the 15%–20% range. A unique index was assigned to each of the 19 doctors, and a special module at each iteration of the neural network training found a vector corresponding to this index. The vector was expanded to the size of the feature map of each level of the feature pyramid,

Received: 15.05.2023 **Accepted:** 05.06.2023 **Published Online:** 10.07.2023

and then attached by separate channels to the maps. Thus, the encoder and the decoder of the detector had access to the information about which doctor labeled the scan. The vectors were updated using the back-propagation method. Three methods were chosen for comparison:

1. Basic model: Combining labels by different doctors using the "voting" method.
2. New stylistic module: For predictions on the test dataset, a single doctor's index was taken, which showed the best metrics on the validation dataset.
3. New stylistic module: The indexes of the five doctors with the best metrics on the validation dataset were used for predictions on the test dataset. Weighted Boxes Fusion was chosen to combine the predictions.

The area under the receiver operating characteristic curve (ROC-AUC) was used as the primary metric on the test dataset (Bi-Rads 3, 4, and 5 categories were referred to pathology). The sum of maximum probabilities of detected malignant objects (malignant masses and calcinates) by cranio-caudal and medio-lateral oblique projections was assumed as the probability of malignancy for each method.

RESULTS: The following ROC-AUC metrics were obtained for the three methods: 0.82, 0.87, and 0.89.

CONCLUSIONS: The information about the labeling doctor allows the neural network to learn and model the labeling style of different doctors more effectively. In addition, this method may obtain an estimate of the uncertainty of the network's prediction. The use of embedding from different doctors, leading to different predictions, may mean that this data is difficult for an artificial intelligence system to process.

Keywords: machine learning; data annotation.

FOR CITATION

Nikitin ED. Learning radiologists' annotation styles with multi-annotator labeling for improved neural network performance. *Digital Diagnostics*. 2023;4(1S):99–101. DOI: https://doi.org/10.17816/DD430358

REFERENCES

1. Jensen MH, Jørgensen DR, Jalaboi R, Hansen ME, Olsen MA. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Vol. 11767. Cham: Springer; 2019. P. 540–548. doi: 10.1007/978-3-030-32251-9_59

2. Jungo A, Meier R, Ermis E, et al. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Vol. 11070. Cham: Springer; 2018. P. 682–690. doi: 10.1007/978-3-030-00928-1_77

3. Guan MY, Gulshan V, Dai AM, Hinton GE. Who Said What: Modeling Individual Labelers Improves Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017;32(1). doi:10.1609/aaai.v32i1.11756

4. Zhu X, Su W, Lu L, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv:2010.04159*. doi: 10.48550/arXiv.2010.04159

5. Solovyev R, Wang W, Gabruseva T. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*. 2021;107:104117. doi: 10.1016/j.imavis.2021.104117

ИНФОРМАЦИЯ ОБ АВТОРЕ

Никитин Евгений Дмитриевич;
ORCID: https://orcid.org/0000-0001-7181-1036;
e-mail: e.nikitin@celsus.ai

AUTHOR'S INFO

Evgeniy D. Nikitin;
ORCID: https://orcid.org/0000-0001-7181-1036;
e-mail: e.nikitin@celsus.ai