**Применение и развитие мультимодальных моделей в медицине в контексте больших моделей**

***Чэнь Чжо***

*Студент (магистр)*

*Московский государственный университет имени М.В.Ломоносова,*

*Институт русского языка и культуры, Москва, Россия*

*E-mail: schatten.cz@gmail.com*

В современной области искусственного интеллекта большие языковые модели (LLM, LLaVA и т.д.) показали очень хорошие результаты, что привело к расширению ряда больших языковых моделей, которые значительно улучшили семантическое понимание текста машинами. Эти модели в основном обучаются на больших наборах изображений или текстов, но такое обучение позволяет машине только делать общее понимание и простую семантическую информацию об изображениях, и не имеет возможности обрабатывать другие специфические данные (например, радиологию, акустические волны). Мультимодальные данные — это способность интегрировать несколько типов данных, включая, в частности, изображения, текст, аудио и т. д. Мультимодальные данные хорошо сочетаются с современными медицинскими данными.

В современной клинической медицине медицинские данные, как правило, являются мультимодальными, и при правильном сочетании этих данных можно принимать более точные клинические решения. Однако сложность заключается в том, что в силу специфики самой медицины в медицинских задачах обычно приходится иметь дело с мультимодальными данными, включая текстовую информацию (электронные медицинские карты, медицинские отчеты), одномерные сигналы (электрокардиограммы), двухмерные изображения (УЗИ, рентгеновские снимки) и трехмерные изображения (КТ, МРТ). Поэтому очень важно создавать полные, точные и большие массивы мультимодальных данных.

**Набор мультимодальных данных**

В качестве примера мы возьмем MedMD. Это крупномасштабный медицинский мультимодальный набор данных, содержащий 3D-данные с текстовыми описаниями, такими как рентгенологические отчеты, визуальные языковые инструкции или соответствующие метки диагностики заболеваний, а также широкий спектр рентгенологических модальностей и анатомических областей человеческого тела, в котором представлены 17 систем здравоохранения и 5000+ заболеваний [3].

**Структура модели**

Обучить мультимодальные данные напрямую очень сложно, поэтому структура модели выбирается таким образом, чтобы внедрить способность понимать изображения, медицинские сигналы в предварительно обученную макромодель медицинского языка. Итак, нам нужно выбрать визуальную модель и модель обработки медицинских сигналов, предварительно обученную для извлечения семантических особенностей изображений и сигналов, особенности изображения могут быть встроены в предварительно обученную языковую модель, здесь мы можем выбрать CLIP, предложенный OpenAI, кодер изображений может быть очень хорошим выходом и текстом для его особенностей. Затем модель настраивается на мультимодальном наборе данных, и поскольку мы по-прежнему используем большую языковую модель, по сути, потери при обучении все еще используют языковую модель.

**Модель реализуется с помощью**

LAION реализует версию Flamingo с открытым исходным кодом (https://laion.ai/blog/open-flamingo/), которая использует LLaMA для своей языковой модели, а также создает наборы данных с перекрывающимися изображениями и текстом для обучения модели.

**Технические трудности и перспективы**

В настоящее время появилось несколько мультимодальных моделей, таких как LLaVA-Med от Microsoft и Med-PaLM от Google, а также Med Flamingo. Однако они по-прежнему ограничены двумерными изображениями, и только Flamingo может поддерживать чередование графических и текстовых данных. В современной медицине большинство распространенных изображений для диагностики и лечения - это 3D-изображения, в то же время задачи диагностики и лечения обычно требуют синтеза нескольких изображений для вынесения точного решения, но с увеличением массивов данных мультимодальные большие модели данных в медицинской сфере обязательно появятся в жизни людей.

**Литература**

1. W. Chen, Y. Li, B. Ou and P. Tan, "Collaborative Multimodal Diagnostic: Fusion of Pathological Labels and Vision-Language Model," 2023 2nd International Conference on Health Big Data and Intelligent Healthcare (ICHIH), Zhuhai, China, 2023, pp. 119-126, doi: 10.1109/ICHIH60370.2023.10396187. keywords: {Pathology;Analytical models;Biological system modeling;Computational modeling;Semantics;X-ray imaging;Biomedical imaging;LangChain;chest X-ray;Multimodal language macromodel},
2. Kumar, A., Kim, J., Cai, W. et al. Content-Based Medical Image Retrieval: A Survey of Applications to Multidimensional and Multimodality Data. J Digit Imaging 26, 1025–1039 (2013).
3. Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023). Towards Generalist Foundation Model for Radiology. ArXiv, abs/2308.02463.
4. Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592.