Разнообразие вместо масштаба: разнообразие целых слайдов изображений обеспечивает обучение фундаментальной модели H&E с меньшим количеством патчей

Источник: Journal of Pathology Informatics

Оригинал: https://www.sciencedirect.com/science/article/pii/S2153353926001069?dgcid=rss_sd_all

vision transformers | гистопатология | диагностика | компьютерная патология | молекулярная подтиповая классификация | фундаментальные модели

Быстрый прогресс в вычислительной патологии всё больше обусловлен моделями-фундаментами компьютерного зрения, предварительно обученными на обширных наборах данных гистопатологии. В то время как недавние усилия были сосредоточены на обучении на всё большем количестве патчей, мы предлагаем альтернативный подход, ориентированный на разнообразие данных.

Наша модель-фундамент Athena была инициализирована из предварительно обученной модели и обучена на 115 миллионах патчей ткани (282 тыс. слайдов), что в несколько раз меньше, чем у недавних моделей-фундаментов в гистопатологии. Вместо того чтобы полагаться на объем патчей или сложные эвристики выборки, мы максимизируем разнообразие данных, случайно выбирая умеренное количество патчей на каждый слайд из нашего разнообразного внутреннего репозитория, который охватывает несколько стран, учреждений и типов сканеров.

Оцененная на одном бенчмарке уровня патчей и четырех задачах уровня слайда (две молекулярные и две морфологические), Athena приближается к передовым показателям и даже превосходит несколько моделей, обученных на значительно больших наборах данных. Это указывает на то, что разнообразие в пределах целых слайдов, а не только количество патчей, определяет обучение в моделях-фундаментах гистопатологии.

Ключевые слова
Вычислительная патология
Гистопатология
Модели-фундаменты
Трансформеры компьютерного зрения
Разнообразие данных
Предсказание молекулярного подтипа
Классификация на уровне слайда

Доступность данных
Обучающие данные, использованные в этой работе, состоят преимущественно из проприетарных гистопатологических слайдов, которые не являются общедоступными, за исключением набора данных GTEx.7 Веса предварительно обученной модели общедоступны по адресу https://huggingface.co/PAICON-GmbH/Athena-0.

Для задачи IDC/ILC использовались общедоступные слайды из TCGA10. Для CAMELYON16 соответствующий набор данных доступен.15

Для задачи микросателлитной нестабильности/стабильности (MSI/MSS) обучающие слайды были получены из TCGA10 и CPTAC,22 тогда как оценка проводилась с использованием PAIP,23 SURGEN,25 и NIB.24

Для задачи HER2 все наборы данных, кроме IXORA, общедоступны: обучающие данные были взяты из TCGA10 и CPTAC,22 а оценочные наборы данных включали YALEHER2,26 IMPRESS,28 BCNB,27 ACROBAT,29 и HEROHE.17