
Emanuele Bugliarello  Ryan Cotterell  Naoaki Okazaki  Desmond Elliott

TAACL / 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

**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

**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

**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