## A Details on Pre-trained V+L Models

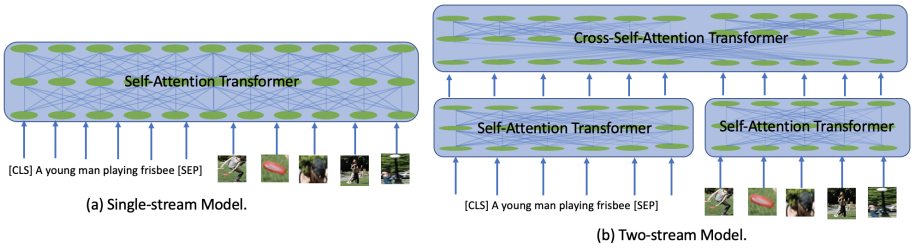



Fig. 5: Comparison between single-stream and two-stream V+L models.

A comparison between single-stream and two-stream V+L models is provided in Figure 5. We choose UNITER-base[10] [6] as the representative model for single-stream, and LXMERT[11] [32] for two-stream. As shown in Figure 5(a), UNITER-base has the same model structure as the BERT-base model [9], which composes of 12 layers of self-attention Transformers. Each layer has 12 self-attention heads, and each hidden representation is a 768-dimensional vector. As shown in Figure 5(b), LXMERT is a two-stream model that performs intra-attention in the same modality first, then cross-attention. We denote $T_t, T_v$, and $T_c$ as the Transformer modules that specifically model text-to-text, image-to-image and cross-modal interactions, respectively. In LXMERT, $T_t$ has 9 layers, $T_v$ has 5 layers, and $T_c$ has 5 layers. Each Transformer's hidden representation is in dimension of 768. Note that each layer in $T_c$ contains one cross-attention layer between two modalities, followed by two self-attention layers for each modality.

## B Results on Untrained Baselines

To measure the gain from learning, we also conducted additional experiments on untrained single-stream (SS) and two-stream (TS) baselines with random weights.

**Multimodal Fusion Probe** The untrained SS model has NMI of 0.99 for all output layers, suggesting that the two modalities are completely separated. The untrained TS model has NMI of 0.56 for all output layers. This is because the cross-modality encoder layers force the two modalities to fuse, even in untrained setting.

**Modality Importance Probe** For the untrained model, the average attention of `[CLS]` token on the image/text modality is 0.66/0.28. Note that the number of tokens in a sentence is usually smaller than that of the visual tokens.

[10] https://github.com/ChenRocks/UNITER
[11] https://github.com/airsplay/lxmert

| Model | VCD | VCC | VRI | VRC |
|---|---|---|---|---|
| Untrained SS | 50.0 | 58.0 | 50.0 | 11.4 |
| Untrained TS | 50.0 | 66.0 | 50.0 | 9.34 |

Table 6: Untrained visual coreference/relation attention baselines.

| Coref Type | people | body parts | scene | clothing | instruments | animals |
|---|---|---|---|---|---|---|
| Ratio | 0.33 | 0.23 | 0.28 | 0.37 | 0.59 | 0.53 |

Table 7: Results on whether the head selected for a specific coreference relation between an image-text pair imposes higher attention scores than all the other pairs.

**Visual Coreference and Relation Probe** We provide additional untrained baselines for visual coreference and visual relation probes in Table 6. Compared to Table 3 and Figure 4(a), for VCD and VRI, untrained baselines for both SS and TS are equivalent to random guess. For VRC, both SS and TS models outperform the baseline by around 10%. For VCC, the SS model outperforms the baseline by 17%; while the TS model performs worse. This may be because after hard-designed multimodal fusion, the direct coreference relationship between a pair of image/text tokens is diluted after training.

Furthermore, we provide an additional evaluation on whether the head selected for a specific coreference relation of an image-text pair imposes higher attention scores for coreference relation than all other pairs. Results are summarized in Table 7, which suggests that these attention heads with maximum attention weight do pay more attention to the coreference image-text pair, compared to other unpaired ones.

## C Additional Guidelines for Future Model Design

In addition to the key takeaways in Sec. 4, we provide a set of guidelines for future model design based on our analysis and observations.

(*i*) Single-stream model is able to capture sufficient intra- and cross-modal knowledge, while the restricted attention structure in two-stream model does not bring additional benefit. For future work, we will further explore single-stream model design, which also exhibits better interpretability as observed.

(*ii*) Initializing V+L model with BERT's weights should be helpful, which can enhance V+L model's capability in language understanding.

(*iii*) It remains unclear how to measure a pre-trained model without evaluating on downstream tasks. Given that finetuning is time consuming, the probing tasks we propose can provide a convenient tool to quickly test intermediate model checkpoints during pre-training.

(*iv*) Explicitly adding extra supervision to probing tasks during model training may lead to more interpretable and robust model.

## D Details on Linguistic Probe

We probe the pre-trained models over nine tasks defined in the SentEval toolkit [8], under three categories:

(*i*) *Surface tasks*: probe for the length of a sentence (`SentLen`);
(*ii*) *Syntactic tasks*: predict the depth of a sentence's syntax tree, consecutive token inversions (`BShift`), and the top constituents sequences (`TopConst`);
(*iii*) *Semantic tasks*: test the tense (`Tense`), the number implied by the subject/object (`SubjNum`/`ObjNum`), the replacement of the noun/verb form (`SOMO`), and the inversion of coordinating conjunctions (`CoordInv`).

## E Additional Results

We provide additional results on multimodal fusion, visual coreference resolution, visual relation detection, and linguistic probing.

### E.1 An t-SNE Visualization of Multimodal Fusion Degree

An t-SNE visualization of multimodal fusion degree of the first and last layer of UNITER (over one image-text pair) is provided in Figure 6. As the layer goes deeper, the two modalities become more intertwined.

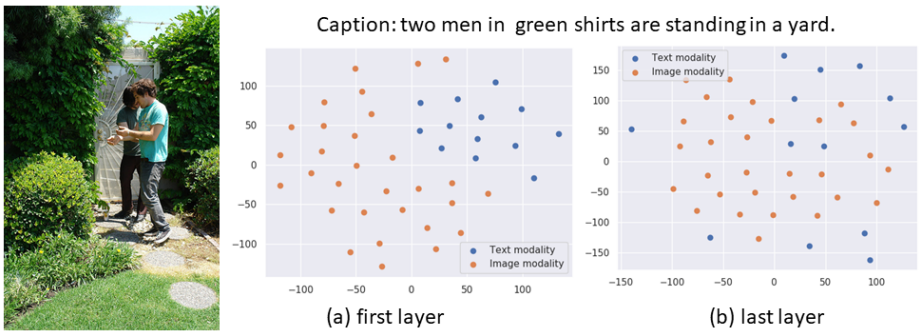



Fig. 6: An t-SNE visualization of multimodal fusion degree of the first and last layer of UNITER over one image-text pair. Each yellow and blue dot corresponds to a visual and textual token, respectively.

### E.2 Visual Coreference Resolution

Due to space limit, we only reported results using the embeddings from Layer 1, 5 and 12 in Table 3. A complete set of results is provided in Table 8. We

| Classifier Input | VCD (SS) | VCD (TS) | VCC (SS) | VCC (TS) |
|---|---|---|---|---|
| 144 Attention Heads | 52.04 | **53.68** | 75.10 | 54.47 |
| Random Guess | 50.00 | 50.00 | 12.50 | 12.50 |
| Layer 1 | 56.86 | **53.68** | 93.51 | 93.35 |
| Layer 2 | 57.58 | 53.49 | 93.91 | **93.36** |
| Layer 3 | 57.81 | 53.32 | 94.11 | 93.32 |
| Layer 4 | 57.97 | 52.92 | 94.10 | 93.12 |
| Layer 5 | **59.12** | 52.59 | 94.05 | 92.62 |
| Layer 6 | 58.58 | / | 94.02 | / |
| Layer 7 | 58.67 | / | **94.26** | / |
| Layer 8 | 58.65 | / | 93.96 | / |
| Layer 9 | 58.15 | / | 93.77 | / |
| Layer 10 | 57.65 | / | 93.77 | / |
| Layer 11 | 57.96 | / | 93.47 | / |
| Layer 12 | 58.40 | / | 93.44 | / |

Table 8: Results of attention and layer-wise embedding probers on Visual Coreference Detection and Classification (VCD and VCC). SS: single-stream; TS: two-stream.

observe that the attention probers work well for VCC, but not for VCD. Our assumption is that task granularity matters to the prober's performance. Attention behavior varies a lot in different coreference relations, thus it performs well on VCC. The dataset for training VCC is built with positive examples from VCD only. Therefore, VCD's settings naturally dilute the distinction between different coreference relations' attentions, which makes it a more challenging task.

### E.3 Visual Relation Detection

Results of the layer-wise embedding probers on the Visual Relation Classification and Identification (VRC and VRI) tasks are visualized in Figure 4(b) and (c), respectively. Detailed numbers corresponding to these two figures are provided in Table 9 and 10.

### E.4 Linguistic Probing

For linguistic probing, we first obtain results from all the layers of a pre-trained model, then report the best number in Table 5. Detailed results for all the layers are provided in Table 11.

### E.5 Visualization of Attention Maps

We show the learned attention maps of one specific relation in the probing tasks: Figure 7 and 8 for visual coreference resolution (Section 3.3.2 of the main paper), and Figure 9 and 10 for visual relation detection (Section 3.4 of the main paper).

| Classifier Input | VRI (SS) | VRC (SS) | VRI (SS mis.) | VRC (SS mis.) |
|---|---|---|---|---|
| Original visual emb. | 76.95 | 36.38 | 76.95 | 36.38 |
| Layer 1 | 77.18 | 37.70 | **77.03** | 37.31 |
| Layer 2 | 79.56 | 42.22 | 76.84 | 37.55 |
| Layer 3 | 82.28 | 43.02 | 76.55 | 37.05 |
| Layer 4 | 83.24 | 45.88 | 76.40 | 37.66 |
| Layer 5 | 84.45 | 47.67 | 76.28 | 37.49 |
| Layer 6 | 84.35 | 46.46 | 75.99 | 37.54 |
| Layer 7 | 84.13 | 45.67 | 75.88 | 37.51 |
| Layer 8 | 83.95 | 45.32 | 75.62 | **37.71** |
| Layer 9 | 85.98 | 54.35 | 74.75 | **37.71** |
| Layer 10 | **86.35** | 55.66 | 74.15 | 37.06 |
| Layer 11 | 86.19 | **56.64** | 73.84 | 36.72 |
| Layer 12 | 86.07 | 55.22 | 72.96 | 36.65 |

Table 9: Accuracies (%) of the layer-wise embedding probers on Visual Relation Identification and Classification (VRI and VRC) tasks for the single-stream (SS) model. mis.: mismatch.

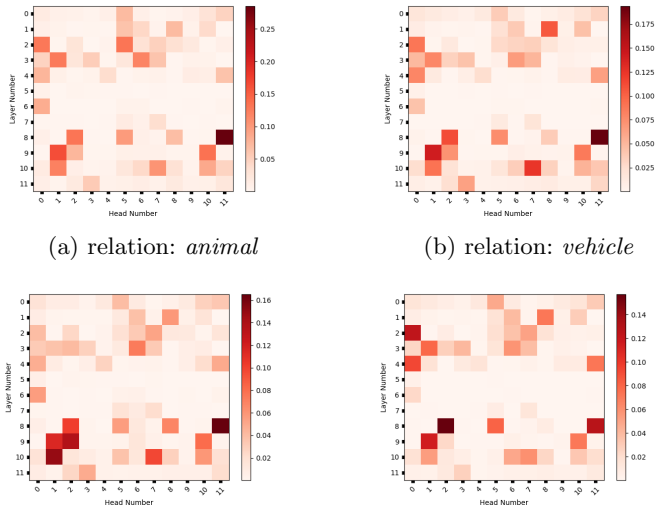
| Classifier Input | VRI (TS) | VRC (TS) | VRI (TS mis.) | VRC (TS mis.) |
|---|---|---|---|---|
| Original visual emb. | 76.95 | 36.38 | 76.95 | 36.38 |
| Layer 1 | 75.92 | 36.61 | 77.58 | **36.57** |
| Layer 2 | 75.42 | 35.86 | 77.44 | 35.82 |
| Layer 3 | 75.01 | 35.72 | 77.38 | 35.66 |
| Layer 4 | 74.67 | 36.01 | 77.19 | 35.99 |
| Layer 5 | 74.75 | 36.45 | 77.05 | 36.43 |
| Layer 6 | **87.00** | 67.82 | **87.03** | 13.20 |
| Layer 7 | 86.59 | 68.06 | 86.24 | 12.83 |
| Layer 8 | 86.14 | 68.67 | 85.86 | 11.93 |
| Layer 9 | 85.50 | 68.57 | 85.36 | 12.11 |
| Layer 10 | 85.43 | **69.66** | 85.26 | 12.07 |

Table 10: Accuracies (%) of the layer-wise embedding probers on Visual Relation Identification and Classification (VRI and VRC) tasks for the two-stream (TS) model. mis.: mismatch.

| Layer | SentLen (Surface) | TreeDepth (Syntactic) | TopConst (Syntactic) | BShift (Syntactic) | Tense (Semantic) |
|---|---|---|---|---|---|
| 1 | 86.5, 75.8, **93.9** | 29.8, 29.3, **35.9** | 36.5, 31.8, **63.6** | 50.0, 50.0, **50.3** | 71.8, 66.0, **82.2** |
| 2 | 88.8, 73.8, **95.9** | 33.6, 27.9, **40.6** | 54.6, 29.4, **71.3** | 50.0, 50.0, **55.8** | 77.8, 70.5, **85.9** |
| 3 | 87.4, 74.4, **96.2** | 34.7, 28.0, **39.7** | 67.5, 32.5, **71.5** | 56.6, 51.8, **64.9** | 83.7, 72.1, **86.6** |
| 4 | 87.7, 76.6, **94.2** | 35.3, 29.0, **39.4** | 69.7, 39.3, **71.3** | 71.2, 54.9, **74.4** | 84.3, 72.5, **87.6** |
| 5 | 86.5, 77.5, **92.0** | 36.4, 29.5, **40.6** | 72.5, 48.6, **81.3** | 73.6, 55.4, **81.4** | 84.1, 74.4, **89.5** |
| 6 | 85.0, 81.8, **88.4** | 36.1, 31.5, **41.3** | 73.5, 48.1, **83.3** | 74.6, 62.3, **82.9** | 83.0, 75.6, **89.8** |
| 7 | 83.6, **83.8**, 83.7 | 36.2, 32.7, **40.1** | 79.0, 63.4, **84.1** | 76.5, 63.4, **83.0** | 83.8, 75.6, **89.9** |
| 8 | 81.7, 81.8, **82.9** | 35.1, 34.0, **39.2** | 78.0, 67.2, **84.0** | 77.3, 64.9, **83.9** | 84.0, 75.2, **89.9** |
| 9 | 79.7, 79.8, **80.1** | 34.5, 32.7, **38.5** | 76.5, 65.7, **83.1** | 78.8, 64.8, **87.0** | 85.3, 75.1, **90.0** |
| 10 | **77.4**, / , 77.0 | 33.9, / , **38.1** | 75.6, / , **81.7** | 81.6, / , **86.7** | 86.4, / , **89.7** |
| 11 | **77.5**, / , 73.9 | 34.1, / , **36.3** | 73.9, / , **80.3** | 80.9, / , **86.8** | 86.6, / , **89.9** |
| 12 | **74.6**, / , 69.5 | 32.2, / , **34.7** | 70.9, / , **76.5** | 80.8, / , **86.4** | 86.2, / , **89.5** |

| Layer | SubjNum (Semantic) | ObjNum (Semantic) | SOMO (Semantic) | CoordInv (Semantic) |
|---|---|---|---|---|
| 1 | 69.0, 70.6, **77.6** | 65.3, 69.1, **76.7** | 49.9, **51.0**, 49.9 | 50.0, 51.2, **53.9** |
| 2 | 74.8, 71.2, **82.5** | 72.5, 70.8, **80.6** | 50.1, 50.0, **53.8** | 50.5, 50.0, **58.5** |
| 3 | 79.6, 70.7, **82.0** | 78.2, 70.0, **80.3** | 50.3, 50.5, **55.8** | 56.8, 50.1, **59.3** |
| 4 | 79.9, 72.6, **81.9** | 77.0, 71.7, **81.4** | 50.9, 50.1, **59.0** | 57.8, 50.0, **58.1** |
| 5 | 80.5, 74.8, **85.8** | 76.6, 74.5, **81.2** | 51.0, 50.1, **60.2** | 59.4, 56.2, **64.1** |
| 6 | 81.1, 74.6, **88.1** | 77.1, 74.6, **82.0** | 52.2, 50.3, **60.7** | 59.4, 56.1, **71.1** |
| 7 | 83.4, 77.3, **87.4** | 78.4, 76.0, **82.2** | 54.3, 51.2, **61.6** | 60.3, 57.4, **74.8** |
| 8 | 82.7, 78.8, **87.5** | 78.2, 76.8, **81.2** | 54.5, 51.4, **62.1** | 60.5, 58.7, **76.4** |
| 9 | 81.8, 78.8, **87.6** | 78.8, 76.7, **81.8** | 55.9, 51.0, **63.4** | 61.8, 58.0, **78.7** |
| 10 | 81.9, / , **87.1** | 78.5, / , **80.5** | 56.3, / , **63.3** | 61.9, / , **78.4** |
| 11 | 83.0, / , **85.7** | 78.2, / , **78.9** | 56.6, / , **64.4** | 62.1, / , **77.6** |
| 12 | 81.8, / , **84.0** | 77.8, / , **78.7** | 57.1, / , **65.2** | 61.7, / , **74.9** |

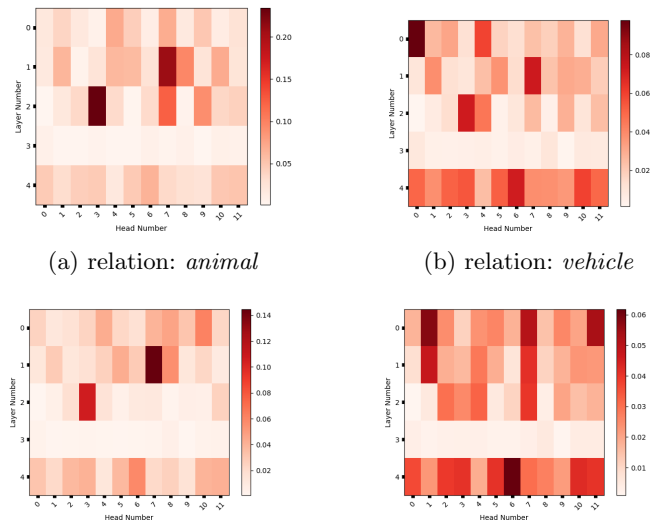
Table 11: Results on the linguistic probing tasks. For each task and each layer, the results are presented in the order of UNITER, LXMERT, and the original BERT.




(a) relation: *animal* (b) relation: *vehicle*

(c) relation: *people* (d) relation: *clothing*

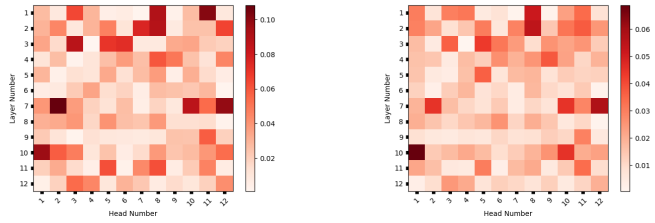
Fig. 7: Visualization of coreference information for all 144 attention heads ($V \rightarrow T$) in the single-stream model (UNITER-base). Note that only a set of attention heads is significant to the $V \rightarrow T$ attention across different coreference relations.



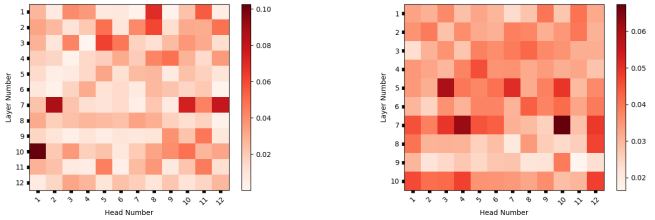

(a) relation: *animal* (b) relation: *vehicle*

(c) relation: *people* (d) relation: *clothing*

Fig. 8: Visualization of coreference information across all attention heads ($V \rightarrow T$) in the two-stream model (LXMERT-base, 5 layers, 12 heads per layer).

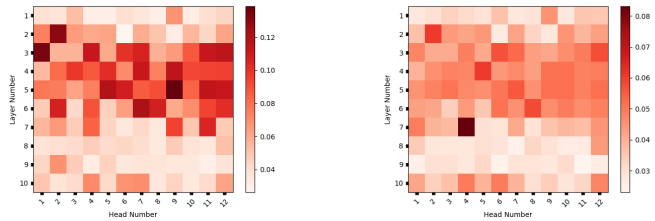



(a) relation: *covering* (b) relation: *at*

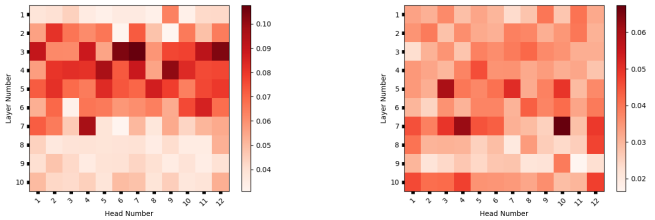



(c) relation: *on* (d) relation: *playing*

Fig. 9: Visualization of the maximum attention between two visually-related tokens across 144 attention heads in single-stream model (12 layers, 12 heads per layer). Note that the spatial relationships (`on`, `at`) have similar attention maps compared to other relations.




(a) relation: *covering* (b) relation: *at*




(c) relation: *on* (d) relation: *playing*

Fig. 10: Visualization of the maximum attention between two visually-related tokens across the attention heads in two-stream model (10 layers: 1-5 layers: self-attention; 6-10 layers: cross-self-attention, 12 heads per layer).