Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs

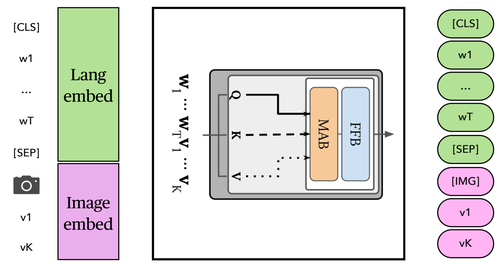
Emanuele Bugliarello **Ryan Cotterell** **Naoaki Okazaki** **Desmond Elliott**

University of Cambridge · ETH zürich · Tokyo Tech · University of Copenhagen

TACL / EMNLP 2021

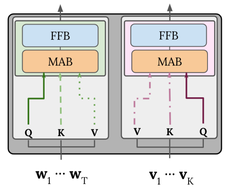
V&L BERT Families

Single-Stream
- Concatenate image—text inputs

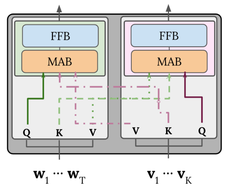


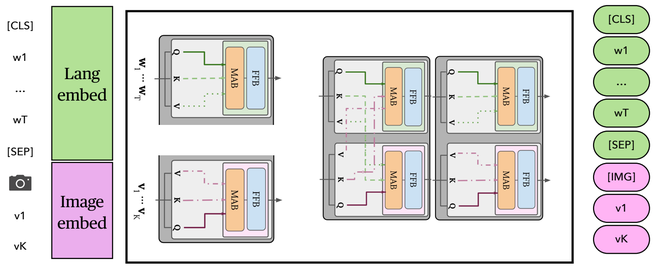
Dual-Stream
1. Model the Image and text independently
2. Two types of cross-modal layers

Intra-modal



Inter-modal



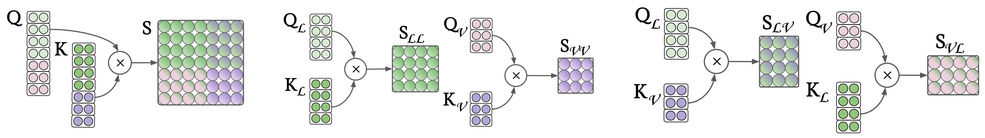


Code, Data & 50 Pretrained Models

- github.com/e-bug/volta
- github.com/e-bug/mpre-unmasked

VOLTA: A Unified Framework

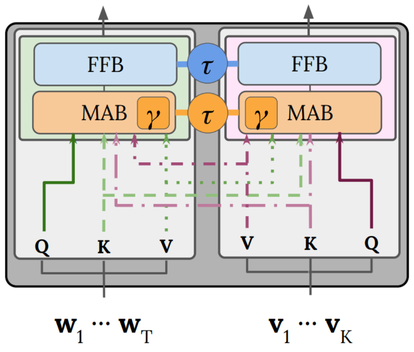
Attention matrices



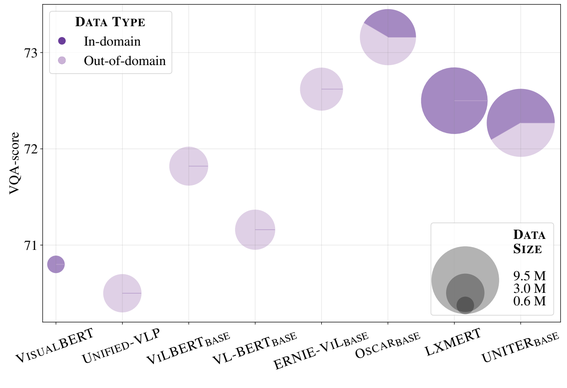
Single-stream | Dual-stream intra-modal | Dual-stream inter-modal

- **New Gated Bimodal Transformer Layer**

- **Express many architectures with this layer**
- Single- & dual-stream layers are special cases
- Takes a set of fixed binary variables $\{\gamma, \tau\}$
  - γ: regulate the cross-modal interactions
  - τ: control if parameters are tied between modalities



Which V&L BERT? 🤷‍♀️



These models are trained on different amounts & types of data

- **In-domain:** also relevant for downstream tasks (e.g. COCO)
- **Out-of-domain**: unrelated to downstream tasks (e.g. CC)

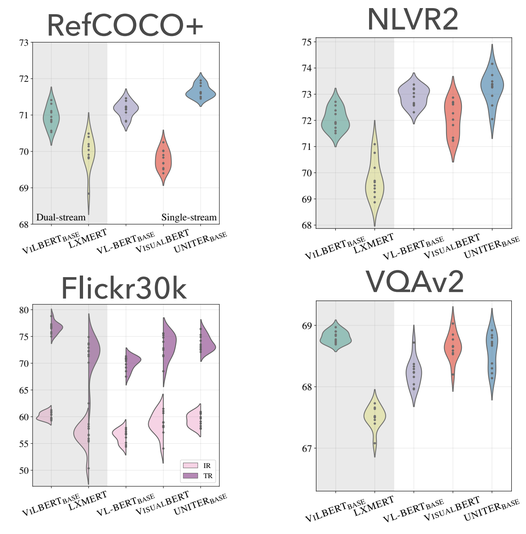
Controlled Experiments

- We implement existing models in **VOLTA**
- Train them on the same setup with multiple seeds and the MLM + MRC-KL + ITM objectives

Models:
- *Single-stream*: VL-BERT, VisualBERT, UNITER
- *Dual-stream*: ViLBERT, LXMERT

Pretraining data: Conceptual Captions (2.77M)

Results



1. Substantial variation due to **seed**
2. **Similar performance** in controlled setup
3. Single- & Dual-Stream are **on par**
4. **Embedding layer** is crucial