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

Introduced and implemented a unified **mathematical framework**, under which recently proposed **V&L BERTs can be specified as special cases**. We conducted a series of controlled studies within this framework to better understand the differences between several models. We found that the performance of the considered models varies significantly due to random initialisation, in both pretraining and fine-tuning. We also found that these models achieve similar performance when trained with the same hyperparameters and data. Notably, some models outperform others but we found that (a) single- and dual-stream model families are on par, and (b) embedding layers play a crucial role towards a model’s final performance

These models can be categorised into either **single-stream** or **dual-stream** encoders. We study the differences between these two categories, and show how they can be unified under a **single theoretical framework**. We then conduct controlled experiments to discern the **empirical differences between five V&L BERTs**

All of these V&L models extend BERT (Devlin et al., 2019) to learn representations grounded in both modalities. They can either be classified as (i) single-stream, where images and text are jointly processed by a single encoder (e.g., Zhou et al. 2020), or (ii) dual-stream, where the inputs are encoded separately before being jointly modelled (e.g., Tan and Bansal 2019)

In order to better understand these models, we conduct a series of controlled studies to investigate whether differences in downstream performance is explained by: (i) **the amount of pretraining data** and the pretraining objectives (e.g., Figure 2); (ii) the **hyperparameters** used to control the learning process; (iii) the variance caused by **random initialisation** when pretraining (e.g., Figure 1), (iv) the variance due to **fine-tuning** multiple times on a downstream task; (v) being **single- or dual-stream** architectures; or (vi) the choice of the **embedding layer**.

**Input Embeddings**

Language input All V&L BERTs adopt the approach of BERT: The input sequence is first tokenized into sub-word units and two special [CLS] and [SEP] tokens are added to generate the text sequence {[CLS], w1, . . . , wT , [SEP]}. The embedding of each token is then given by the sum of three learnable vectors, corresponding to its form, position in the sequence and segment In addition, VL-BERT also adds the visual feature of the entire image to each token.

**Encodings**

**Single-stream encoders** A standard BERT architecture is given the concatenation of the visual and linguistic features of an image–text pair as input. This design allows for an early and unconstrained fusion of cross-modal information.

**Dual-stream encoders** Here, the visual and linguistic features are first processed by two independent stacks of Transformer layers. The resulting representations are then fed into cross-modal Transformer layers where intra-modal interactions are alternated with inter-modal interactions. Both VILBERT and LXMERT modelled **inter-modal interactions** in the same way: **each stream first computes its query, key, and value matrices, before passing the keys and values to the other modality**. By doing so, these models explicitly constrain interactions between modalities at each layer, inhibiting some of the interactions that are possible in a singlestream encoder while increasing their expressive power by separate sets of learnable parameters.

Diagram

Description automatically generated

**Single-stream Multimodal Transformers**

A close-up of a document

Description automatically generated with low confidence

**Dual-stream Multimodal Transformers**

**A screenshot of a computer

Description automatically generated with low confidence**

**Text, letter

Description automatically generated**

**Dual-stream Attentions as Restricted Single-stream Attention**

**Text, letter

Description automatically generated**

**Text

Description automatically generated with medium confidence**

**Gated Bimodal Transformer Layers**

We now introduce a general gated bimodal Transformer layer (Figure 3d), in which both single- and dual-stream layers are special cases. By doing so, we can define existing V&L BERTs within a single architecture, which allows us to implement and evaluate several of these models in a controlled. In addition to textual XL and visual embeddings XV, this layer takes a set of fixed binary variables {γ, τ } as part of its input: γ = {γLV, γVL, γLL, γVV}, and τ = {τMHA, τLN1 , τFF , τLN2 }. The γ values act as gates that regulate the cross-modal interactions within a layer, while the τ values control whether the parameters are tied between modalities.

**Text, letter

Description automatically generated**

**Fine-tuning variance**

It was observed that the fine-tuning of BERT is sensitive to randomness in initialisation and data ordering. We also report an average standard deviation of 0.3 points for these models across both tasks. However, the minimum and the maximum scores of a given model often differ by 1 or more points, showing how a single fine-tuning run of these models can lead to incorrect conclusions.

**Pretraining variance**

we find that some of these architectures are less prone to variance caused by pretraining seed, such as VILBERT for VQA and retrieval tasks, and UNITER for referring expression. Nevertheless, the performance of all of these models can vary by more than 1 point in several tasks solely due to random initialisation.

**Diagram

Description automatically generated**

**Single or Dual stream**

s. Lu et al. (2019) showed how their single-stream baseline performed worse than their dual-stream VILBERT architecture, while Chen et al. (2020) claimed single-stream UNITER outperformed VILBERT. Our controlled study across several tasks and different pretraining initialisations allows us to provide an answer grounded with statistical tests. To do so, we split the models in dual- and single-stream architectures17 and run a one-way ANOVA (Table 3). After Bonferroni correction, we only find statistical difference at p < 0.005 (Benjamin et al., 2018) between these two groups for the Flickr30K text retrieval task.

**Table

Description automatically generated**

**Importance of the embeddings**

**Chart, box and whisker chart

Description automatically generated**

**Multimodal Natural Language Inference Final Report**

Graphical user interface, text

Description automatically generated

Table

Description automatically generated

Diagram

Description automatically generated

Table

Description automatically generated

A picture containing chart

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**Evaluating Multimodal Representations on Visual Semantic Textual Similarity**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**Visual Entailment: A Novel Task for Fine-Grained Image Understanding**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Graphical user interface, diagram

Description automatically generated**

**Table

Description automatically generated**

**Graphical user interface, website

Description automatically generated**

**Learning Transferable Visual Models From Natural Language Supervision**

**Table

Description automatically generated with low confidence**

**Chart, line chart

Description automatically generated**

**SemVLP: Vision-Language Pre-training by Aligning Semantics at Multiple Levels**

**Diagram

Description automatically generated**

Graphical user interface, application

Description automatically generated

<https://huggingface.co/docs/transformers/model_doc/lxmert#transformers.LxmertForQuestionAnswering>

<https://huggingface.co/docs/transformers/main_classes/trainer>