Олександр О. Паладієв

АРХІТЕКТУРИ ЕНКОДЕР–ДЕКОДЕР ТА ЇХ УНІВЕРСАЛЬНІСТЬ

У статті здійснено системний аналіз encoder–decoder архітектур як базового підходу до задач умовної генерації типу «вхід X → вихід Y». Уточнено зміст поняття універсальності цих архітектур у трьох взаємопов'язаних вимірах: як універсальності архітектурного контракту між репрезентацією та генерацією, як універсальності апроксимаційної здатності для широкого класу sequence-to-sequence відображень і як практичної універсальності, обмеженої обчислювальною вартістю, масштабуванням і довжиною контексту. Простежено еволюцію підходу від RNN-based seq2seq моделей із фіксованим векторним bottleneck до attention-механізмів, Transformer encoder–decoder архітектур і попередньо натренованих seq2seq моделей типу T5, BART, PEGASUS. Окрему увагу приділено формалізації умовного розкладу p(y|x), ролі self-attention, cross-attention і positional encodings, а також обчислювальним обмеженням повної уваги. Розглянуто сучасні розширення парадигми, зокрема long-context моделі, retrieval-augmented generation та мультимодальні системи. На основі порівняння encoder–decoder, encoder-only, decoder-only і retrieval-augmented підходів сформульовано практичні рекомендації щодо вибору архітектури залежно від типу задачі, структури вхідних даних і ресурсних обмежень.

Парадигму кодер-декодер слід розуміти як універсальний архітектурний контракт між представленням та генерацією: він послідовно реалізується на різних мовах, мультимодальних та орієнтованих на пошук системах через явний інтерфейс (увага / перехресна увага).

«Універсальність» у сенсі експресивності має суворі теоретичні формулювання для трансформаторних моделей (універсальна апроксимація класів функцій seq2seq за певних умов), але ця теорія не замінює врахування даних, оптимізації та індуктивних зміщень

Практична універсальність обмежена вартістю уваги в тривалих контекстах; сучасні підходи розвиваються у трьох основних напрямках: топологічні модифікації уваги (Longformer/LED), алгоритмічне прискорення точної уваги (FlashAttention) та альтернативні основи моделювання послідовностей (наприклад, Mamba).

Конкуренція між парадигмами кодер-декодер та лише декодер залишається відкритим питанням: масштабні порівняльні дослідження показують, що моделі кодер-декодер можуть демонструвати конкурентні властивості масштабування та переваги логічного висновку за певних режимів, навіть якщо моделі лише декодер домінують як стандартний інтерфейс у багатьох сценаріях LLM .

Ключові слова: енкодер–декодер, sequence-to-sequence, attention, cross-attention, transformer, універсальне наближення, довгий контекст, retrieval-augmented generation, мультимодальні моделі.