



# Sondage des représentations Vision-Langage

Badreddine Farah

Encadrants Lis-Lab : Benoit Favre

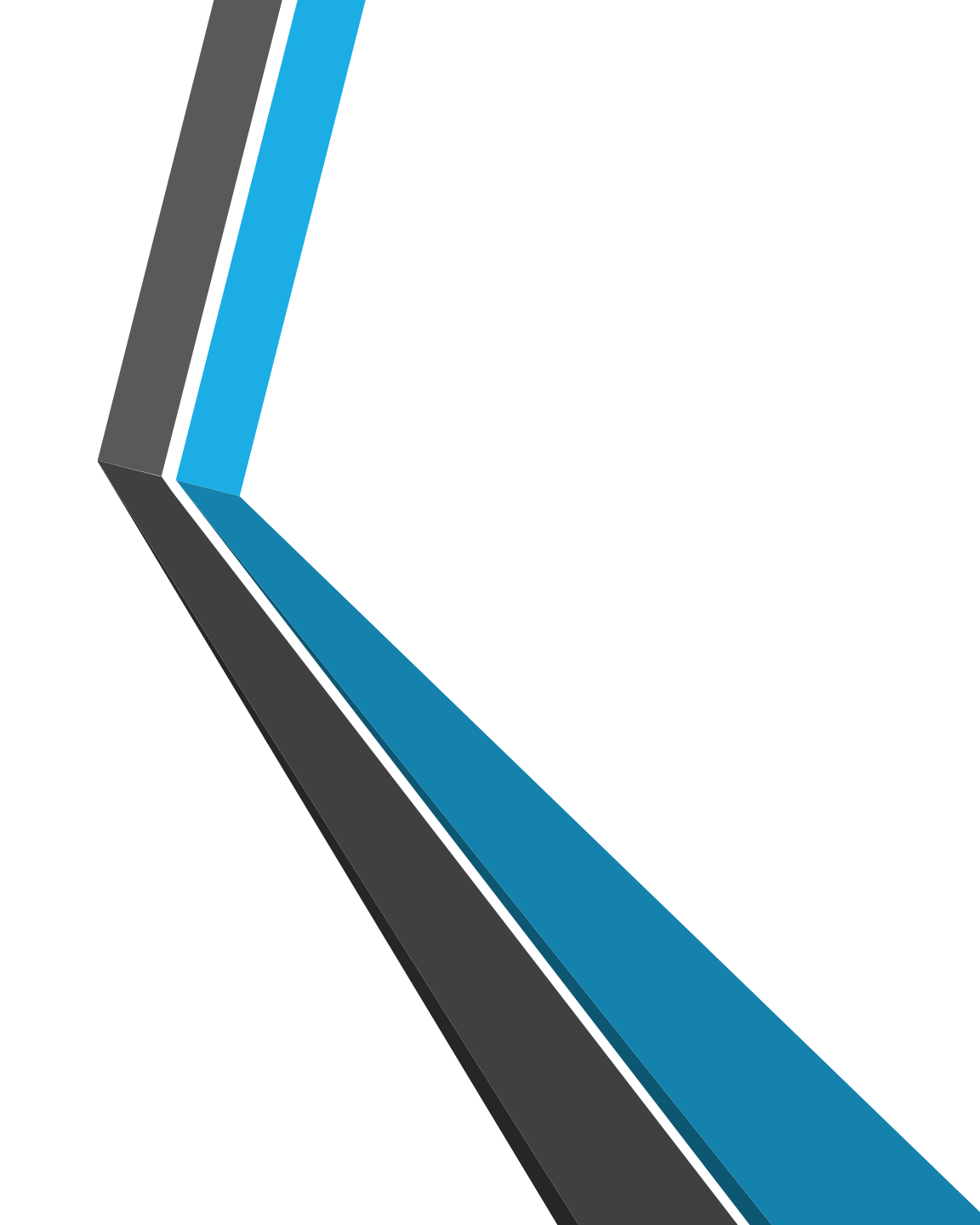
Emmanuelle Salin

Stéphane Ayache

# Plan

---

- Contexte du stage
- Sondage (Probing)
- Expériences et Résultats
- Conclusion



# Contexte du Stage

- Introduction.
- Le Transformer
- Modèles vision et langage.

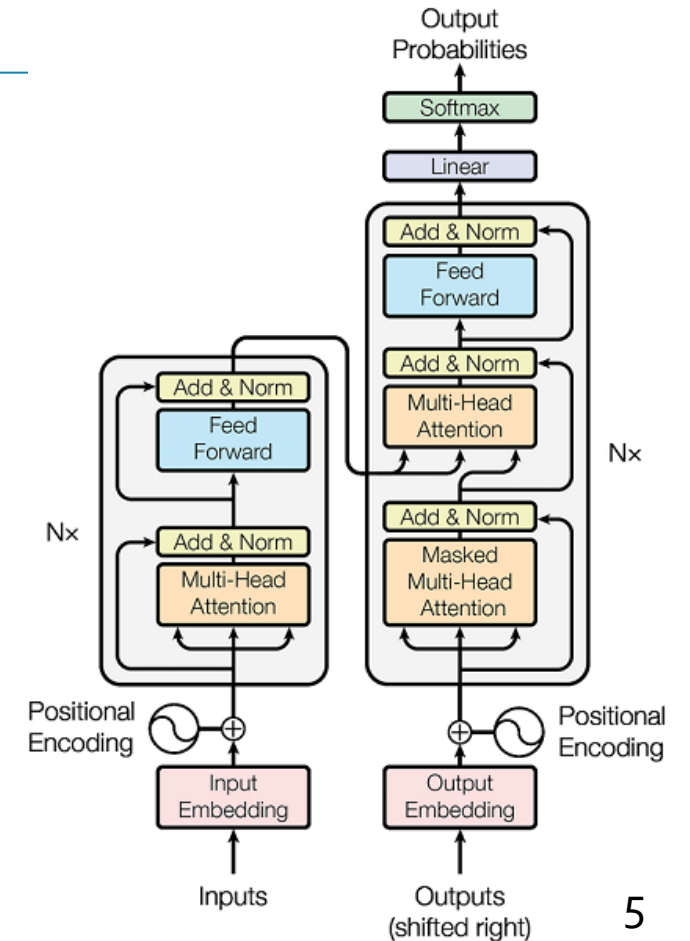
# Introduction

---

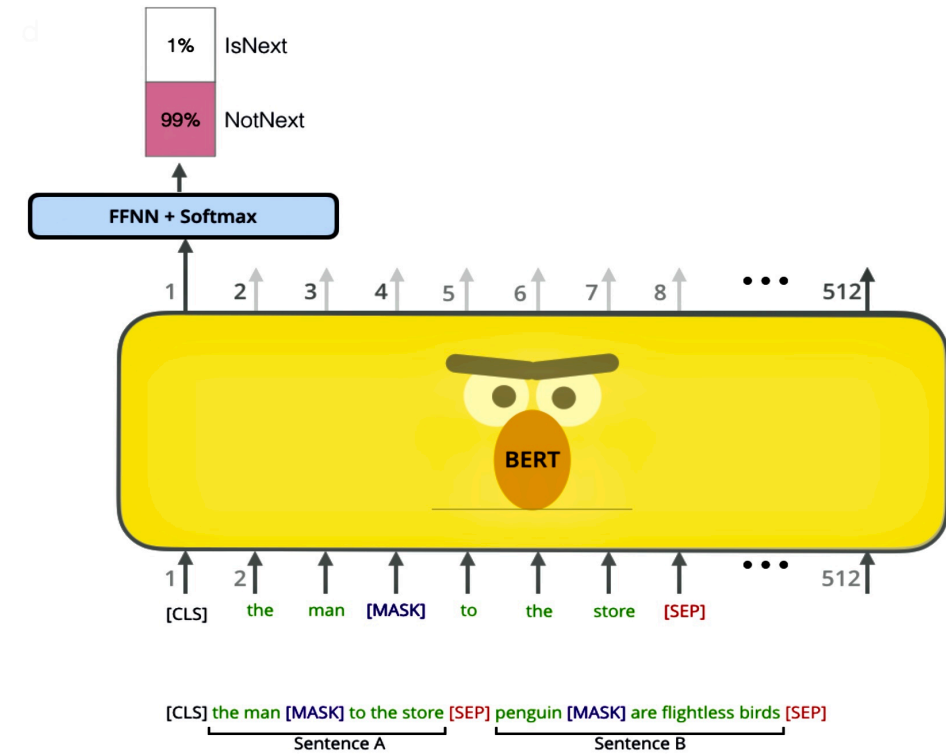
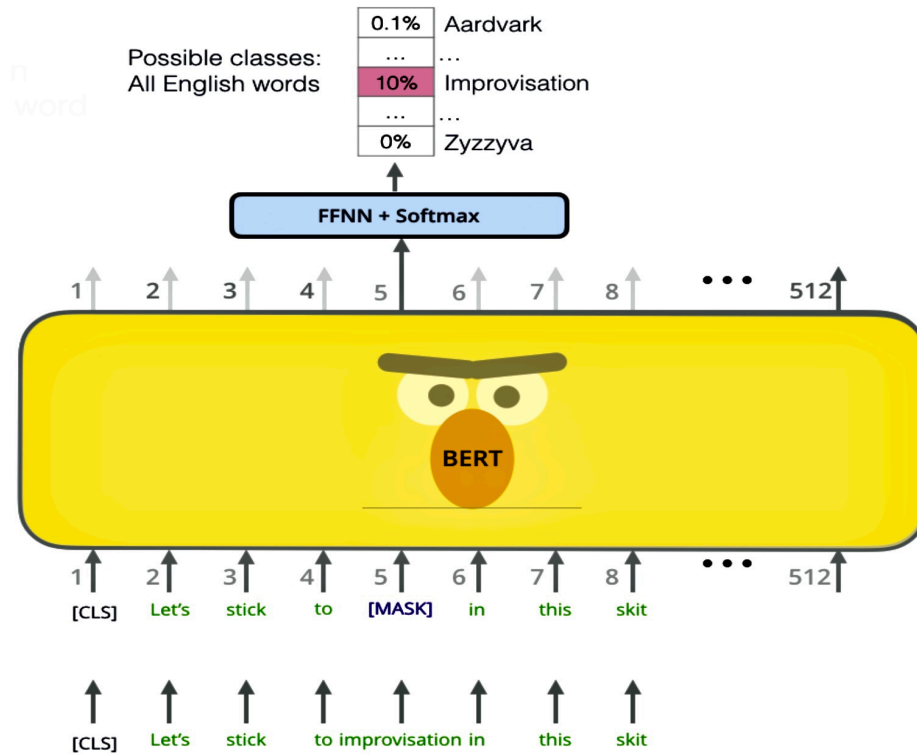
- Analyses des modèles vision-langage
  - Analyse des modèles prés-entraînés.
  - Analyse des modèles ajustés (finetuned).
  - Études des différences entre les modèles.
- Utilisation du sondage pour mieux comprendre les modèles.

# Le Transformer

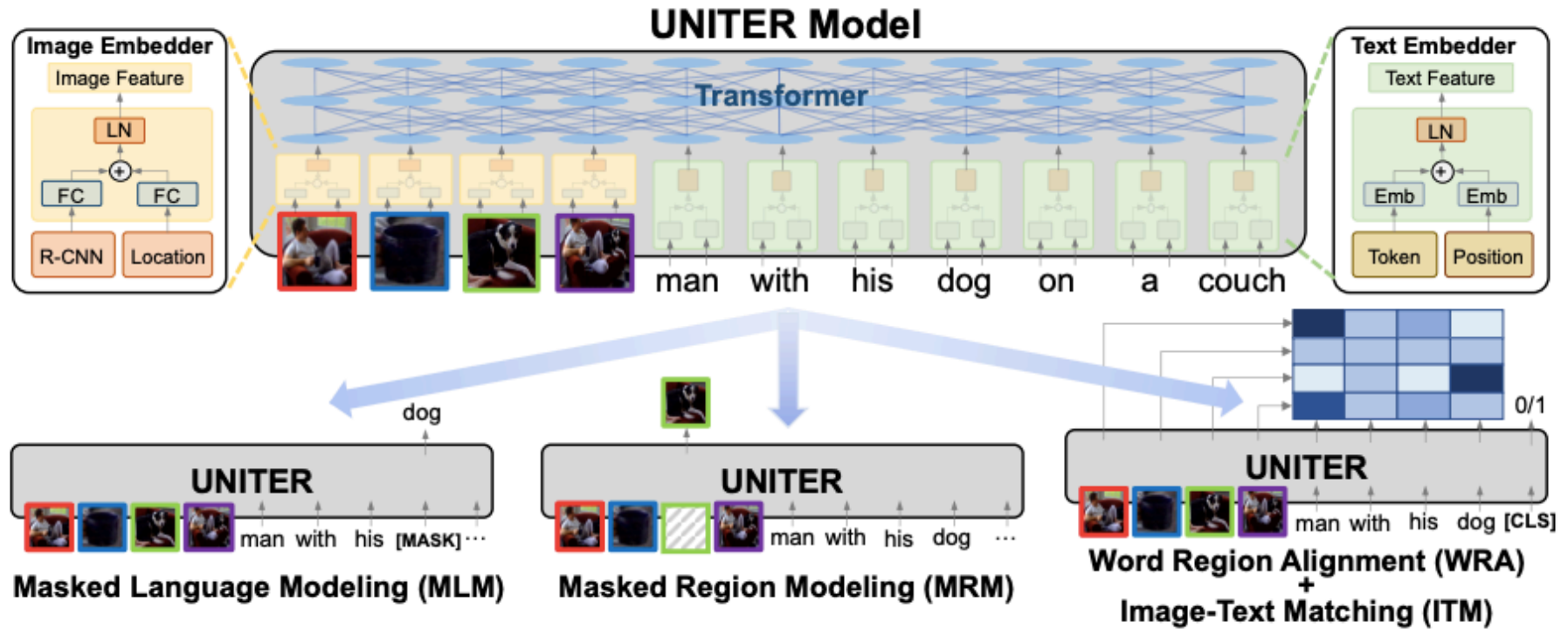
- Introduit en 2017 pour l'utilisation en traduction automatique.
- Généralisation en TAL avec BERT.
- Entraînement avec de nouvelles méthodes (auto-supervision).



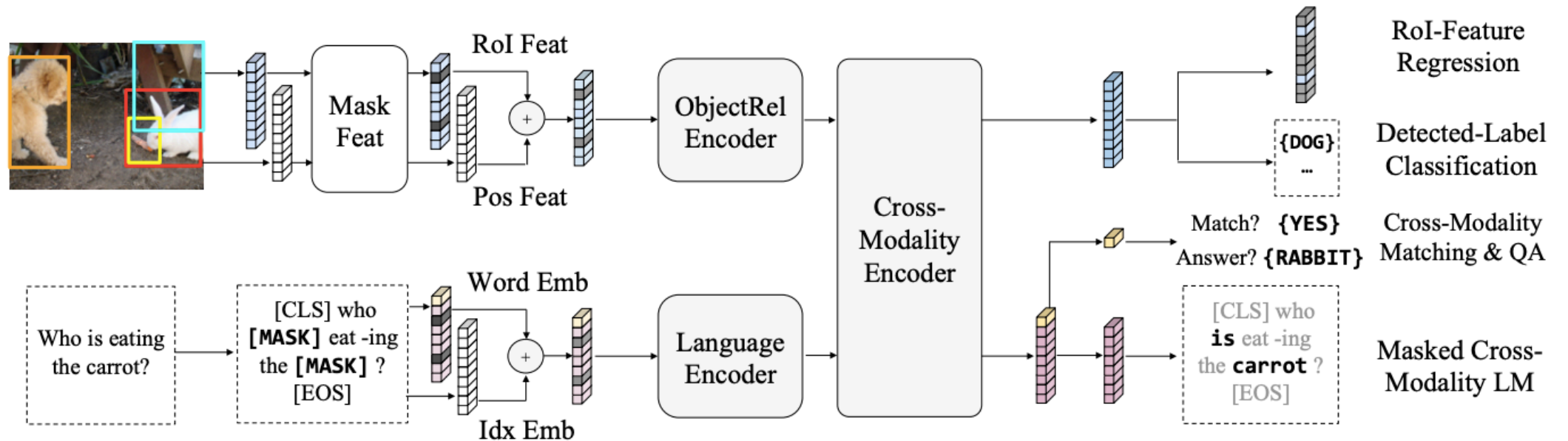
# BERT



# Modèles Vision-langage (UNITER)



# Modèles Vision-langage (LXMERT)



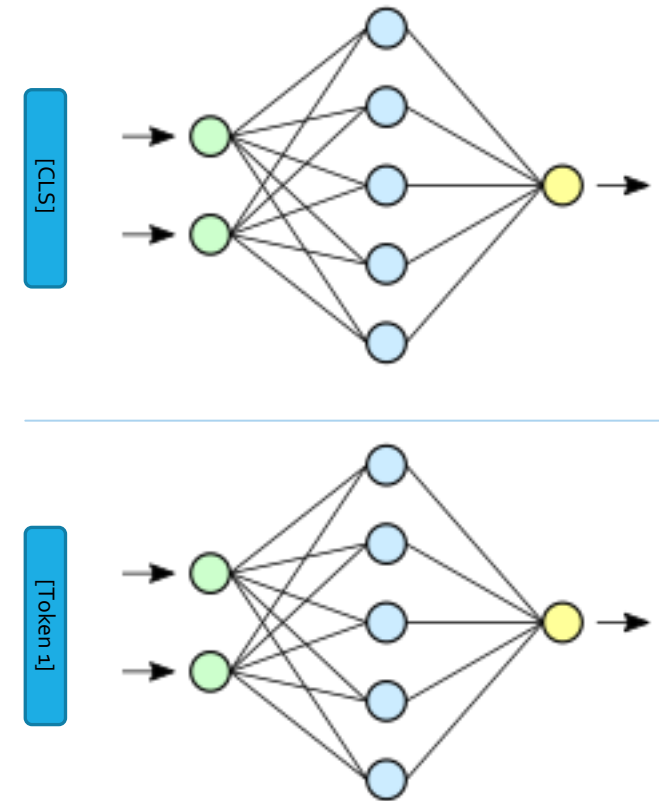
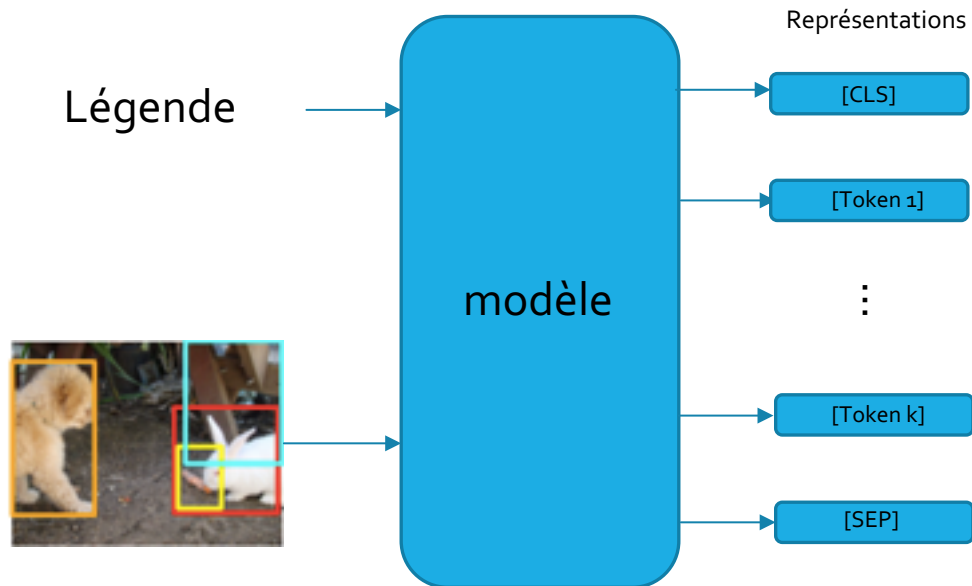




# Le sondage (probing)

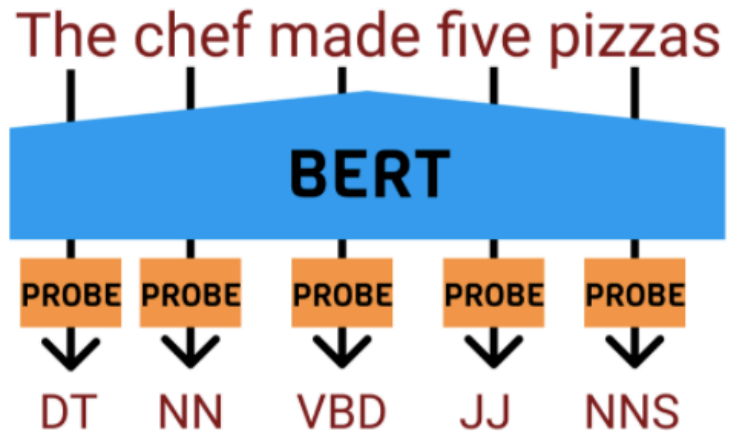
- Définition du sondage.
- Pourquoi le sondage?
- Sondage et modèles V&L

# Comment peut-on analyser les représentations?

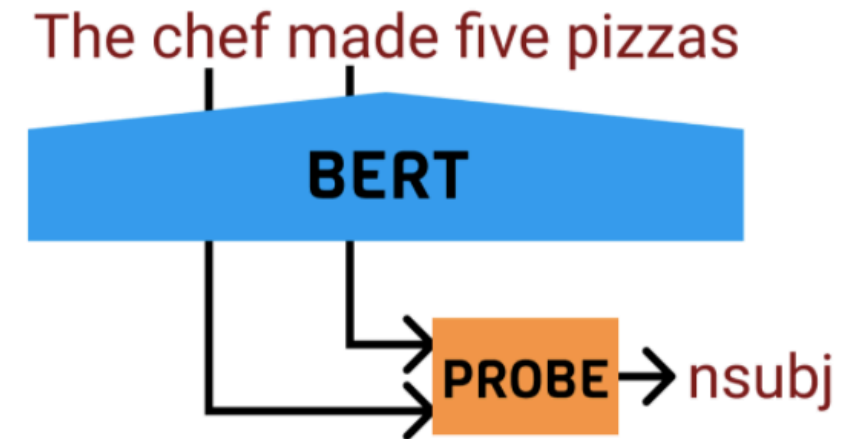


## Sondage (Exemples)

### Part-of-speech!



### Partial dependency info!



# Pourquoi le sondage ?

---

- Une meilleure compréhension de la nature des représentations non-supervisées.
- Comprendre pourquoi les modèles réussissent ou échouent.
- Tester les intuitions sur la concordance entre les propriétés des représentations et les propriétés linguistiques ou les propriétés Vision.

# Sondage et modèles V&L

---

- Pour bien comprendre le fonctionnement des modèles V&L, le sondage doit se faire sur trois niveaux :
  - Vison.
  - Langage.
  - Multimodalité.



# Expériences et résultats

- Méthodologie du Sondage.
- Sondage au niveau Langage.
- Sondage au niveau Vision.
- Sondage au niveau multimodal.

# Méthodologie du Sondage

- Pour sonder un modèles :
  - On calcule les représentations avec les modèles pré-entraînés et ajustés
  - On entraine des classifieurs linéaire (sondes) .
  - On note les performances de testes sur ces classifieurs.
- Utiliser des données décorrélées ( image ne correspond pas au texte).
- Utilisation de BERT, RESNET et ViT comme références.

# Taches de sondage

| Task              | Representation | Type (Metric)                | Dataset    | Instances |
|-------------------|----------------|------------------------------|------------|-----------|
| L-Part-Of-Speech  | Word tokens    | Classification (acc.)        | Flickr     | 2000/1000 |
| L-Bigram Shift    | [CLS] token    | Binary classification (acc.) | Flickr     | 2000/1000 |
| V-Flower          | [CLS] token    | Classification (acc.)        | Flower-102 | 6149/1020 |
| V-Object Counting | [CLS] token    | Regression (MSE)             | MS-COCO    | 1800/624  |
| M-Color           | [MASK] token   | Classification (acc.)        | Flickr     | 2000/1000 |
| M-Size            | [CLS] token    | Binary Classification (acc.) | Flickr     | 1800/752  |
| M-Position        | [CLS] token    | Binary Classification (acc.) | Flickr     | 1800/824  |
| M-Adversarial     | [CLS] token    | Binary Classification (acc.) | MS-COCO    | 500/200   |



## Résultats du Sondage Langage

|                   | UNITER       | LXMERT       | BERT         |
|-------------------|--------------|--------------|--------------|
| L-Part-Of-Speech  | 94.66        | 95.13        | <b>95.57</b> |
| L-POS wrong image | 90.86        | <b>95.36</b> | -            |
| L-Bigram Shift    | 80.89        | 70.65        | <b>86.33</b> |
| L-BS wrong image  | <b>76.25</b> | 72.22        | -            |

|                   | UNITER       |              | LXMERT       |       |
|-------------------|--------------|--------------|--------------|-------|
| Fine-tuning task  | VQA          | NLVR2        | VQA          | NLVR2 |
| L-Part-Of-Speech  | 93.84        | <b>94.37</b> | 94.14        | 88.44 |
| L-POS wrong image | 93.80        | 94.38        | <b>94.73</b> | 88.49 |
| L-Bigram Shift    | <b>79.48</b> | 72.74        | 65.40        | 57.10 |
| L-BS wrong image  | <b>76.92</b> | 72.34        | 62.74        | 57.82 |

# Résultats du Sondage Vision

|          | UNITER | LXMERT | Resnet | VIT          |
|----------|--------|--------|--------|--------------|
| V-Flower | 71.82  | 75.56  | 86.83  | <b>99.66</b> |

|             | UNITER       |       | LXMERT |       |
|-------------|--------------|-------|--------|-------|
| Fine-tuning | VQA          | NLVR2 | VQA    | NLVR2 |
| V-Flower    | <b>82.91</b> | 82.78 | 78.80  | 74.23 |

|          | UNITER | LXMERT      | BERT | Resnet      | VIT  |
|----------|--------|-------------|------|-------------|------|
| V-OC     | 5.49   | 5.49        | 6.27 | <b>4.96</b> | 5.67 |
| V-OC wr. | 7.20   | <b>7.31</b> | -    | -           | -    |

|                    | UNITER      |             | LXMERT |       |
|--------------------|-------------|-------------|--------|-------|
| Fine-tuning        | VQA         | NLVR2       | VQA    | NLVR2 |
| V-Object Counting  | <b>4.98</b> | 4.95        | 5.13   | 5.65  |
| V-OC wrong caption | 6.49        | <b>6.22</b> | 6.85   | 6.94  |

# Sondage au niveau multimodal



| Task | Input                                            |
|------|--------------------------------------------------|
| M-C  | Two men standing behind a tall [MASK] fence.     |
| M-S  | Two men standing behind a short black fence.     |
| M-P  | Two men standing in front of a tall black fence. |
| M-A  | Two men running behind a tall black fence.       |

# Résultats du Sondage multimodale

|                 | UNITER       | LXMERT       | BERT  | Resnet | ViT   | UNITER       |              | LXMERT       |       |
|-----------------|--------------|--------------|-------|--------|-------|--------------|--------------|--------------|-------|
|                 |              |              |       |        |       | VQA          | NLVR2        | VQA          | NLVR2 |
| M-Color         | <b>86.27</b> | 71.21        | 37.02 | 38.97  | 41.19 |              |              |              |       |
| M-C wr.         | 34.80        | <b>39.33</b> | -     | -      | -     |              |              |              |       |
| M-Size          | 57.15        | <b>58.43</b> | 55.66 | 52.93  | 51.76 |              |              |              |       |
| M-S wr.         | <b>56.06</b> | 55.05        | -     | -      | -     |              |              |              |       |
| M-P             | <b>55.92</b> | 54.62        | 56.52 | 53.03  | 52.78 |              |              |              |       |
| M-P wr.         | 54.06        | <b>54.68</b> | -     | -      | -     |              |              |              |       |
| M-A             | <b>79.71</b> | 72.60        | 53.46 | -      | -     |              |              |              |       |
| M-A wr.         | 51.92        | <b>61.25</b> | -     | -      | -     |              |              |              |       |
| Fine-tuning     |              |              |       |        |       | VQA          | NLVR2        | VQA          | NLVR2 |
| M-Color         |              |              |       |        |       | <b>83.39</b> | 82.52        | 82.60        | 78.00 |
| M-C wrong image |              |              |       |        |       | 35.75        | 36.17        | <b>37.18</b> | 37.02 |
| M-Size          |              |              |       |        |       | <b>64.23</b> | 63.19        | 60.85        | 59.28 |
| M-S wrong image |              |              |       |        |       | 55.96        | <b>59.28</b> | 56.60        | 54.57 |
| M-Position      |              |              |       |        |       | <b>57.55</b> | 56.99        | 56.25        | 56.23 |
| M-P wrong image |              |              |       |        |       | <b>55.27</b> | 54.42        | 54.51        | 55.09 |
| M-Adversarial   |              |              |       |        |       | <b>78.37</b> | 77.50        | 74.90        | 68.46 |
| M-A wrong image |              |              |       |        |       | 49.42        | 53.17        | <b>61.06</b> | 53.46 |

# Conclusion

---

- Les performances multimodales sont très dépendantes du jeux d'entraînement.
- Les modèles s'appuient beaucoup plus sur le langage que la vision.
- UNITER et LXMERT arrivent a extraire des connaissances multimodales sur certains concepts .
- On peut émettre l'hypothèse que le Faster R-CNN est un facteur qui limite les performances multimodales.