Исследование контекста статьи MDETR - Modulated Detection for End-to-End Multi-Modal Understanding

Автор исследования: Данг Куинь Ньы

Статья была выложена на arXiv **26/04/2021**; вторая версия загружена 12/10/2021. В июле была принята на **ICCV2021 (oral presentation)**, презентация прошла в октябре 2021.

Авторы статьи:

- **Aishwarya Kamath**: PhD-студентка (NYU CS), на момент написания работы работала в Facebook AI, сейчас проходит стажировку в GoogleAI
- **Mannat Singh**: Facebook AI Research
- **Yann LeCun**: Chief AI Scientist at Facebook Silver Professor at the Courant Institute, New York University
- **Gabriel Synnaeve**: Research scientist at Facebook AI Research
- **Ishan Misra**: Research Scientist, Facebook AI Research
- **Nicolas Carion**: Postdoc at NYU Courant Institute

MDETR является продолжением статьи **End-to-End Object Detection with Transformers (N. Carion et al, 2020)** и использует модель DETR для мультимодальных задач. Авторы статьи MDETR Nicolas Carion и Gabriel Synnaeve также являются авторами статьи про DETR.

На данный момент у статьи 18 цитирований в работах про Visual Grounding, Referring Understanding задачи. Среди них выделяется конкурент MDETR в задаче zero-shot relation classification - **SORNet: Spatial Object-Centric Representations for Sequential Manipulation (W. Yuan et al, 2021)**. SORNet[1] на выборке CLEVR-CoGenT превосходит MDETR, которая до этой статьи показывала SOTA-результаты.

[1]Основой этой модели является Visual Transformer, в отличие от MDETR вместо текстовых запросов модели подаются canonic object views.

Возможные исследования / дальнейшее направление работы:

- в работу можно было бы добавить эксперименты с текстовым энкодером (в статье используется предобученная DistilRoberta из Huggingface)
- модели можно было бы добавить и исследовать генеративные способности
- авторы на данный момент продолжают работу в направлении self-supervised обучения

В статье приводится описание применения модели в следующих downstream задачах:

- Phrase Grounding
- Referring Expression Comprehension
- Referring Expression Segmentation
- Visual Question Answering