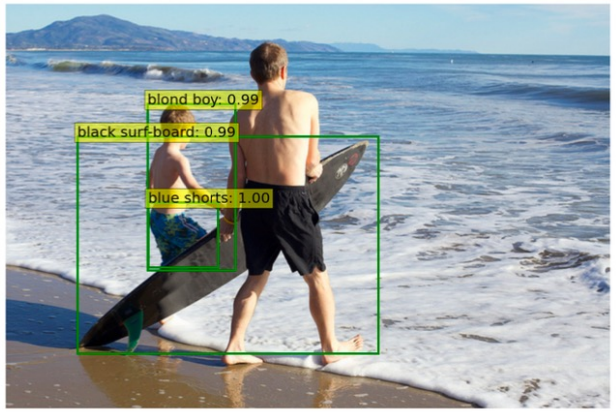
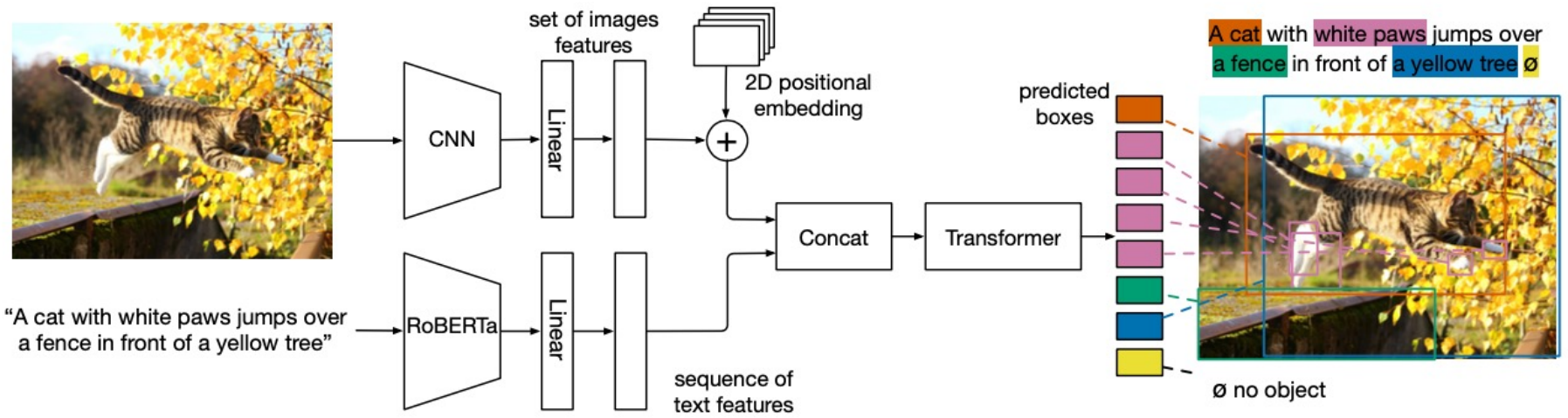
MDETR [Aishwarya Kamath+, 21]

- 検出した物体に自然言語でラベル生成
  - VQA等、様々なタスクにfine-tuning
- 画像とテキストを照合し学習
  - 画像、テキストそれぞれ特徴抽出
  - ConcatしてTransformerへ
  - 物体のbounding boxとテキストを予測





Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, Nicolas Carion.(2021). MDETR -- Modulated Detection for End-to-End Multi-Modal Understanding. arXiv:2104.12763