基于编码模型解析大脑的语义加工机制

一、课题背景及相关研究

语言是人类沟通与思维的核心工具，也是区别于其他物种的关键认知能力。语言学家 Chomsky 提出，人脑内存在先天的语言习得机制，使人类能够在有限的经验输入中，从大量随机噪声中迅速构建一种语言。关于语言在大脑中如何表征，尤其是不同语义成分如何映射到特定的脑区或网络，一直是认知神经科学领域的重要研究问题。

功能磁共振成像（fMRI）技术常常被用于研究大脑的认知过程。在一般的磁共振实验范式中，刺激和响应之间的关系可以通过**一般线性模型（GLM）**来拟合（**具体请参考《磁共振成像信号分析指南》2.4 和 4.2 节**），即用数值强度或类别表示不同的刺激，训练一个刺激到 fMRI 响应的线性模型，根据体素（voxel）对应不同刺激的权重，研究大脑对特定刺激的响应模式（如图 1）。然而，诸如图像或语言等复杂的刺激无法直接表示为简单的数值，需要采用更有效的表征方式。

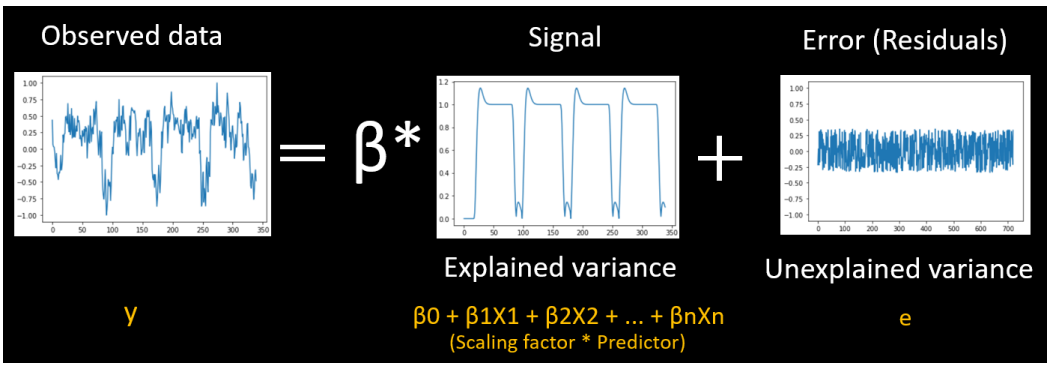

图 1 单一体素的 GLM 示意图

为解决这一问题，Gallant 课题组在语言研究中系统发展了 GLM 方法，提出“体素编码模型”（voxelwise encoding model）。他们利用自然语言处理中的特征提取方法，将复杂刺激表征为高维特征向量，再通过 GLM（岭回归）拟合各体素的 BOLD 响应，从而实现对全脑响应模式的预测与分析。这一方法能够准确预测初级感觉区及高级语义区的响应，进而揭示语义沿皮层的分布。例如，Huth 等人[1]采集了被试听叙事文本的 fMRI 数据，利用词语共现矩阵构建文本语义特征，结合 GLM 和主成分分析（principal component analysis，PCA），首次在全

脑尺度揭示了大脑的语义地图。

随着深度学习的发展，刺激特征提取逐渐从传统的特征工程转向预训练深度神经网络模型。大量研究表明，预训练音频或语言模型（如语言模态的 GPT，音频模态的 Wav2Vec 等）能够提取更高层次、语义抽象性更强的特征，从而显著提升编码模型的预测性能。Toneva 等人[2]研究了不同语言模型对大脑 fMRI 信号的预测性能，发现 Transformer 模型的中间层能更好地预测被试的大脑响应；LeBel 等人[3]将故事文本中的每个单词与之前的 25 个单词一同输入预训练 GPT 模型，得到单词的上下文表征，再结合其他不同级别的语言表征，证明了小脑在语言加工中主要参与高级语义和概念的加工；Antonello 等人[4]通过实验表明不同规模的 LLM 对大脑的预测性能，同样遵循 Scaling Law。

综上，本课题的目标是基于人类被试听故事的 fMRI 响应，结合预训练语言模型和音频模型，构建岭回归编码模型，研究不同层级的语义信息的皮层分布，以及音频与文本语义在大脑内的整合模式。

二、数据集介绍

本课题使用的数据来自 Narratives 数据集[5]，包含 345 名英语母语者收听总长 4.6 h 的自然叙事时的 fMRI 记录，其中共有 **25 名被试**收听了故事 **21st year**，总长约 56 min。fMRI 数据采集的时间分辨率（repetition time，TR）为 1.5s，全长 2249 TRs。附加材料包含：

(1) 故事文本，包含每个单词的起始和结束时刻（21styear_align.csv）。

(2) 多被试的 fMRI 数据，已按照 MMP 图谱（360 ROIs）提取各脑区信号（21styear_all_subs_rois.npy），包含 25 个被试，每个被试对应一个2249×360的数组。

三、课题内容

本课题的数据集提供了人类听故事时的 fMRI 信号，同时提供了相应的刺激音频和文本。提供的示例代码（assignment.ipynb 和 utils.py）中包含了单被试分析的全流程：(1) 提取每个 TR 对应的刺激 (2) 利用预训练模型提取特征 (3) 拟

合编码模型并在测试集上测试 (4) 可视化 correlation map。

大家需要在参考代码的基础上，完成：

(1) 音频特征的提取（参考文本特征提取，**音频下载链接[6]**，文件大小约 568MB，已提供划分音频的代码）。

(2) 多被试分析，对**每个被试训练编码模型**，报告测试集准确率的**均值和标准差**（参考代码中只包含单被试的分析流程），不要求做显著性检验。

(3) 尝试用**不同的预训练模型**提取特征，包括文本模型（GPT2、BERT、Llama、Qwen 等）、音频模型（Wav2vec、WavLM）和多模态模型（CLAP、Whisper 等），**对比不同模型、不同层的特征**对大脑的预测性能（与大脑的对齐程度）。

(4) 尝试**调整并优化特征提取方式**，如调整文本特征提取的上下文窗口（示例代码中为 200 tokens）和 pooling 方式（示例代码中采用平均每个 TR 内的单词特征）、音频特征提取的音频窗口时长（示例代码中为 2 TRs = 3 s）等，以提高编码模型的性能。

(5) 研究不同特征对**皮层各脑区**的预测性能，分析各脑区的语义偏好性。

(6) （可选）拼接不同模态的特征，找出最优预测性能的文本-音频特征组合。

(7) （可选）构建非线性编码模型，对比线性/非线性模型的性能。

四、评价标准

(1) 分析过程的严谨性和完整性，每种模态至少对比**三个模型**（GPT2-base 和 GPT2-large 算一个模型），对每个模型不同层的特征也需进行分析。

(2) 对语义表征机制或语义分布模式的可视化和解释。

(3) 上述标准为基本要求，不一定要严格执行，但需要自己有独立的思考和创新点。

五、说明

(1) 关于利用编码模型研究自然叙事 fMRI，可参考[7]。

(2) 同样基于 Narratives 数据集的工作：[8]，[9]（有代码），[10]（有代码）。

(3) 大脑结果可视化可使用 Brainspace 库[11]。

参考文献

[1] Huth A G, De Heer W A, Griffiths T L, et al. Natural speech reveals the semantic maps that tile human cerebral cortex[J]. Nature, 2016, 532(7600): 453-458.

[2] Toneva M, Wehbe L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)[J]. Advances in neural information processing systems, 2019, 32.

[3] LeBel A, Jain S, Huth A G. Voxelwise encoding models show that cerebellar language representations are highly conceptual[J]. Journal of Neuroscience, 2021, 41(50): 10341-10355.

[4] Antonello R, Vaidya A, Huth A. Scaling laws for language encoding models in fMRI[J]. Advances in Neural Information Processing Systems, 2023, 36: 21895-21907.

[5] Nastase S A, Liu Y F, Hillman H, et al. The “Narratives” fMRI dataset for evaluating models of naturalistic language comprehension[J]. Scientific data, 2021, 8(1): 250.

[6] https://datasets.datalad.org/labs/hasson/narratives/stimuli/21styear_audio.wav

[7] Binhuraib T, Gao R, Ivanova A A. LITcoder: A General-Purpose Library for Building and Comparing Encoding Models[J]. arXiv preprint arXiv:2509.09152, 2025.

[8] Millet J, Caucheteux C, Boubenec Y, et al. Toward a realistic model of speech processing in the brain with self-supervised learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 33428-33443.

[9] Oota S R, Gupta M, Toneva M. Joint processing of linguistic properties in brains and language models[J]. Advances in Neural Information Processing Systems, 2023, 36: 18001-18014.

[10] Kumar S, Sumers T R, Yamakoshi T, et al. Shared functional specialization in transformer-based language models and the human brain[J]. Nature communications, 2024, 15(1): 5523.

[11] https://brainspace.readthedocs.io/en/latest