Vision-and-Language or Vision-for-Language?
On Cross-Modal Influence in Multimodal Transformers

Stella Frank*
University of Trento
stella.frank@unitn.it

Emanuele Bugliarello*
University of Copenhagen
emanuele@di.ku.dk

Desmond Elliott
University of Copenhagen
de@di.ku.dk
V&L Transformers
V&L Transformers

Model Zoo

- LXMERT (Tan & Bansal, 2019)
- ViLBERT (Liu+, 2019)
- VL-BERT (Su+, 2020)
- ...
V&L Transformers

Model Zoo

- LXMERT (Tan & Bansal, 2019)
- ViLBERT (Liu+, 2019)
- VL-BERT (Su+, 2020)
- ...

But how multimodal are they really?
V&L Transformers

Model Zoo
- LXMERT (Tan & Bansal, 2019)
- ViLBERT (Liu+, 2019)
- VL-BERT (Su+, 2020)
- ...

But how multimodal are they really?
- Downstream performance might be misleading
V&L Transformers

Model Zoo

- LXMERT (Tan & Bansal, 2019)
- ViLBERT (Liu+, 2019)
- VL-BERT (Su+, 2020)
- ...

But how multimodal are they really?

- Downstream performance might be misleading
- Previous work: Cao+(2020) Li+(2020) Parcalabescu+(2021)
V&L Transformers

Model Zoo

- LXMERT (Tan & Bansal, 2019)
- ViLBERT (Liu+, 2019)
- VL-BERT (Su+, 2020)
- ...

But how multimodal are they really?

- Downstream performance might be misleading
- Previous work: Cao+(2020) Li+(2020) Parcalabescu+(2021)
- Ours: An easy way of assessing cross-modal influence within these models
V&L Transformers

Model Zoo

- LXMERT (Tan & Bansal, 2019)
- ViLBERT (Liu+, 2019)
- VL-BERT (Su+, 2020)
- ...

But how multimodal are they really?

- Downstream performance might be misleading
- Previous work: Cao+(2020) Li+(2020) Parcalabescu+(2021)
- Ours: An easy way of assessing cross-modal influence within these models
Cross-Modal Input Ablation
Cross-Modal Input Ablation

How does a missing modality affect model predictions?
Cross-Modal Input Ablation

How does a missing modality affect model predictions?

Based on the same objectives used during pretraining: what the model is trained to do
Cross-Modal Input Ablation

How does a missing modality affect model predictions?

Based on the same objectives used during pretraining: what the model is trained to do

e.g. Masked Language Modelling
Cross-Modal Input Ablation

How does a missing modality affect model predictions?

Based on the same objectives used during pretraining: what the model is trained to do

e.g. Masked Language Modelling

• How much does the model rely on vision to predict a masked token?
  1. With vision inputs
  2. Without vision inputs
Cross-Modal Input Ablation

How does a missing modality affect model predictions?

Based on the same objectives used during pretraining: what the model is trained to do

e.g. Masked Language Modelling

• How much does the model rely on vision to predict a masked token?
  1. With vision inputs
  2. Without vision inputs

Falsifiable hypothesis
Ablating Vision-for-Language
Ablating Vision-for-Language

How much does the model rely on visual inputs for text predictions?
Ablating Vision-for-Language

How much does the model rely on visual inputs for text predictions?

- No ablation (None)
Ablating Vision-for-Language

How much does the model rely on visual inputs for text predictions?

• No ablation *(None)*

[MASK] playing tennis
Ablating Vision-for-Language

How much does the model rely on visual inputs for text predictions?

- No ablation (**None**)

**Vision-for-Language Diagnostic**

**Language-for-Vision Diagnostic**

$p([\text{MASK}] = \text{girl})$

[V&L BERT] playing tennis
Ablating Vision-for-Language

How much does the model rely on visual inputs for text predictions?

• No ablation (None)
Ablating Vision-for-Language

How much does the model rely on visual inputs for text predictions?

• No ablation (None)

• Full ablation (All)
Ablating Vision-for-Language

How much does the model rely on visual inputs for text predictions?

• No ablation (None)

• Full ablation (All)

Expect decreased performance
Ablating Vision-for-Language

How much does the model rely on visual inputs for text predictions?

• No ablation (None)

• Object ablation (Object)

• Full ablation (All)

Expect decreased performance
Ablating Vision-for-Language

How much does the model rely on visual inputs for text predictions?

• No ablation (None)

• Object ablation (Object)

• Full ablation (All)

 Expect decreased performance

Expect intermediate performance

FRANK, BUGLIARELLO & ELLIOTT | VISION-AND-LANGUAGE OR VISION-FOR-LANGUAGE?

EMNLP 2021
Ablating Language-for-Vision
Ablating Language-for-Vision

How much does the model rely on textual inputs for vision predictions?
Ablating Language-for-Vision

How much does the model rely on textual inputs for vision predictions?

• No ablation (None)

Girl playing tennis
Ablating Language-for-Vision

How much does the model rely on textual inputs for vision predictions?

• No ablation (None)

Girl playing tennis
Ablating Language-for-Vision

How much does the model rely on textual inputs for vision predictions?

• No ablation (None)

• Full ablation (All)
Ablating Language-for-Vision

How much does the model rely on textual inputs for vision predictions?

• No ablation (None)

• Phrase ablation (Phrase)

• Full ablation (All)
Experimental Setup
Experimental Setup

Data

- Flickr30k Entities (validation)
  - Human-annotated phrase-image alignments
Experimental Setup

Data
- Flickr30k Entities (validation)
  - Human-annotated phrase-image alignments

Models
- 5 V&L BERTs from VOLTA (Bugliarello+, 2021)
Experimental Setup

Data

• Flickr30k Entities (validation)
  ■ Human-annotated phrase-image alignments

Models

• 5 V&L BERTs from VOLTA (Bugliarello+, 2021)
• Vision inputs from Faster R-CNN (Anderson+, 2018)
Experimental Setup

Data
- Flickr30k Entities (validation)
  - Human-annotated phrase-image alignments

Models
- 5 V&L BERTs from VOLTA (Bugliarello+, 2021)
- Vision inputs from Faster R-CNN (Anderson+, 2018)
- Prediction tasks
  - Vision-for-language: MLM
  - Language-for-vision: MRC-KL
Vision-for-Language Ablation
Vision-for-Language Ablation

![Graph showing the evaluation of different vision-for-language models across different conditions.](image)

- BERTCC
- ViLBERT
- VisualBERT
- LXMERT
- VL-BERT
- UNITER

The graph compares the performance of various vision-for-language models under different conditions: None, Object, and All. The x-axis represents the conditions, and the y-axis represents the bit error rate. The full image condition shows the highest performance, followed by the object condition, and then the none condition.
Vision-for-Language Ablation

Performance degrades (increased MLM perplexity) as visual inputs are removed
Vision-for-Language Ablation

Performance degrades (increased MLM perplexity) as visual inputs are removed
Language-for-Vision Ablation
Language-for-Vision Ablation

![Graph showing the performance of different models (BERT_{CC}, ViLBERT, VisualBERT, LXMERT, VL-BERT, UNITER) across different scenarios (Full text, Phrase, No text) in terms of bit accuracy.]
Language-for-Vision Ablation

Performance *barely* degrades (increased MRC KL) as textual inputs are removed.
Language-for-Vision Ablation

Performance *barely* degrades (increased MRC KL) as textual inputs are removed.

MODELS DO NOT USE LANGUAGE FOR THE VISION TASK
Why No Language-for-Vision?
Why No Language-for-Vision?

- Model architectures
Why No Language-for-Vision?

- Model architectures
Why No Language-for-Vision?

- Model architectures
- Form of MRC loss
Why No Language-for-Vision?

- Model architectures
- Form of MRC loss
Why No Language-for-Vision?

- Model architectures
- Form of MRC loss
- Pretraining regime
Why No Language-for-Vision?

- Model architectures
- Form of MRC loss
- Pretraining regime
Why No Language-for-Vision?

- Model architectures
- Form of MRC loss
- Pretraining regime
Why No Language-for-Vision?

- Model architectures
- Form of MRC loss
- Pretraining regime
- Visual leakage
Why No Language-for-Vision?

- Model architectures
- Form of MRC loss
- Pretraining regime
- Visual leakage
Why No Language-for-Vision?

- Model architectures
- Form of MRC loss
- Pretraining regime
- Visual leakage
The Devil's in the Data
The Devil's in the Data

MRC is based on silver data
The Devil's in the Data

MRC is based on silver data

- Faster R-CNN object category predictions
The Devil's in the Data

MRC is based on silver data

- Faster R-CNN object category predictions
- They often do not match the text description
The Devil's in the Data

MRC is based on silver data

- Faster R-CNN object category predictions
- They often do not match the text description

Analysis by category:
The Devil's in the Data

MRC is based on silver data

• Faster R-CNN object category predictions
• They often do not match the text description

Analysis by category:
• people = \{man, woman, \ldots\}
MRC is based on silver data
• Faster R-CNN object category predictions
• They often do not match the text description

Analysis by category:
• people = \{man, woman, \ldots\}
MRC is based on silver data
• Faster R-CNN object category predictions
• They often do not match the text description

Analysis by category:
• people = \{man, woman, \ldots\}
• Most confusion is within categories
MRC is based on silver data
• Faster R-CNN object category predictions
• They often do not match the text description

Analysis by category:
• people = {man, woman, …}
• Most confusion is within categories

Object label–text label mismatch hinders learning language-for-vision
Conclusions
Conclusions

Present **cross-modal input ablation**
Conclusions

Present **cross-modal input ablation**

- Straightforward to perform + easy to interpret + no intervention in the model 😊
Conclusions

Present **cross-modal input ablation**

- Straightforward to perform + easy to interpret + no intervention in the model 😊

Pretrained V&L Transformers are **asymmetric**
Conclusions

Present **cross-modal input ablation**

- Straightforward to perform + easy to interpret + no intervention in the model 😊

Pretrained V&L Transformers are **asymmetric**

- They better integrate vision-for-language than language-for-vision
Conclusions

Present *cross-modal input ablation*

- Straightforward to perform + easy to interpret + no intervention in the model 😊

Pretrained V&L Transformers are *asymmetric*

- They better integrate vision-for-language than language-for-vision

- Are current downstream tasks more vision-for-language?
Conclusions

Present **cross-modal input ablation**
- Straightforward to perform + easy to interpret + no intervention in the model 😊

Pretrained V&L Transformers are **asymmetric**
- They better integrate vision-for-language than language-for-vision

- Are current downstream tasks more vision-for-language?
- How do we avoid the silver data trap?
Conclusions

Present **cross-modal input ablation**
- Straightforward to perform + easy to interpret + no intervention in the model 😊

Pretrained V&L Transformers are **asymmetric**
- They better integrate vision-for-language than language-for-vision

- Are current downstream tasks more vision-for-language?
- How do we avoid the silver data trap?

Code, models and data available online
- [github.com/e-bug/cross-modal-ablation](https://github.com/e-bug/cross-modal-ablation)
- [github.com/e-bug/volta](https://github.com/e-bug/volta)
Conclusions

Present **cross-modal input ablation**

- Straightforward to perform + easy to interpret + no intervention in the model 😊

Pretrained V&L Transformers are **asymmetric**

- They better integrate vision-for-language than language-for-vision

- Are current downstream tasks more vision-for-language?
- How do we avoid the silver data trap?

Code, models and data available online

- [github.com/e-bug/cross-modal-ablation](https://github.com/e-bug/cross-modal-ablation)
- [github.com/e-bug/volta](https://github.com/e-bug/volta)