Towards a better Understanding of Vision-Language Transformer Models

Emmanuelle Salin

TALEP
Laboratoire d'Informatique et Systèmes

JTT - 11 mai 2023

Plan de la Présentation

Introduction

Vision-Language Multimodality

Vision-Language models combine information from the visual and textual modalities to create multimodal representations.

Figure – A flying specimen of Amegilla cingulata, Australia. © Jenny Dettrick, Getty Images

Some real-world applications : computer aided diagnosis, image-text retrieval for the news domain, aid for visually impaired people

Since 2019, varied models have been developed, with different architectures and pre-training tasks.

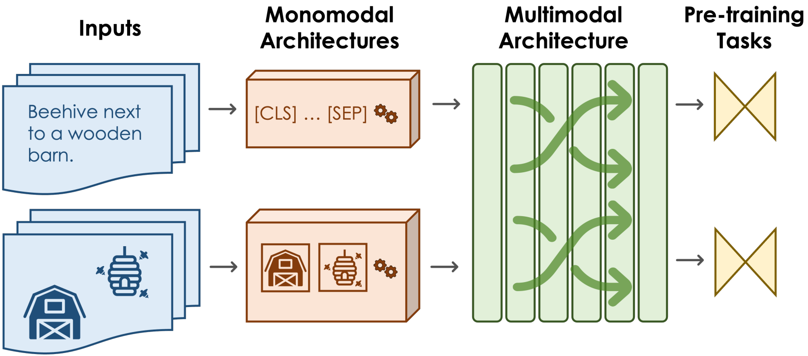


Figure – Pre-training of a Vision-Language Transformer Model

Architecture

Some models use single-stream architectures with early fusion, while others use other types of multimodal fusion.

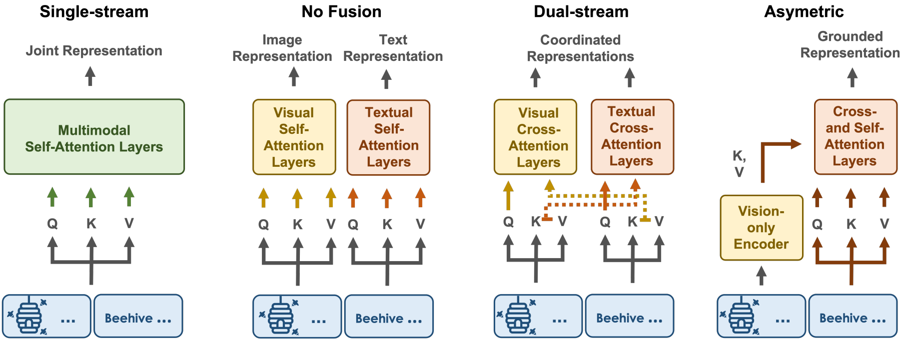


Figure – Different types of cross-modal architecture for vision-language models

Pre-training Tasks

Vision-Language models are pre-trained on textual, visual and multimodal tasks.

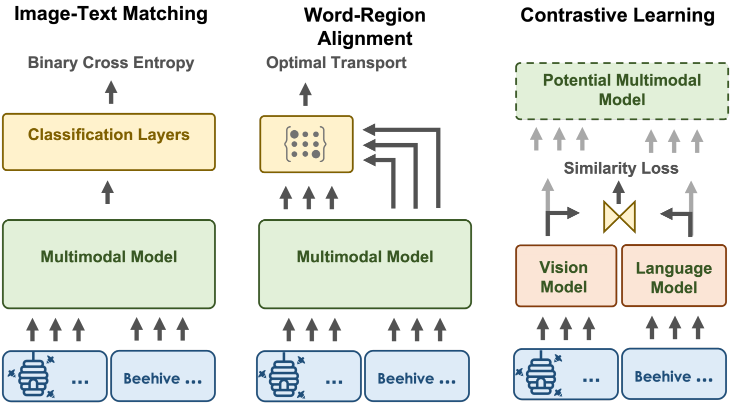


Figure – Examples of multimodal pre-training tasks

Conclusion

It is difficult to compare different models due to the different pre-training protocols.

Thus, our understanding of the strengths and weaknesses of vision-language models is still limited, and several questions remain :

- Architecture : single-stream vs dual-stream
- Pre-training tasks : Image-Text Matching or Contrastive Learning
- Pre-training dataset : descriptive datasets or web-crawled datasets

We aim to reach a better explainability of vision-language models.

Vision-Language Pre-training Datasets

Introduction

Large-scale datasets have been used in Natural Language Processing and Computer Vision.

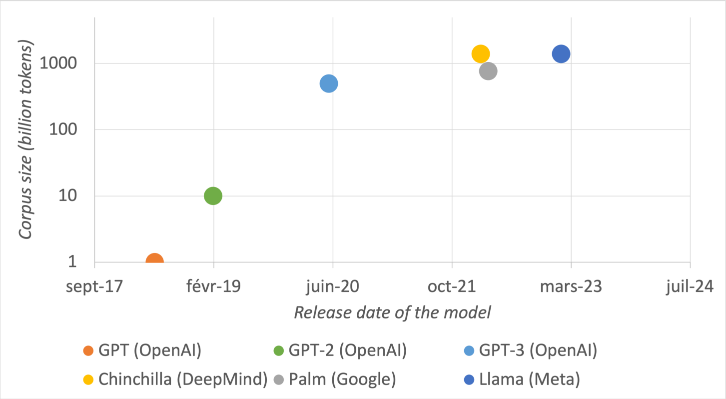


Figure – Evolution of the size of pre-training datasets of transformer-based language models

Manually Annotated Datasets

Vision-language models were at first annotated manually by human annotators.

Name	MS COCO [Lin et al., 2014]	Visual Genome [Krishna et al., 2017]
Nb Images	111k	103k
Nb Texts	558 k	5 millions
Ex. Image		
Ex. Text	A horse carrying a large load of hay and two people sitting on it.	Park bench is made of gray weathered wood

Table – Manually annotated datasets

Automatically Created Datasets

Due to their cost, human annotated datasets are limited in size, which has led to the development of automatically annotated datasets.

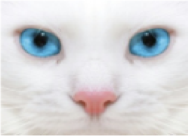
Name	SBU [Ordonez et al., 2011]	Conceptual Captions [Sharma et al., 2018]	LAION [Schuhmann et al., 2021]
Size	1 million	3/12 millions	0.4/6 billions
Ex. Image			
Ex. Text	Man sits in a rusted car buried in the sand on Waitarere beach	a worker helps to clear the debris.	cat, white, and eyes image

Table – Automatically created datasets

Automatic Filtering of Image-Text Datasets

Automatically collected datasets filter data to ensure the quality of images and their annotations.

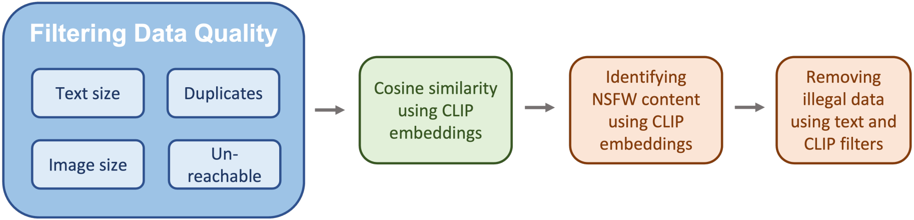


Figure – Example : Filtering process of the Laion-400 dataset

Collecting protocols have also used object detectors to ensure the compatibility of the text-image pair, limiting the size of the dataset.

Analysis : Pre-training Data and Model Performance

Compiling different studies on vision-language models and their pretraining datasets, we determine several factors that impact performance :

Analysis : Pre-training Data and Model Performance

Compiling different studies on vision-language models and their pretraining datasets, we determine several factors that impact performance :

- **Variability** : Dataset should have a diverse range of object categories

Analysis : Pre-training Data and Model Performance

Compiling different studies on vision-language models and their pretraining datasets, we determine several factors that impact performance :

- **Variability** : Dataset should have a diverse range of object categories
- **Accuracy** : Text and image of a same instance should correctly match

Analysis : Pre-training Data and Model Performance

Compiling different studies on vision-language models and their pretraining datasets, we determine several factors that impact performance :

- **Variability** : Dataset should have a diverse range of object categories
- **Accuracy** : Text and image of a same instance should correctly match
- **Compositionality** : Text and image should not only represent simple scene structures

Analysis : Pre-training Data and Model Performance

Compiling different studies on vision-language models and their pretraining datasets, we determine several factors that impact performance :

- **Variability** : Dataset should have a diverse range of object categories
- **Accuracy** : Text and image of a same instance should correctly match
- **Compositionality** : Text and image should not only represent simple scene structures
- **Bias** : Datasets should avoid bias as much as possible

Analysis : Pre-training Data and Model Performance

Compiling different studies on vision-language models and their pretraining datasets, we determine several factors that impact performance :

- **Variability** : Dataset should have a diverse range of object categories
- **Accuracy** : Text and image of a same instance should correctly match
- **Compositionality** : Text and image should not only represent simple scene structures
- **Bias** : Datasets should avoid bias as much as possible
- Similarity between pre-training and fine-tuning

Improving Pre-training Dataset Quality

- Metrics can be computed to evaluate the quality of a subset of the dataset :
  - Vocabulary variability ratio
  - Evaluation of syntax structure
  - N-gram frequency in the dataset
  - Number of objects in the images
  - Proportion of object categories in the dataset

Improving Pre-training Dataset Quality

- Metrics can be computed to evaluate the quality of a subset of the dataset :
  — Vocabulary variability ratio
  — Evaluation of syntax structure
  — N-gram frequency in the dataset
  — Number of objects in the images
  — Proportion of object categories in the dataset

- Filtering methods can improve dataset quality :
  — Named Entity pseudonymisation to limit bias
  — Social media text cleaning methods
  — Use of computer vision models to assess image quality
  — Use of models pre-trained on different domains to assess text-image matching

Comparison of MS COCO and LAION

We study a subset of the MS COCO and LAION-400 datasets :

- Most frequent set of 2 words :
  - LAION : (stock, photo)
  - COCO : (group, people)
- Most frequent detailed part of speech tags (Spacy) :
  - LAION : Noun, Proper noun, Adjective
  - COCO : Noun, Determiner, Conjunction
- Most frequent dependency labels (Spacy) :
  - LAION : Compound modifier, Punctuation
  - COCO : Determiner, Prepositional modifier
- Vocabulary factor (Vocabulary | Number of words) : LAION : 0.10, COCO : 0.04
- Median number of objects per image : LAION : 1, COCO : 4

Ethical issues

Beyond the quality of datasets, several ethical issues are raised with the use of such datasets :

Ethical issues

Beyond the quality of datasets, several ethical issues are raised with the use of such datasets :

- Privacy and consent

Ethical issues

Beyond the quality of datasets, several ethical issues are raised with the use of such datasets :

- Privacy and consent
- Annotator well-being

Ethical issues

Beyond the quality of datasets, several ethical issues are raised with the use of such datasets :

- Privacy and consent
- Annotator well-being
- Harmful biases (ex : representational bias)

Ethical issues

Beyond the quality of datasets, several ethical issues are raised with the use of such datasets :

- Privacy and consent
- Annotator well-being
- Harmful biases (ex : representational bias)

Conclusion

- Several characteristics of a pre-training dataset can impact model performance : Variability, Accuracy, Compositionality, Bias.

- The use of minimal filtering methods can create less than optimal datasets, at high economical and environmental cost. Finer filtering methods should be developed to help improve the quality of multimodal pre-training datasets.

Conclusion

- Several characteristics of a pre-training dataset can impact model performance : Variability, Accuracy, Compositionality, Bias.
- The use of minimal filtering methods can create less than optimal datasets, at high economical and environmental cost. Finer filtering methods should be developed to help improve the quality of multimodal pre-training datasets.
- More attention should be paid to the quality of automatically created datasets, as well as their societal impact.

A Taxonomy of Vision-Language Capabilities

Introduction

There is no consensus on what capacities a vision-language model should be evaluated on, which can make it difficult to identify unknown weaknesses.

Indeed, vision-language models can be used in a multitude of applications requiring various skills :

Figure – Multimodal applications for vision-language models

Evaluation Methods

Evaluation methods have been used to compare the performance of the models.

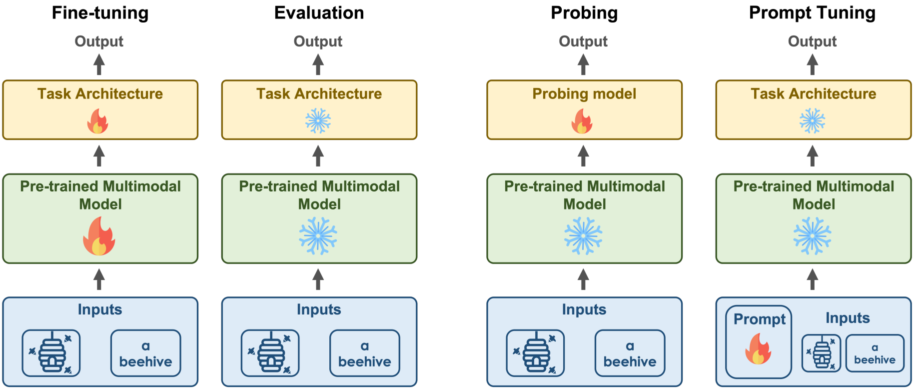


Figure – Different methods of evaluation for Vision-Language Transformers

Multimodal Capabilities

We create a set of various vision-language capabilities from various sources :

- Existing work studying vision-language capabilities (Ex : Position, counting tasks)
- Existing vision-language evaluation tasks assessing specific skills (Ex : Medical VQA)
- Analysis of vision-language datasets for specific applications (Ex : News datasets)
- Natural Language Processing tasks applied to multimodal data (Ex : Natural Language Inference)

Categorization

In order to establish a categorization of vision-language capabilities, we use as inspiration terminology from Visual Literacy : *Denotation* and *Connotation*.

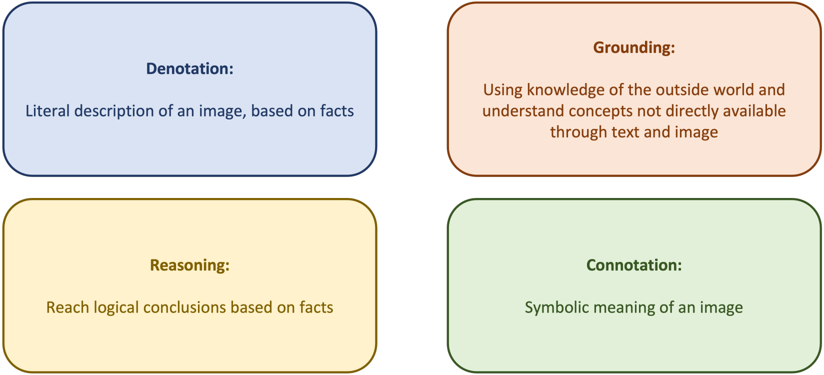


Figure – Classification of vision-language skills into four categories

Denotation

Local	Structural	Global
Basic Element Detection	Syntactic Understanding	Understanding Document Type
Object Perception	Scene Structure Understanding	Focus Identification
Attribute Association	Positional understanding	Context Understanding
Body Language Recognition	Co-reference Resolution	Contradiction detection
Optical Character Recognition	Multimodal Dependency Understanding	

Table – Denotation Skills

Grounding

Temporal Grounding	Spatial Grounding	Knowledge Grounding
Action Classification		Object Role Understanding
Object State Understanding	Spatial Understanding	Named Entity Recognition
Motion Detection	Spatial Extrapolation	Technical Term Recognition
Temporal extrapolation	Location Recognition	Cultural Grounding
Time or Period Identification		

Table – Grounding Skills

Reasoning

Visual Semantic Reasoning	Logical Reasoning	Complex Reasoning
Abnormality Detection	Logical Operations	Extrapolation
Taxonomy Understanding	Multimodal Inference	Multi-hop Reasoning
Polysemy	Comparison	Interactive Reasoning
Structural Inconsistency Detection		Explainability

Table – Reasoning Skills

Connotation

Meaning Understanding	Quality Evaluation
Iconography Understanding	Stylistic Appreciation Evaluation
Style Understanding	Consistency Evaluation
Ambiguity Understanding	Effectiveness Evaluation
Sentiment Understanding	
Cultural Meaning Understanding	

Table – Connotation Skills

Taxonomy and Evaluation Tasks

The taxonomy can be used to highlight gaps in the evaluation of vision-language models.

Reasoning	Logical	E-SNLI-VE [22]	Vision-Language inference
		NLVR2 [68]	Natural Language Visual Reasoning on two images
	Complex	E-vil [35], VCR [81]	Natural language explanations for Visual Question Answering
		VQA-HAT [18]	Visual explanations for Visual Question Answering
		GuessWhat?! [20]	Visual Guess What? Game
		Visual Dialog [19]	Dialog with visual context
		IQUAD, EmbodiedVQA [77], Spoon [3]	Interactive Visual Question Answering
		FashionIQ [78]	Dialog for Fashion recommendation

Figure – Existing datasets translated on part of the taxonomy

Conclusion

- We create a taxonomy of vision-language capabilities
- This taxonomy can be used to highlight lacking aspects of vision-language evaluation
- The goal is to develop more exhaustive evaluation methods for vision-language foundation models

Creating an Evaluation Task

Guidelines to Creating a Multimodal Evaluation Task

- Choose a task to evaluate the appropriate level of multimodal understanding : for example, create positive and negative pairs with minimal differences
- Pay attention to textual bias : create balanced datasets that avoid spurious biases
- Carefully select difficulty levels (image quality, details, complexity)
- Try to avoid representational bias (do an analysis of the dataset)
- Be aware of the possible subjectivity of the task (i.e. label depends on the annotator)

Evaluation Task : Hypernym understanding

Using the previous taxonomy, we identified gaps with no existing dataset. For example, few tasks evaluate skills related to *Semantic Reasoning*.

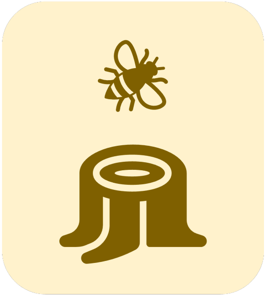


Figure – Caption 1 : A bee above tree trunk.
Caption 2 : An insect above a tree trunk

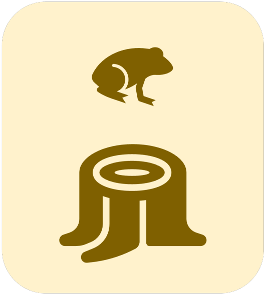


Figure – Caption 1 : A frog above tree trunk.
Caption 2 : An amphibian above a tree trunk

The Checklist Methodology [Ribeiro et al., 2020]

A new method to test model capabilities inspired by and behavior.

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** — Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** — Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos / neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral / neg	X
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR*** — Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg / neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg / neutral	X
...			
Failure rate = 34.6%			

Figure – Examples of minimal function, invariance and directional tests applied to a sentiment analysis model [Ribeiro et al., 2020]

Robustness

Robustness consists in checking whether a model's performance in a specific task is not affected by unrelated information.

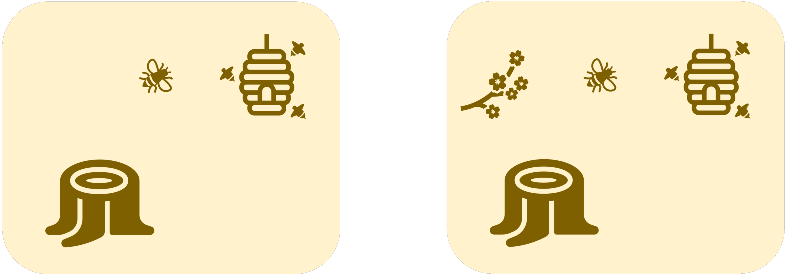


Figure – Question : 'How many bees are in the picture ?' Answer : 4
The flower branch should not impact the prediction

Consistency

Consistency consists in checking whether a model shows coherent behavior by correctly interpreting the change in semantics between two slightly semantically different instances.

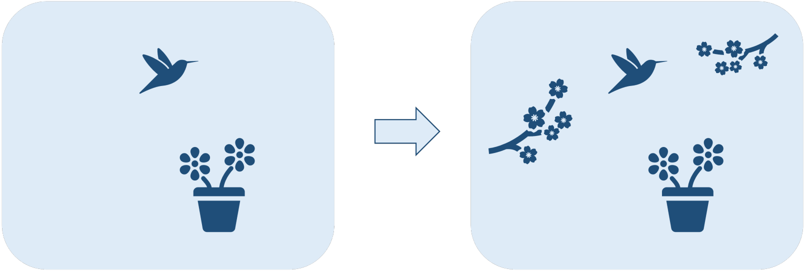


Figure – Caption : 'A bird flying surrounded by flowers'
The second image should be more related to the caption than the first.

Future Work : Dataset Creation

The use of a synthetic images enables a greater control of the content of each image, in particular for smaller variations of data.

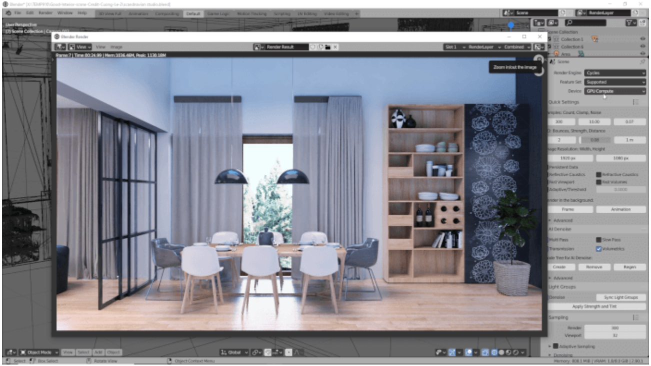


Figure – Synthetic scene created through blender [1]

1. https ://www.blender3darchitect.com/furniture-models/scandinavian-studio-free-interior-scene-with-settings-for-cycles/

Conclusion

- Using the taxonomy, we can identify gaps in the evaluation of vision-language models
- When possible, it is interesting to check the behavior of the model in terms of robustness and consistency
- We want to create a synthetic dataset to precisely evaluate capabilities and check the robustness and consistency of the models

References

Blanchard, G., Neuvial, P., and Roquain, E. (2020).
Post hoc confidence bounds on false positives using reference families.
pages 1281–1303.

Durand, G., Blanchard, G., Neuvial, P., and Roquain, E. (2020).
Post hoc false positive control for structured hypotheses.
pages 1114–1148.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017).
Visual genome : Connecting language and vision using crowdsourced dense image annotations.
International journal of computer vision, 123 :32–73.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014).
Microsoft coco : Common objects in context.
In *Computer Vision–ECCV 2014 : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Ordonez, V., Kulkarni, G., and Berg, T. (2011).
Im2text : Describing images using 1 million captioned photographs.
Advances in neural information processing systems, 24.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020).
Beyond accuracy : Behavioral testing of NLP models with CheckList.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. (2021).
Laion-400m : Open dataset of clip-filtered 400 million image-text pairs.
arXiv preprint arXiv :2111.02114.