Apprentissage Automatique

Questions contemporaines

Stéphane Herbin

stephane.herbin@onera.fr

sommaire

- Algorithmes: « Beyond supervised learning »
  - Apprentissage faible quantité de données (« few shot »)
  - Transfert Multi-modalité (vision & langage → « zero-shot learning »)
  - IA hybride

- Modèles de fondation
  - Une autre architecture: transformer
  - LLM et GPT
  - Vision Language Models

- IA de confiance
  - Processus de ML, normes, éthique
  - Robustesse
  - Explicabilité

« BEYOND SUPERVISED LEARNING »

Autres modalités d’apprentissage

- Fine tuning
- « Few shot learning »
- Incrémental
- Semi-supervisé
- Active learning
- Adaptation de domaine

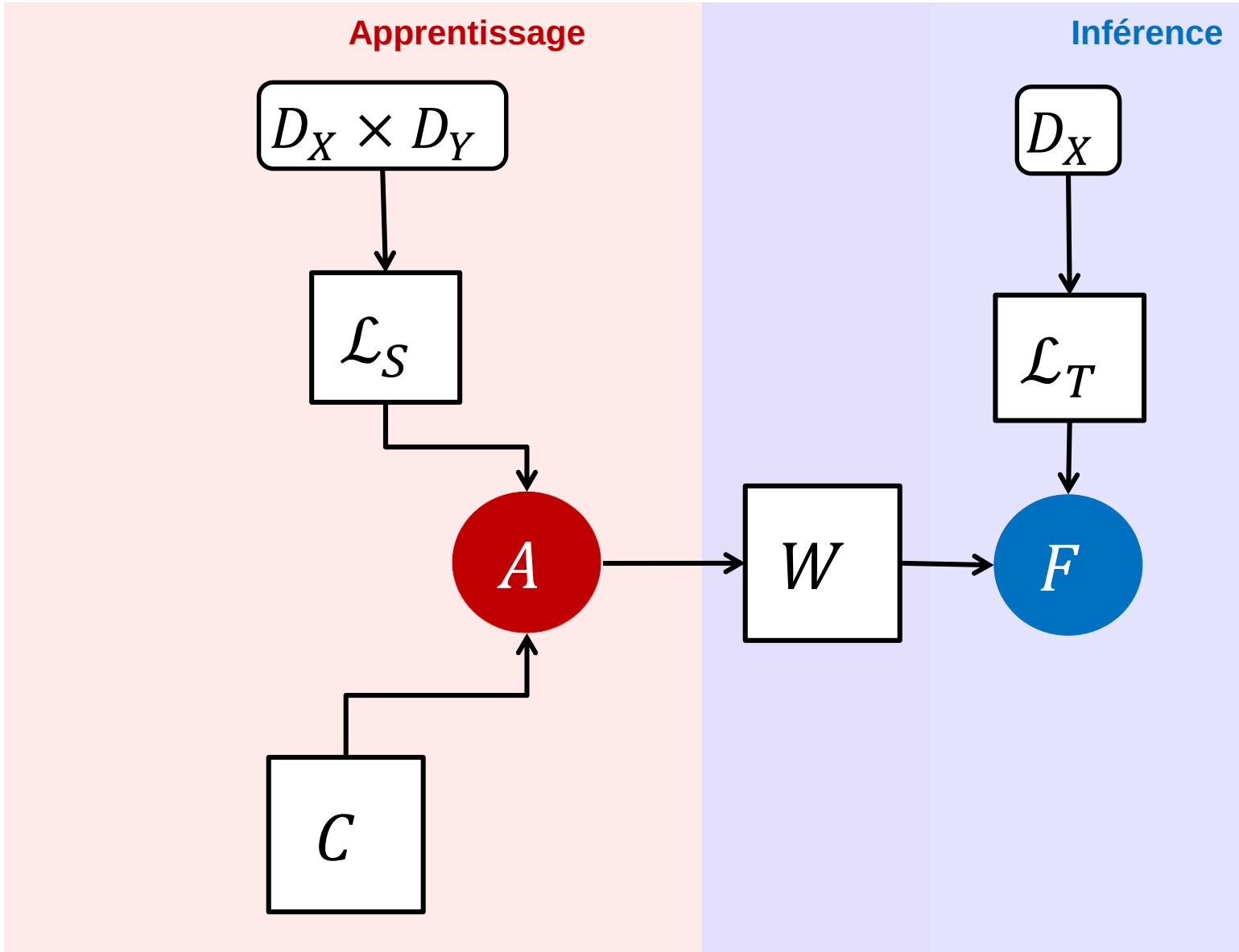
- « Zero-shot learning »
- Open vocabulary

- Meta learning

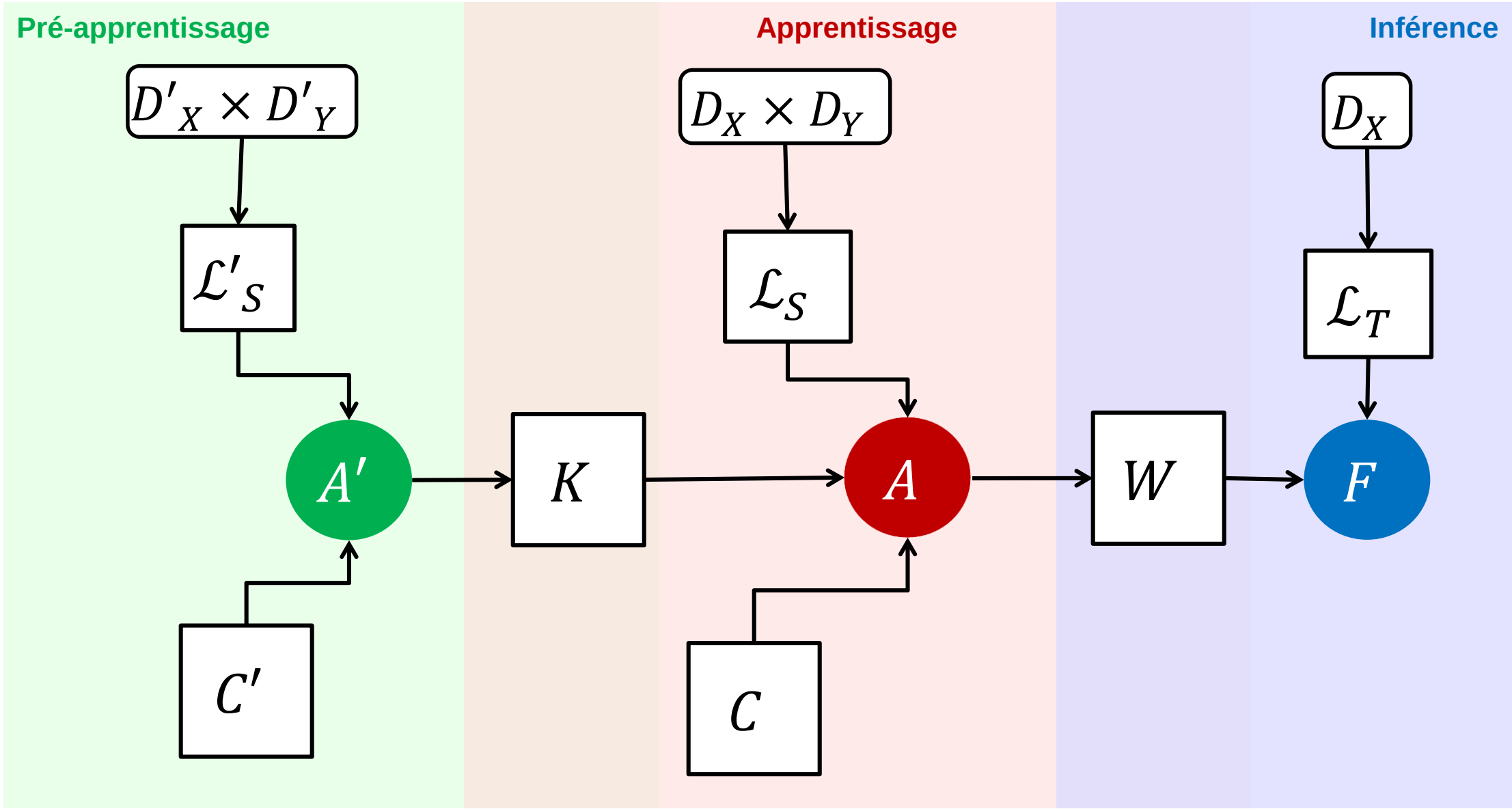
Supervised learning: le cadre standard

Hypothèses

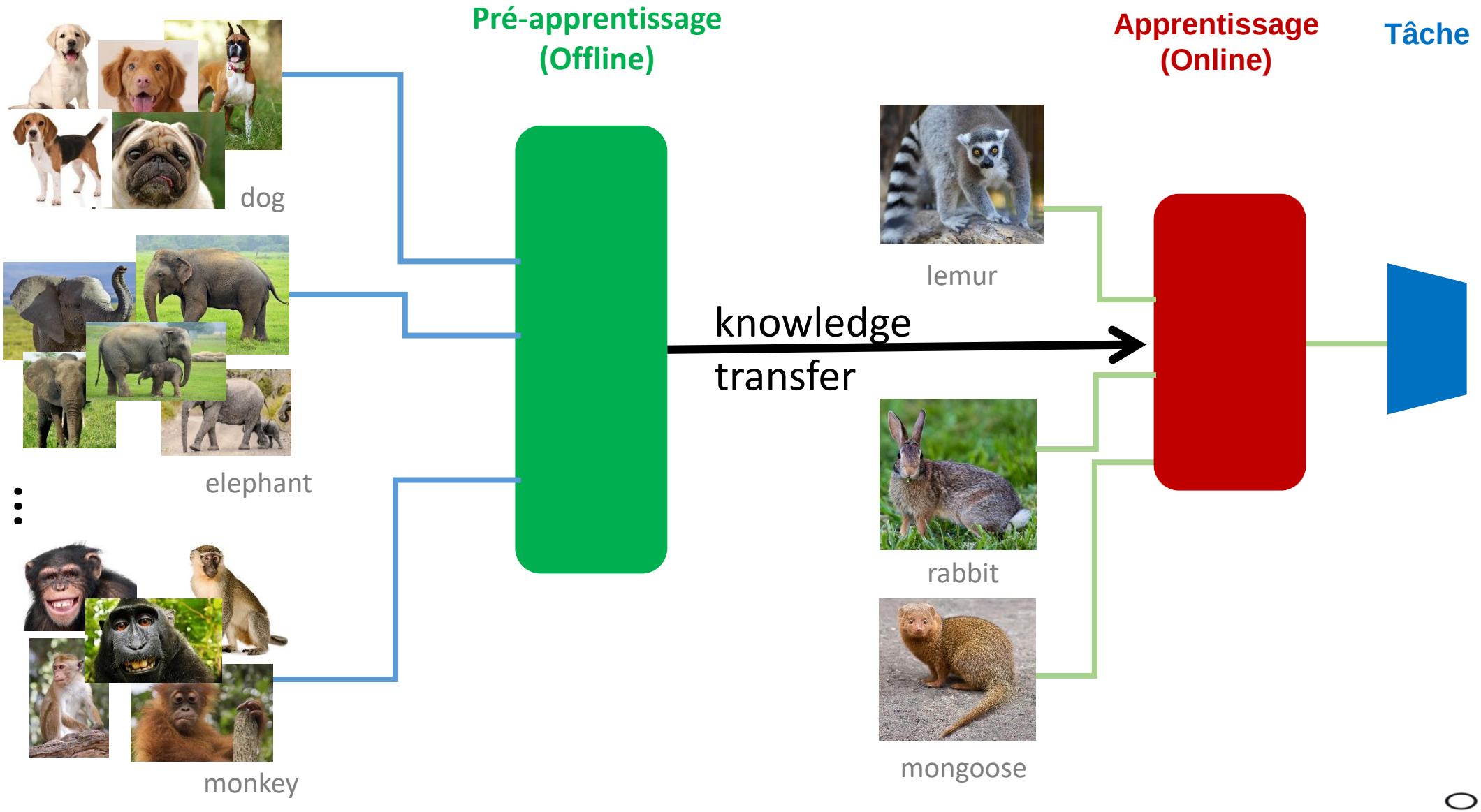
- Même distribution train et test
- Quantité de données « suffisante » pour l'apprentissage



Transfer learning / Fine Tuning / Few-shot learning / Domain adaptation



Few-shot learning



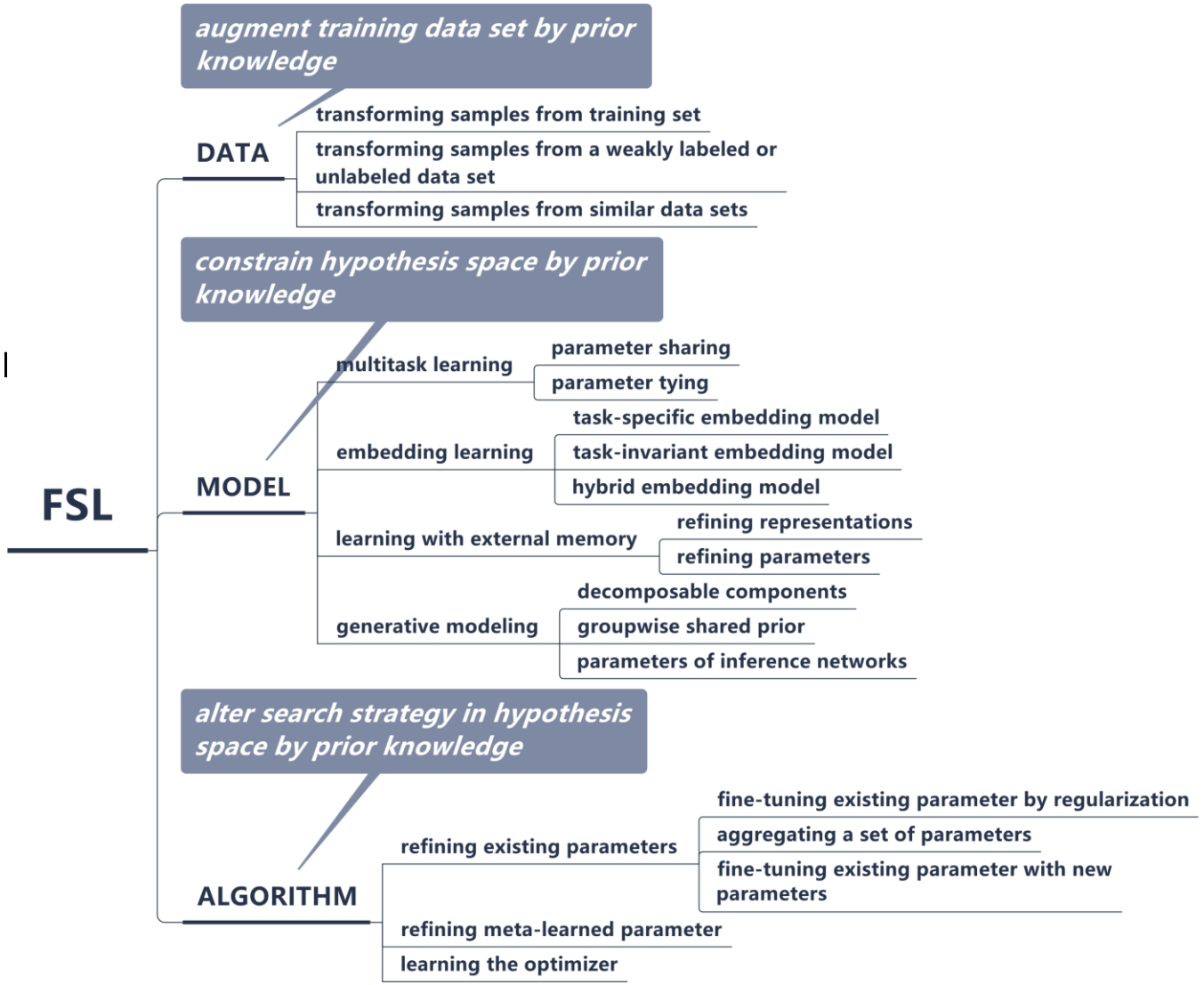
Few shot learning

Few-shot = peu de données

➔ on va manquer d'échantillons pour estimer correctement l'erreur théorique par l'erreur empirique

Trois familles de stratégies pour compenser

- Plus de données (génération, augmentation…)
- Meilleure représentation de données (auto-supervision, multi-tâche…)
- Espace de fonctions régularisé (apprentissage de métrique, meta-learning…)



Meta-Learning: « Learning how to learn »

Algorithm 1 Model-Agnostic Meta-Learning

Require: $p(\mathcal{T})$: distribution over tasks

Require: α, β: step size hyperparameters

1: randomly initialize θ

2: **while** not done **do**

3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$

4: **for all** $\mathcal{T}_i$ **do**

5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to K examples

6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$

7: **end for**

8: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$

9: **end while**

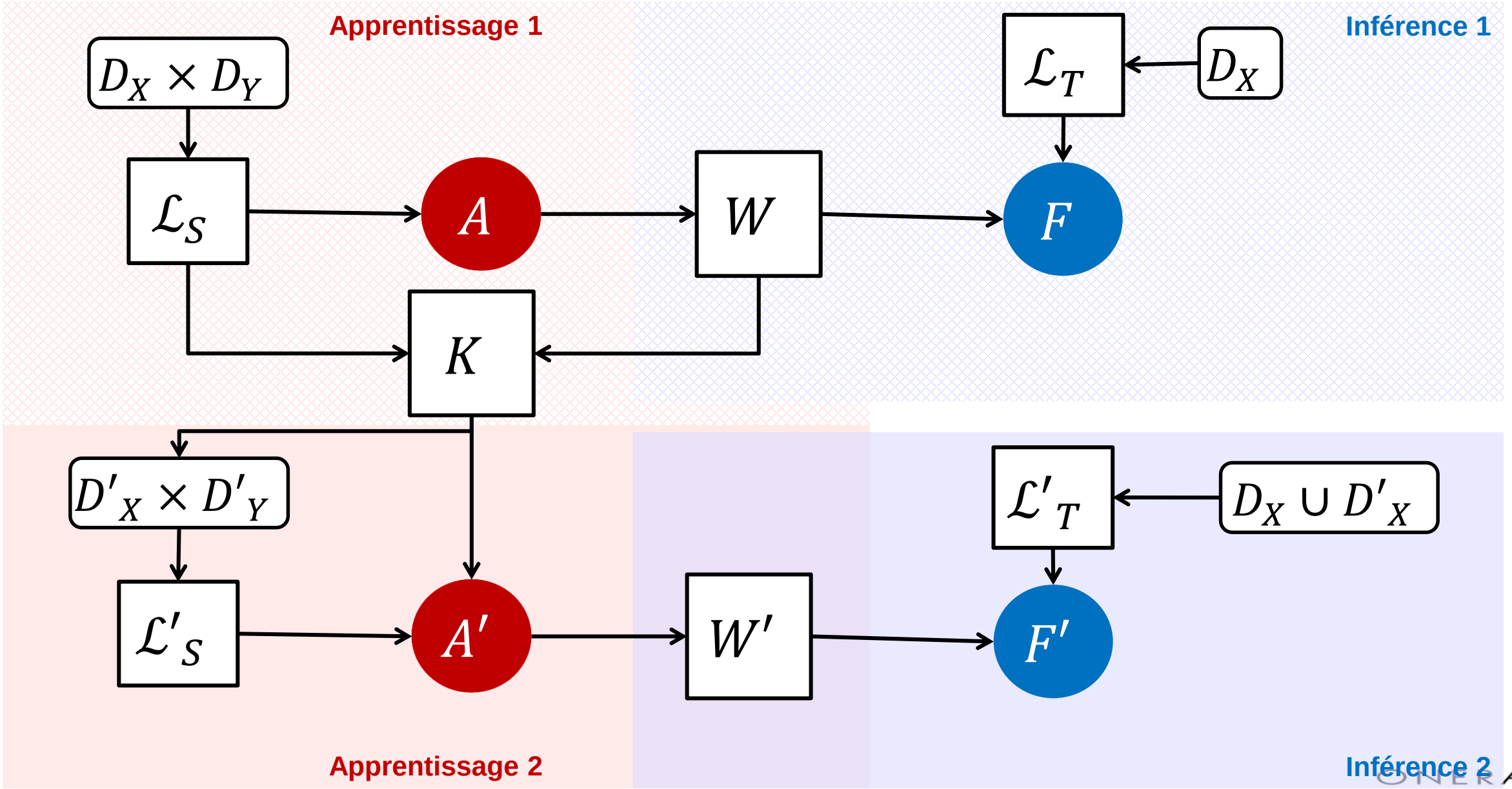
Inner Level/ Adaptation Step

Outer Level/ Meta-Optimization Step

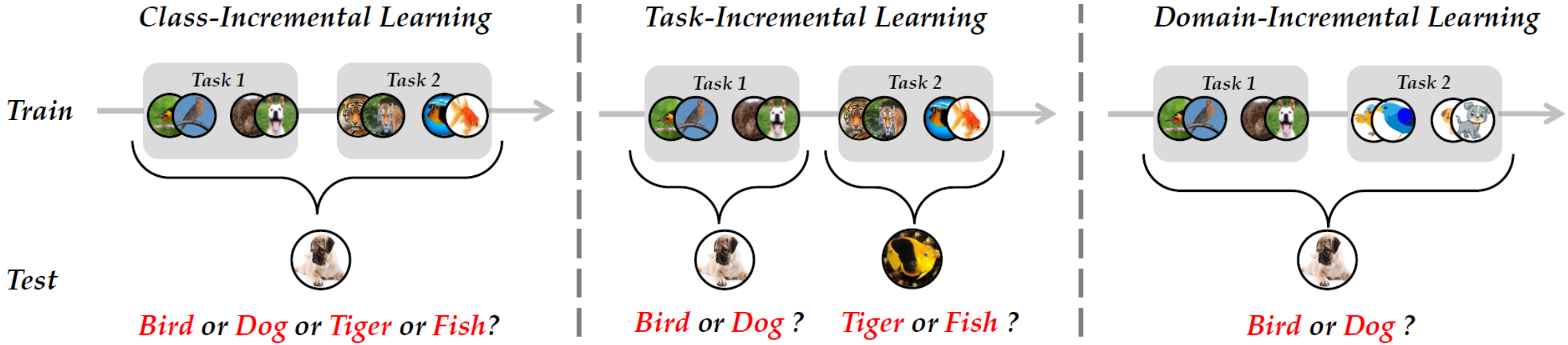
- Model Agnostic Meta Learning (MAML)
- Learn a parameter initialization that can reach a generalizable state for a particular task, after a number of updates
- Episodic learning
- Very unstable!

Antoniou, A., Edwards, H., & Storkey, A. (2019). How to train your MAML. International Conference on Learning Representations.

Apprentissage incrémental / Apprentissage actif



Apprentissage incrémental



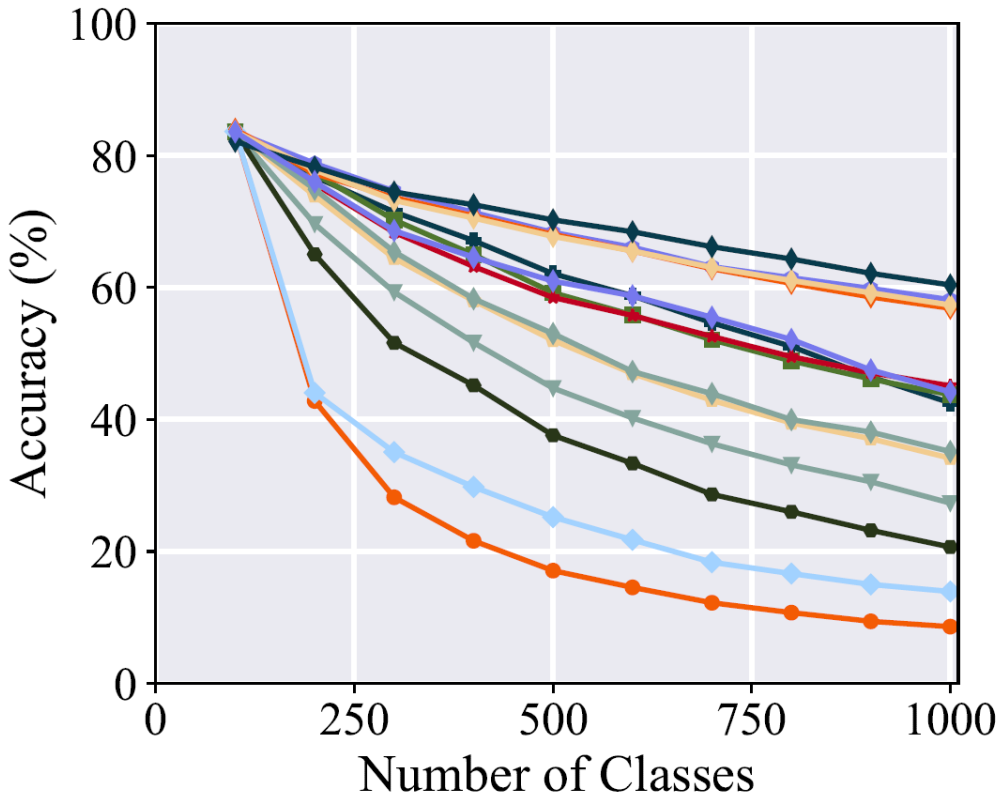
Un danger: l'oubli « catastrophique »

Trois types de stratégies

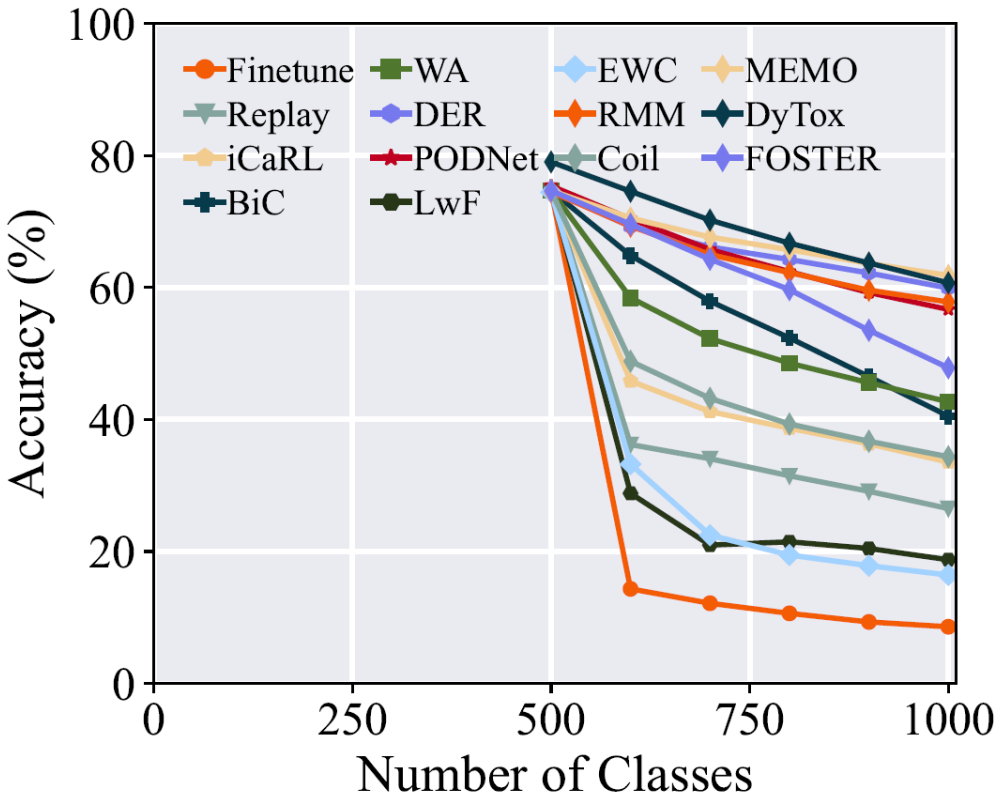
- Régularisation: on fait en sorte de ne pas trop modifier les paramètres précédents
- Mémoire: on garde quelques données passées (replay) ou un utilise un générateur
- Architecture: on modifie la structure des réseaux pour ajouter de nouvelles capacités

Zhou, D.-W., Wang, Q.-W., Qi, Z.-H., Ye, H.-J., Zhan, D.-C., & Liu, Z. (2023). Deep Class-Incremental Learning : A Survey (arXiv:2302.03648). arXiv. https://doi.org/10.48550/arXiv.2302.03648

Benchmark sur apprentissage incrémental

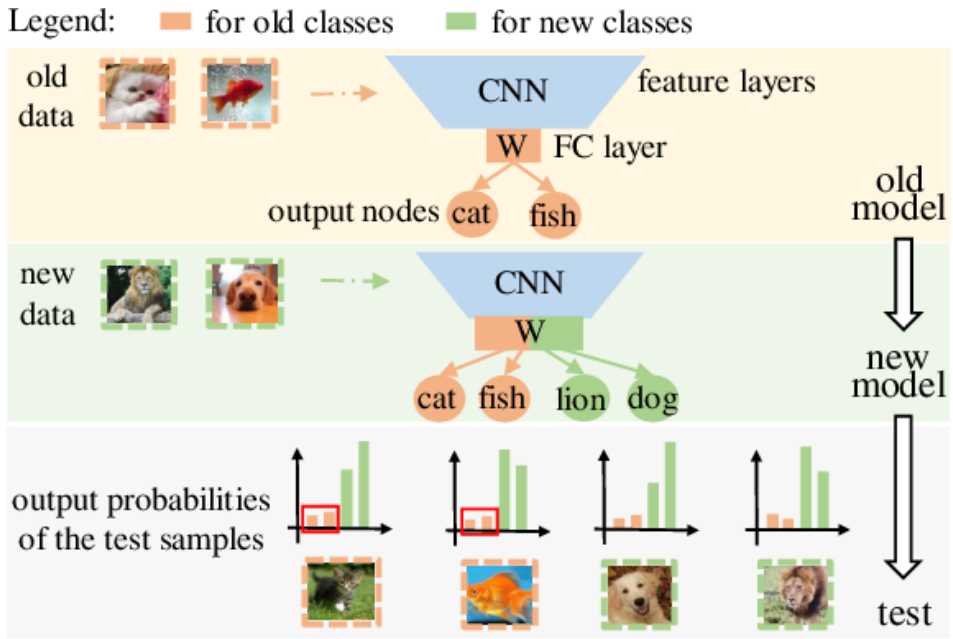


(a) ImageNet1000 Base0 Inc100

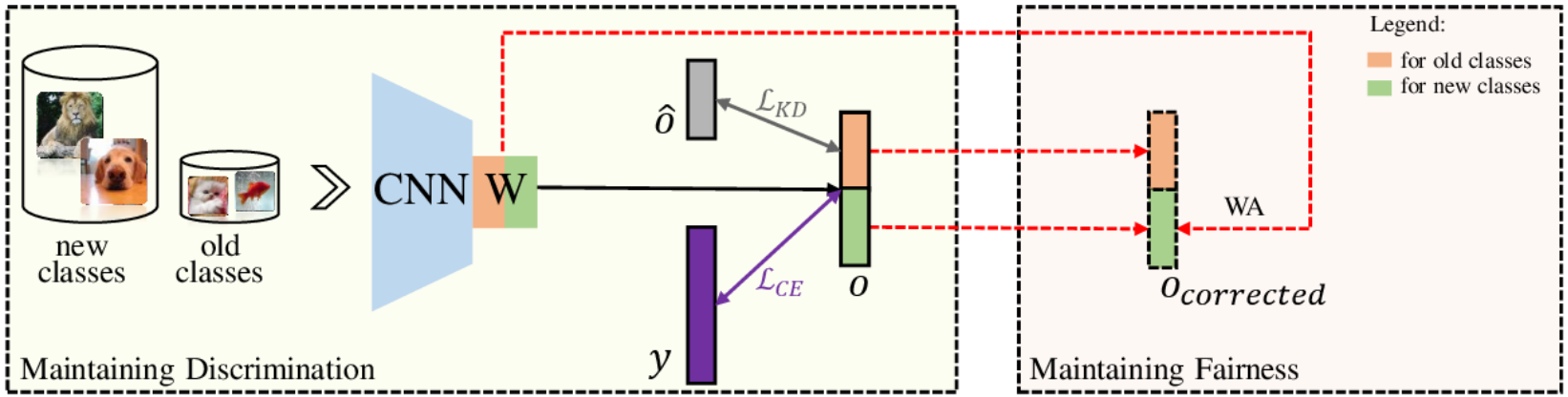


(b) ImageNet1000 Base500 Inc100

Une solution simple

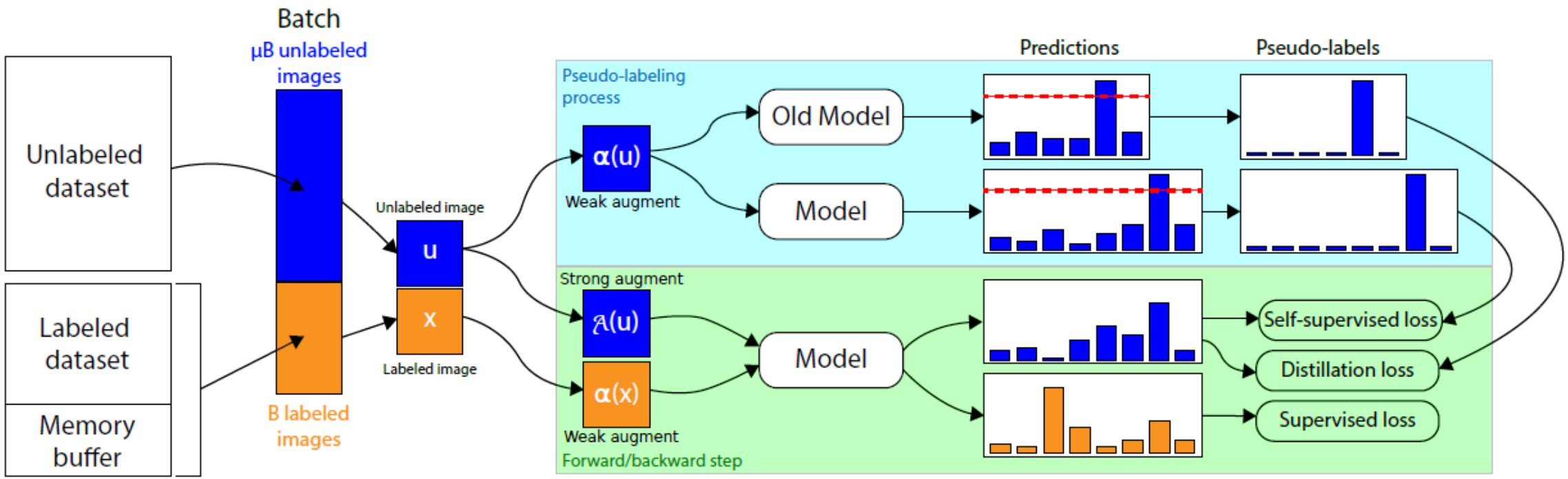


Replay + distillation de connaissance (= régularisation) + re-normalisation a posteriori des poids



Zhao, B., Xiao, X., Gan, G., Zhang, B., & Xia, S. T. (2020). Maintaining discrimination and fairness in class incremental learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13208-13217).

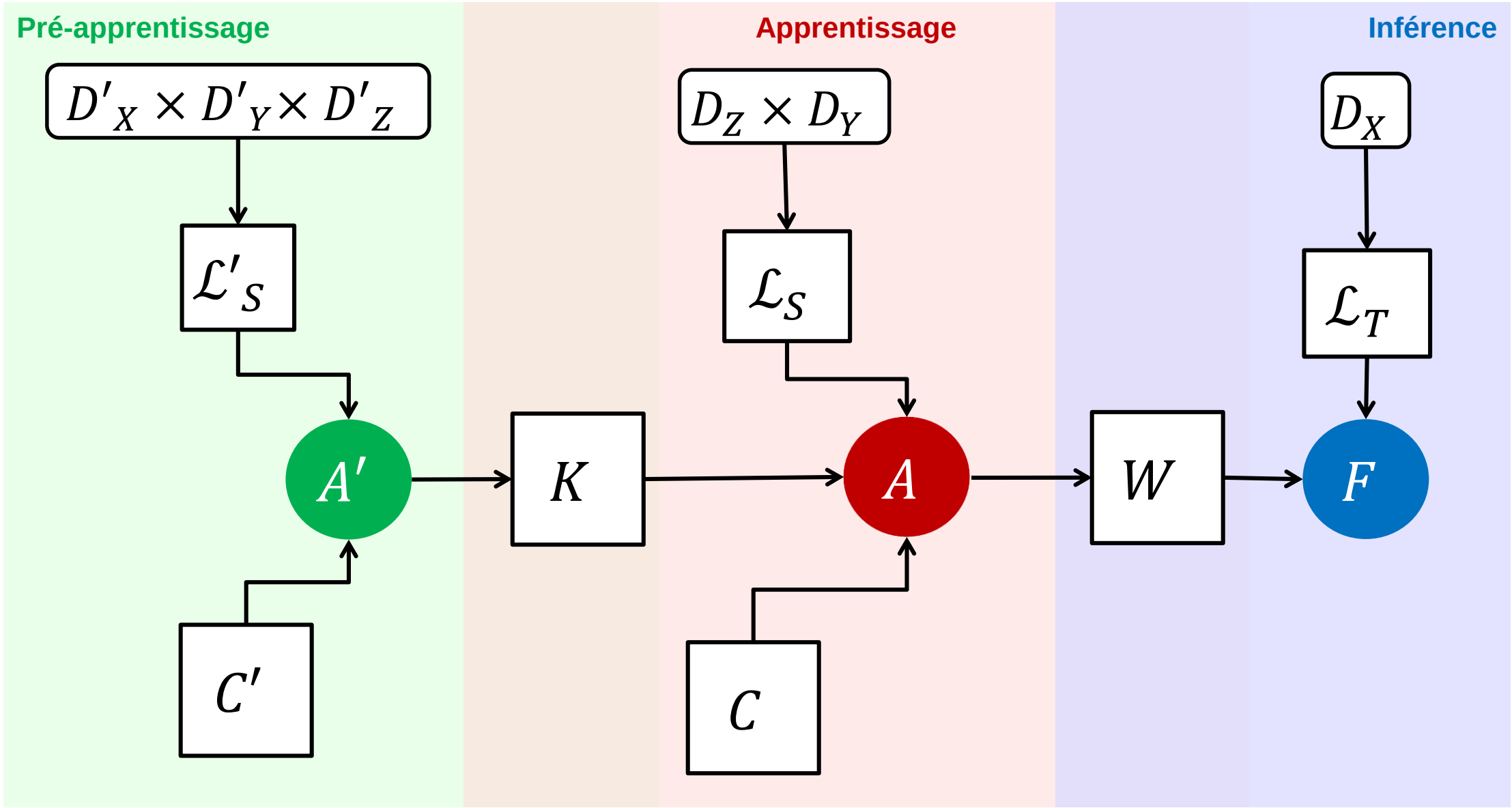
Apprendre au long cours avec un flot de données



Apprentissage incrémental semi-supervisé

- Lechat, A., Herbin, S., & Jurie, F. (2021). Semi-supervised class incremental learning. ICPR.
- Lechat, A., Herbin, S., & Jurie, F. (2021). Pseudo-Labeling for Class Incremental Learning. BMVC.

Zero shot learning / Transfert de modalité



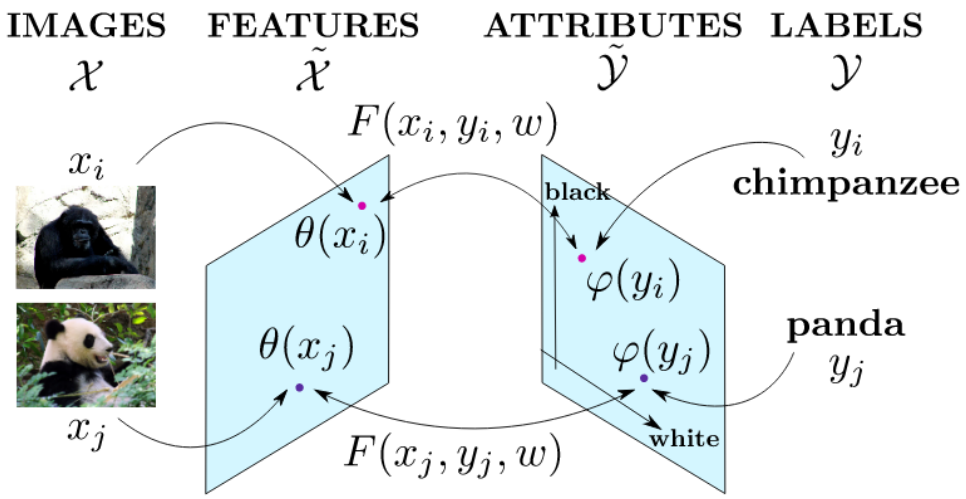
Zero-shot classification

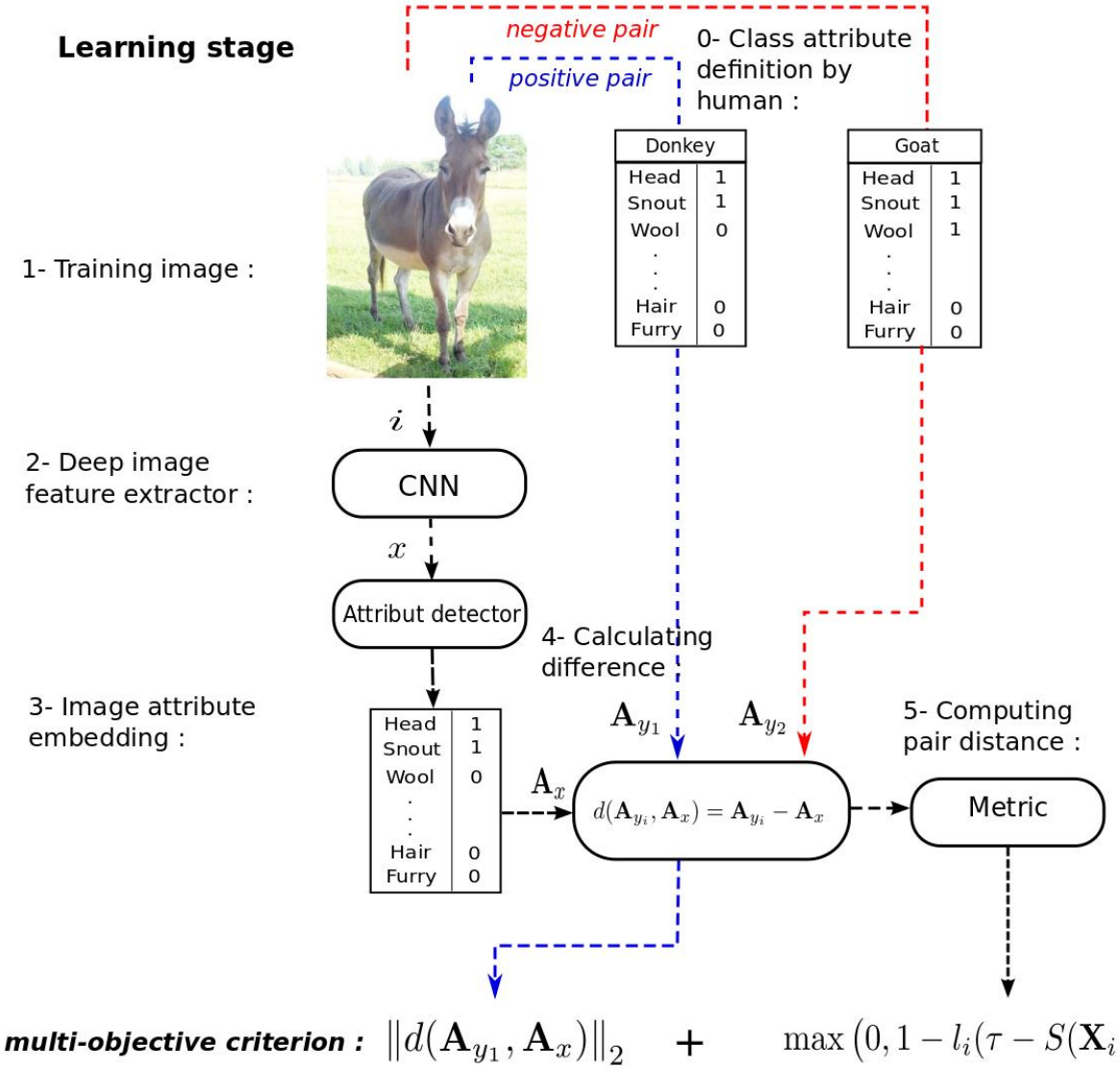
- Apprendre sans référence visuelle !
- Mais description textuelle de la catégorie
  - Attributs
  - Dictionnaire/encyclopédie

= représentation de connaissance

- On dispose d'images X, de descriptions Z des classes Y « vues »
- Pour les classes non vues, on n'a que les descriptions de classes
- Comment construire les associations: $X \leftrightarrow Y \leftrightarrow Z$?

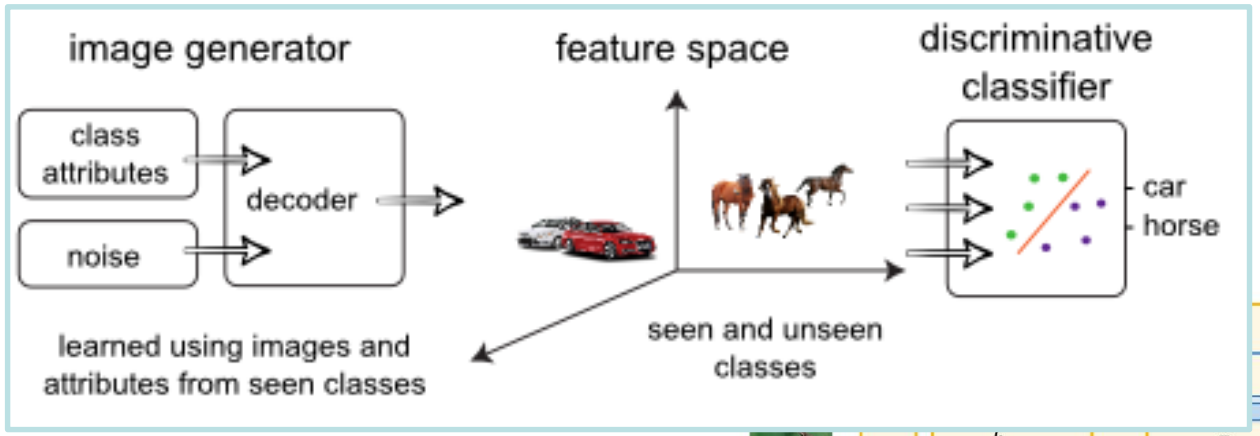
Calculs dans espace de représentation sémantique + fonction de compatibilité

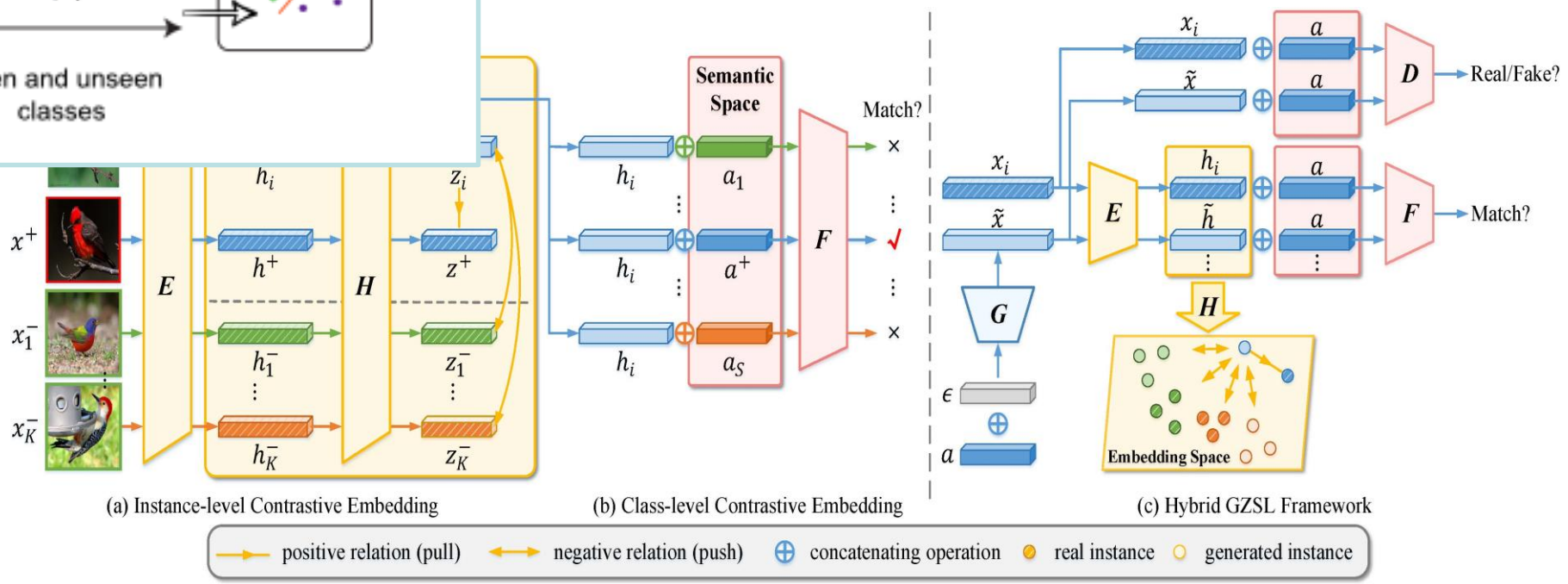




- Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2015). Label embedding for image classification. IEEE transactions on pattern analysis and machine intelligence, 38(7):1425–1438.
- Maxime Bucher, Stéphane Herbin, Frédéric Jurie: Improving Semantic Embedding Consistency by Metric Learning for Zero-Shot Classiffication. ECCV (5) 2016: 730-746

Génération conditionnelles des caractéristiques des classes « non vues »





- Maxime Bucher, Stéphane Herbin, Frédéric Jurie (2017) : Generating Visual Representations for Zero-Shot Classification. ICCV workshop Task-CV (Best Paper)
- Han, Z., Fu, Z., Chen, S., & Yang, J. (2021). Contrastive embedding for generalized zero-shot learning. In IEEE/CVF conference on computer vision and pattern recognition.

Vision & Langage - I

Captioning:
a cat staring out the window at a group of birds.

FOIL:
a ~~dog~~ cat staring out the window at a group of birds.

Referring Expression Recognition:
Bird to the left of the feeder.

Visual Question Answering:
Q: How many birds are there?
A: four

NLVR:
Q: Left image has twice as many cats as the right image, and at least two cats are black. **A:** True

Visual Dialog:

A-Bot: Image shows a cat staring out the window at a group of birds.

Q-Bot: How many cats are there ?
A-Bot: 1

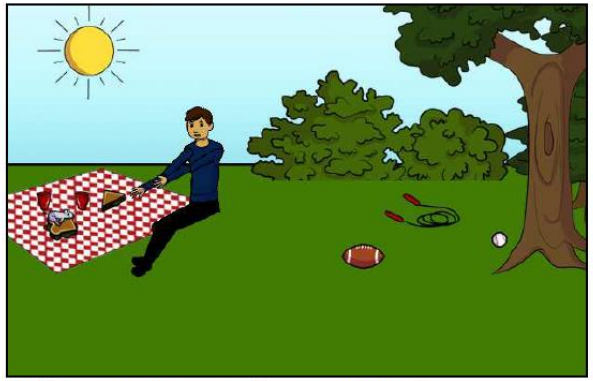
Q-Bot: Can you see its face? [*it = cat; visual coreference*]
A-Bot: no

Q-Bot: I think we were talking about **Image 2**.

Vision & Langage - II

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

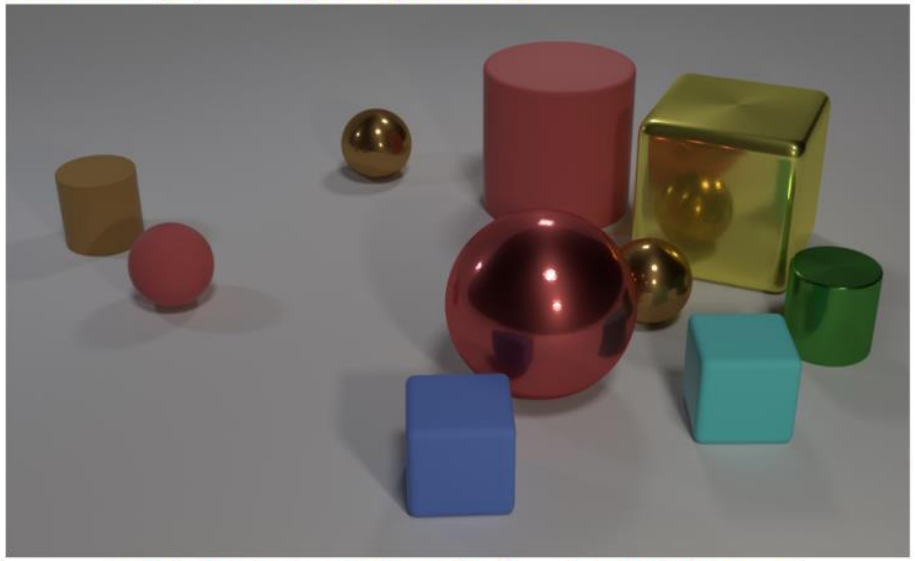


Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Visual Question Answering

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.



Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
Q: **How many** objects are **either** **small cylinders** or **red** things?

VQA + visual reasoning

Antol, S. et al. Vqa: Visual question answering. CVPR 2015.
CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR 2017

Vision & Langage - III

Q: Which American president is associated with the stuffed animal seen here?

A: Teddy Roosevelt

Outside Knowledge

Another lasting, popular legacy of Roosevelt is the stuffed toy bears—teddy bears—named after him following an incident on a hunting trip in Mississippi in 1902.

Developed apparently simultaneously by toymakers ... and named after President Theodore "Teddy" Roosevelt, the teddy bear became an iconic children's toy, celebrated in story, song, and film.

At the same time in the USA, Morris Michtom created the first teddy bear, after being inspired by a drawing of Theodore "Teddy" Roosevelt with a bear cub.

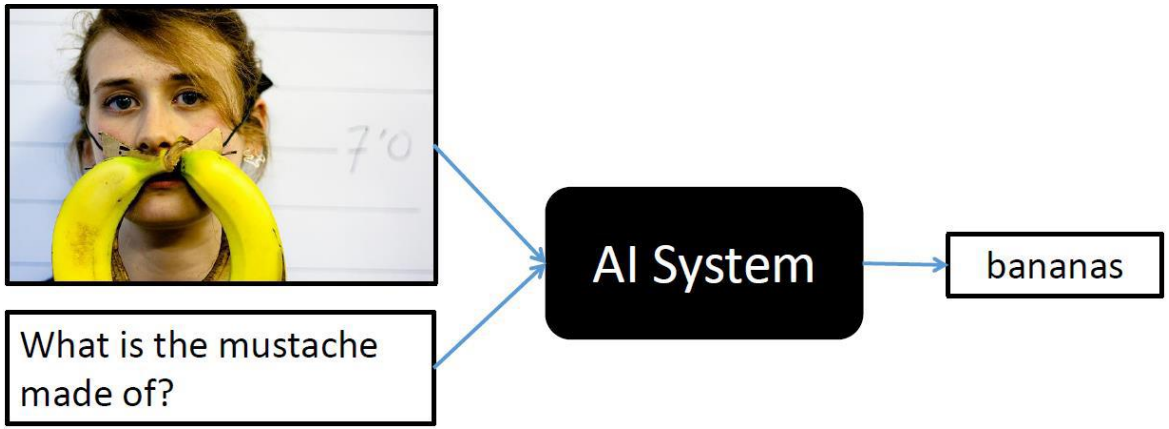
VQA + external knowledge

A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

Visual grounding

OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge, EMNLP 2014
Flickr30K Entities: Collecting Region -to-Phrase Correspondences for Richer Image-to-Sentence Models, IJCV 2017

VQA: tâche paradigmatique



- « Visual Question Answering »
  - Réponse à une question de format libre sur le *contenu visuel* d'une image ou d'une scène
  - Un « Visual Turing Test »
- La tâche
  - Comment est-elle définie?
  - Comment est-elle résolue?

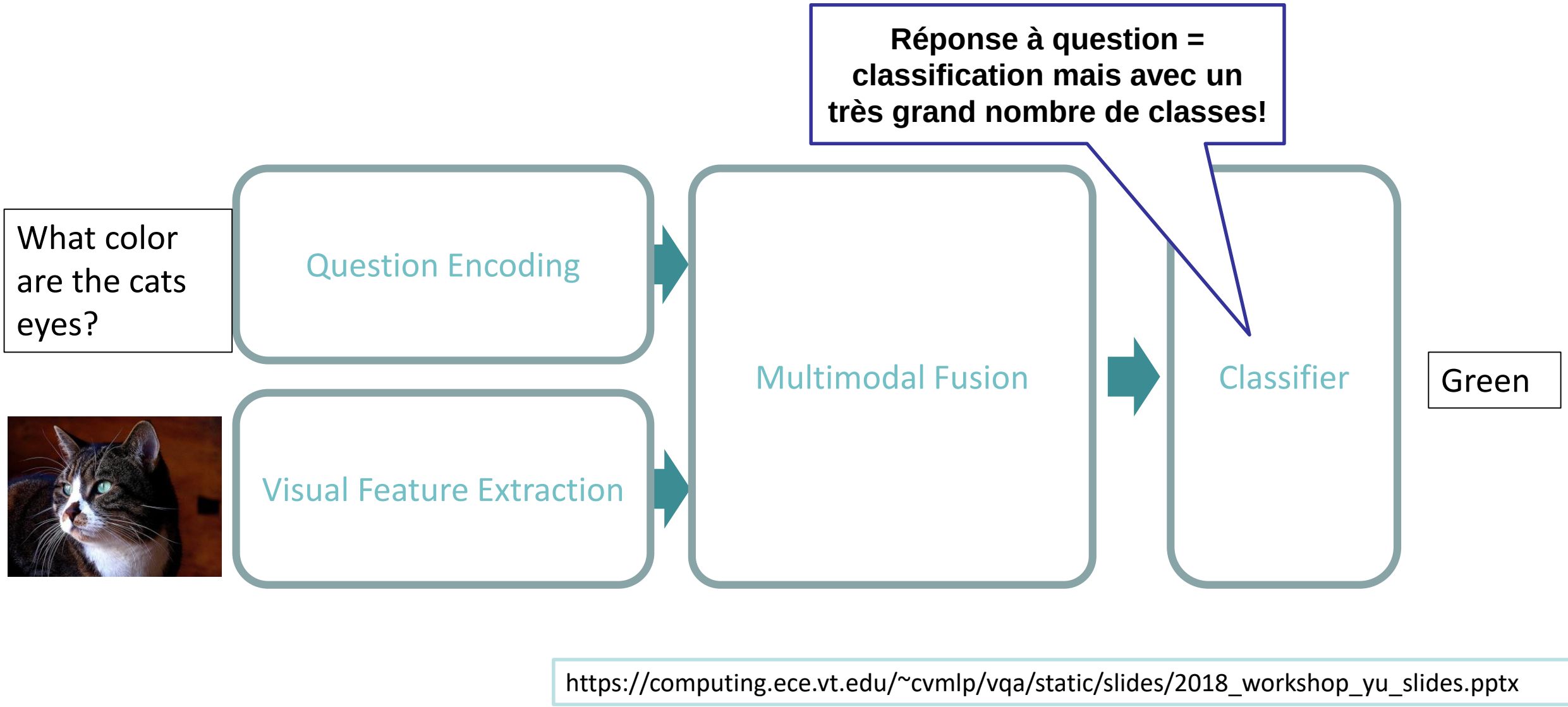
Définition de la tâche = Dataset

- 265,016 images (COCO & “abstract scenes”)
- > 3 questions (5.4 questions en moyenne) par image
- 10 réponses par question (« crowdsourcing »)
- 3 réponses plausibles par question (mais fausses)
- Une même question pour deux images différentes et deux réponses différentes

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

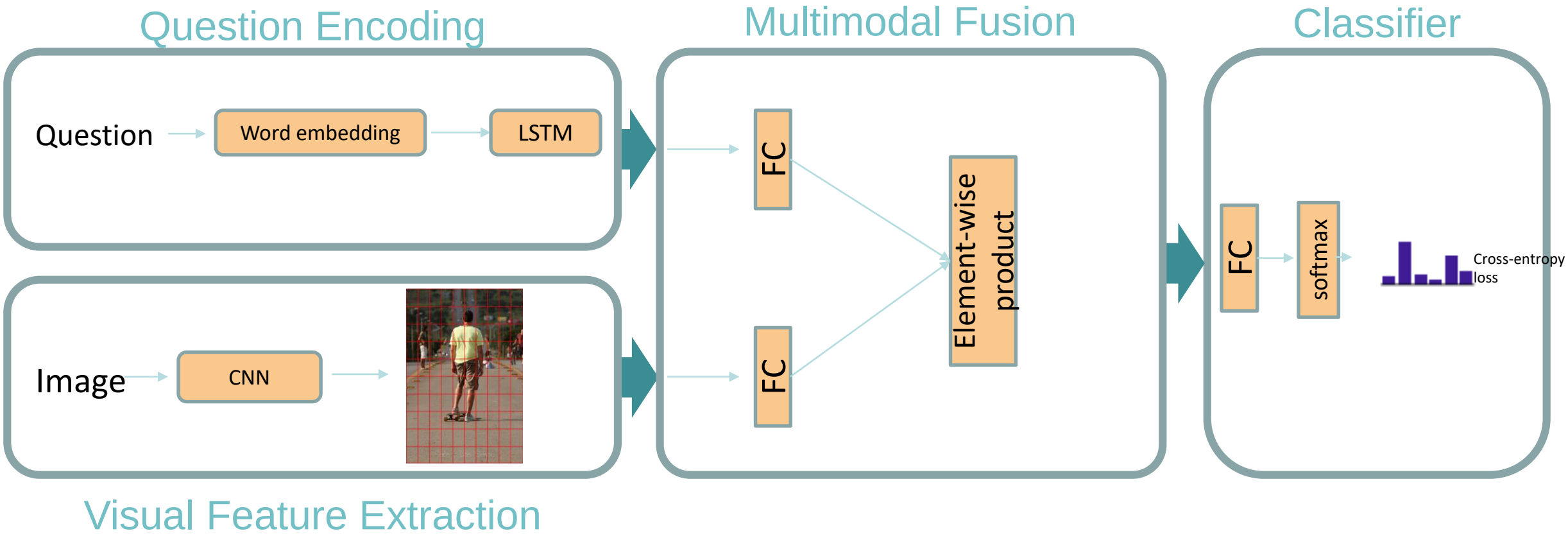
Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, CVPR 2017

Architecture VQA standard



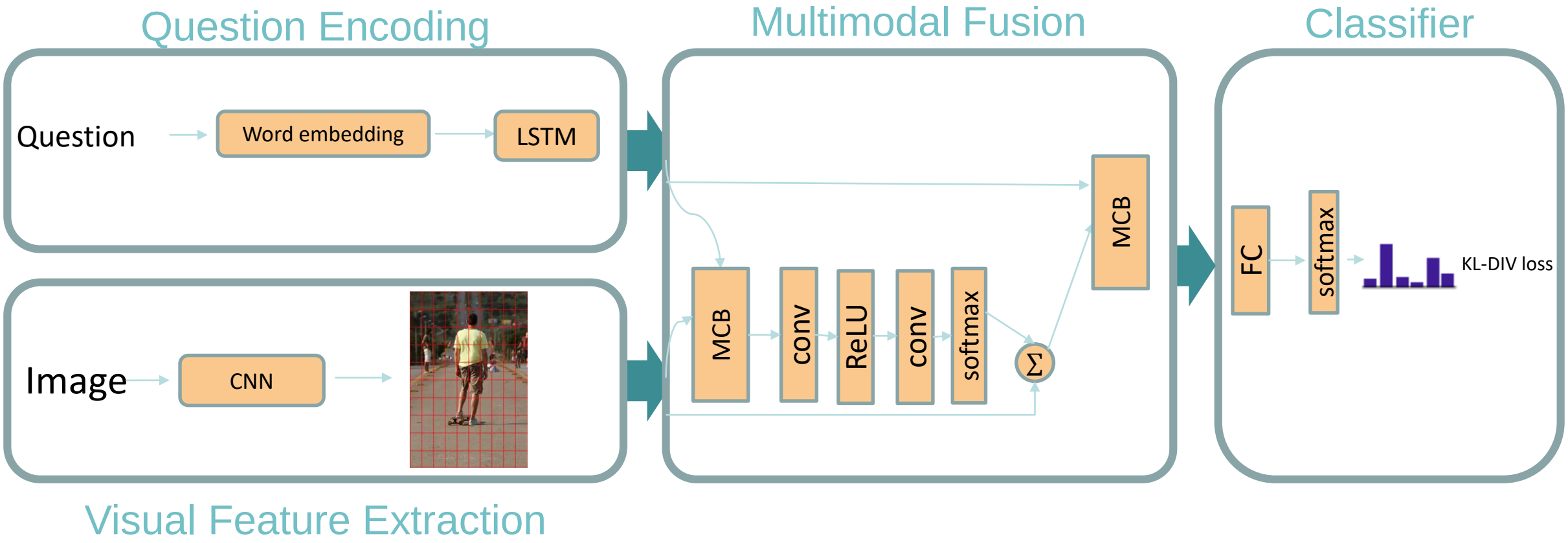
https://computing.ece.vt.edu/~cvmlp/vqa/static/slides/2018_workshop_yu_slides.pptx

VQA Baseline Architecture: CNN + LSTM



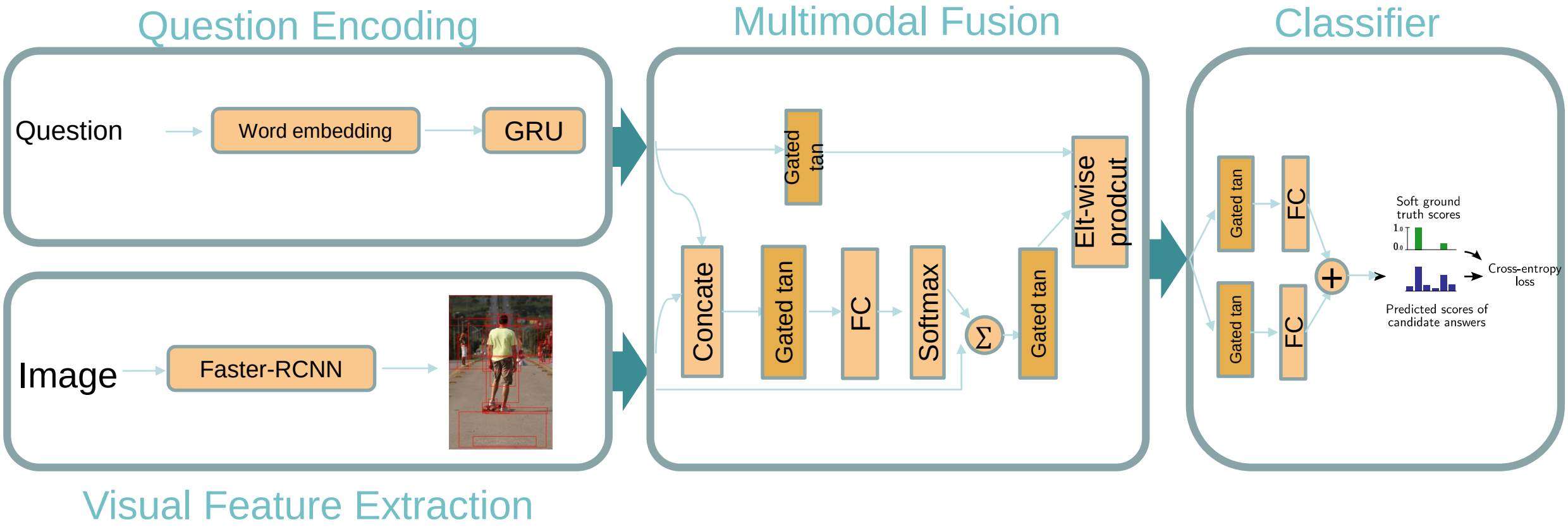
Agrawal et al. 2016

Multimodal Compact Bilinear Pooling



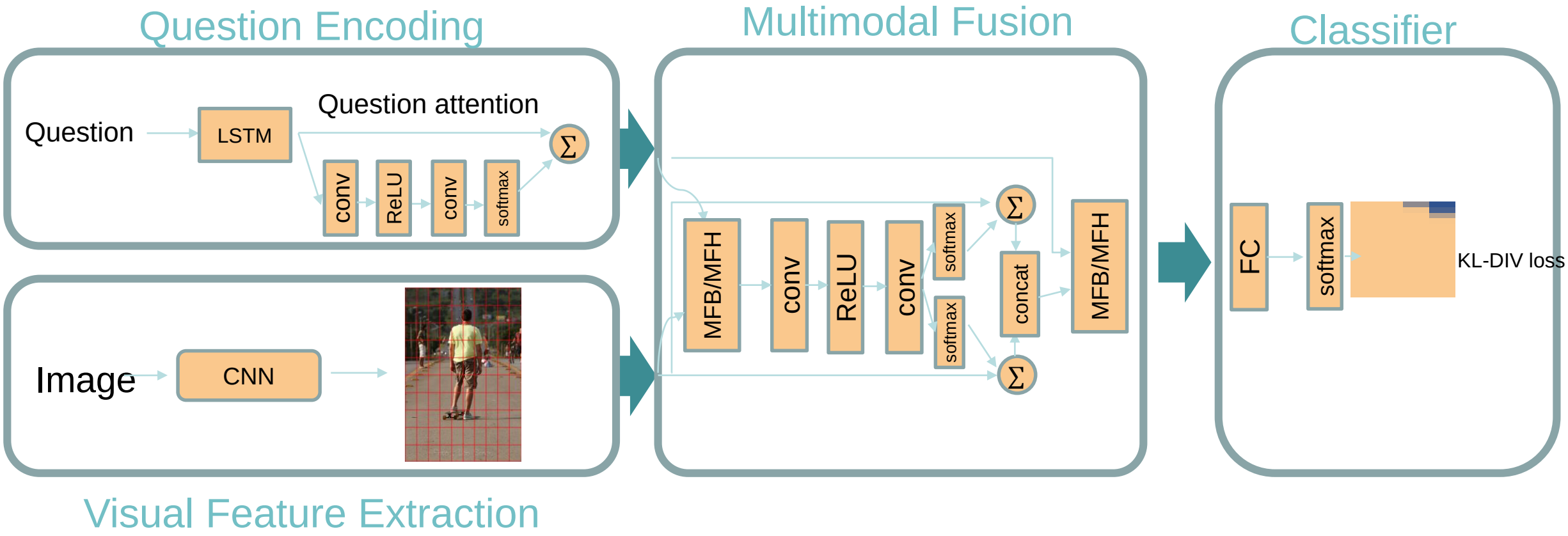
Fukui et al. 2016

Bottom-up and Top-down Attention



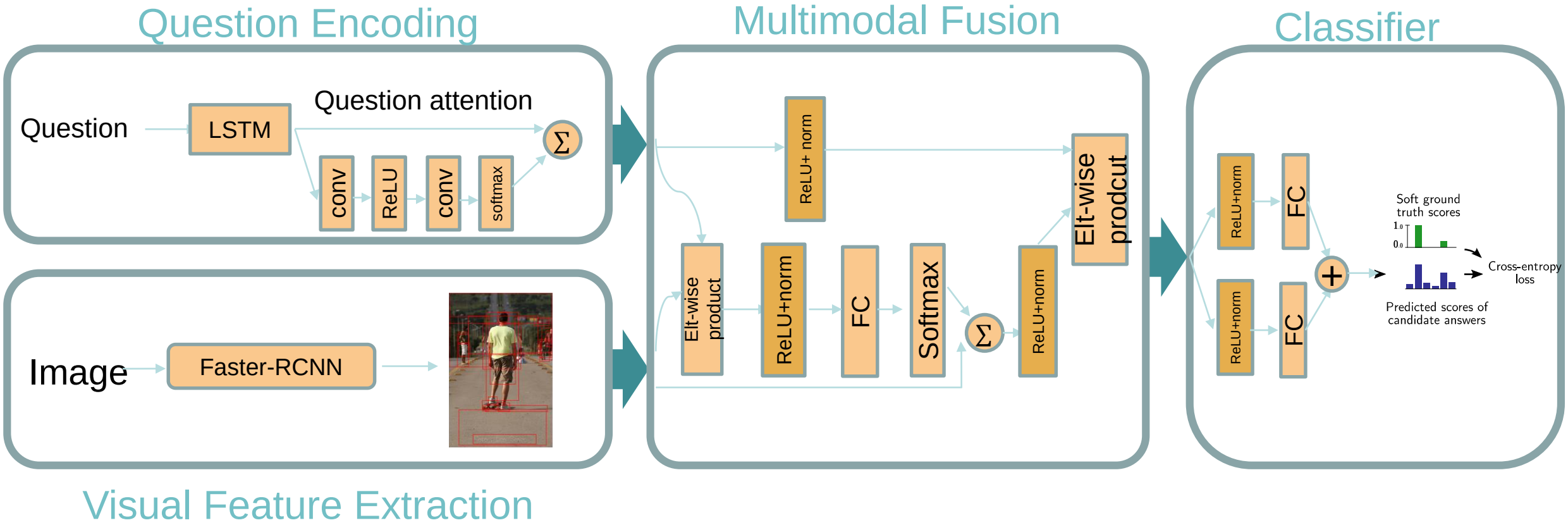
Teney et al. 2017

Multi-modal Factorized Bilinear Pooling with Co-Attention



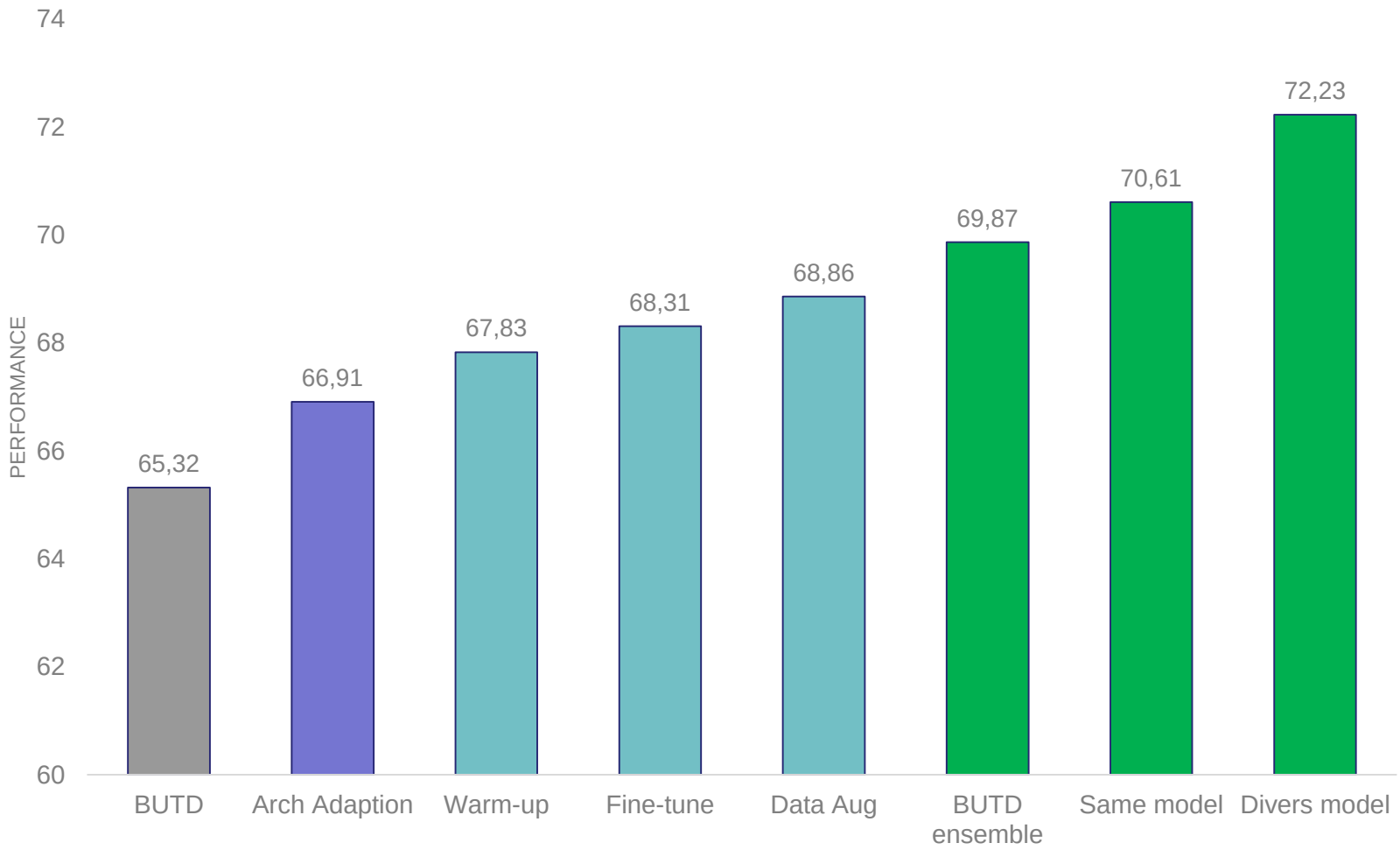
Yu et al 2017

VQA-suite: Architecture Adaptation (winner of 2018 VQA challenge)



https://github.com/hengyuan-hu/bottom-up-attention-vqa

Impact des activités sur les performances



VQA Challenge:

- test-dev : 72.12
- test-standard : 72.25
- test-challenge: 72.41

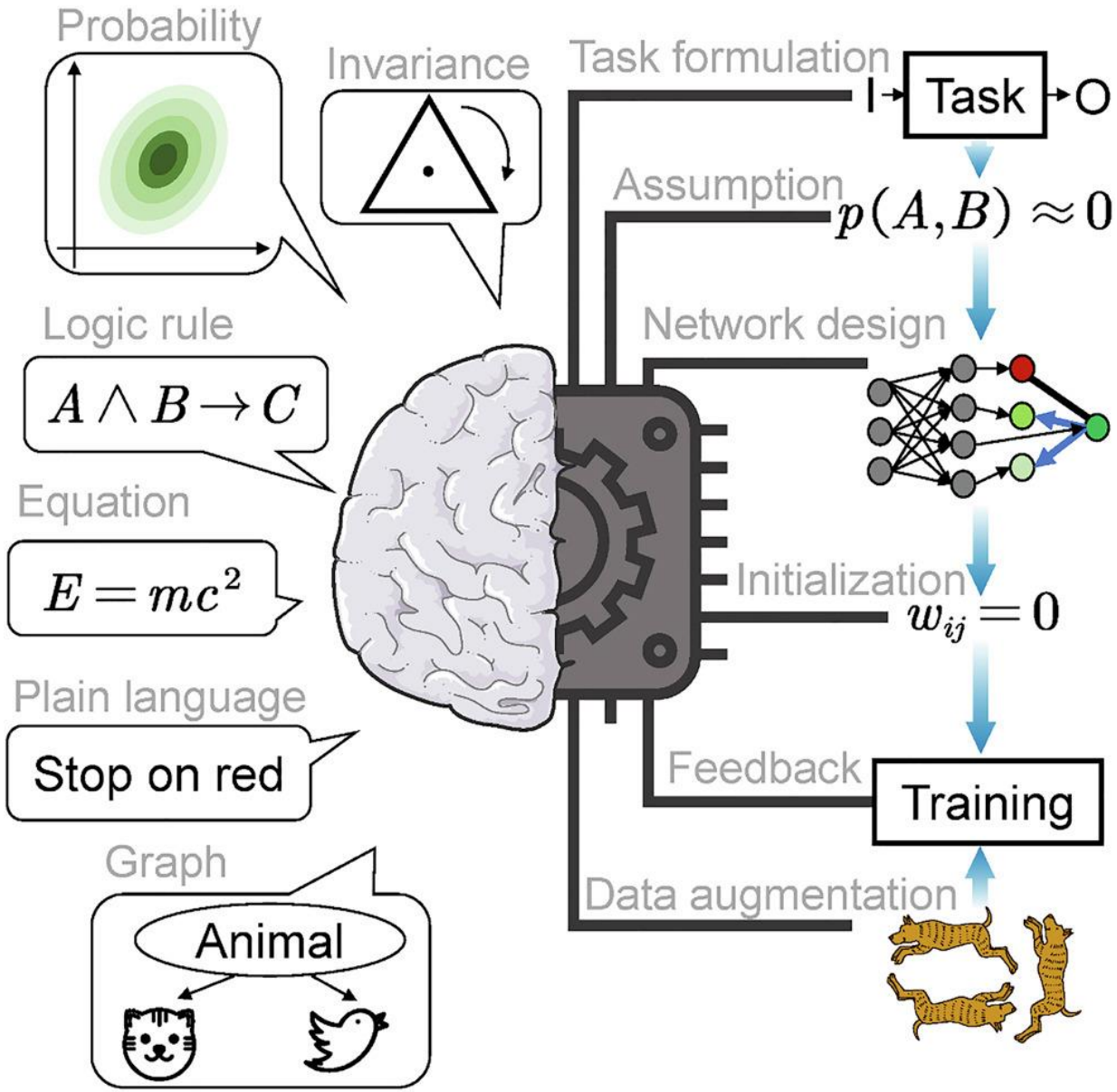
ONERA
THE FRENCH AEROSPACE LAB

Introduire du raisonnement?

- Base d'apprentissage VQA v2 est « biaisée »:
  - On peut construire un prédicteur aveugle correct pour
    - 67% questions oui/non
    - 27.4% questions ouvertes

→ Questions demandant une compréhension structurée des données
  - Raisonnement spatial
  - Raisonnement relationnel (composition)
  - Comparaisons

IA hybride



Deng, C., Ji, X., Rainey, C., Zhang, J., & Lu, W. (2020). Integrating Machine Learning with Human Knowledge. Iscience, 101656.

Deux univers compatibles?

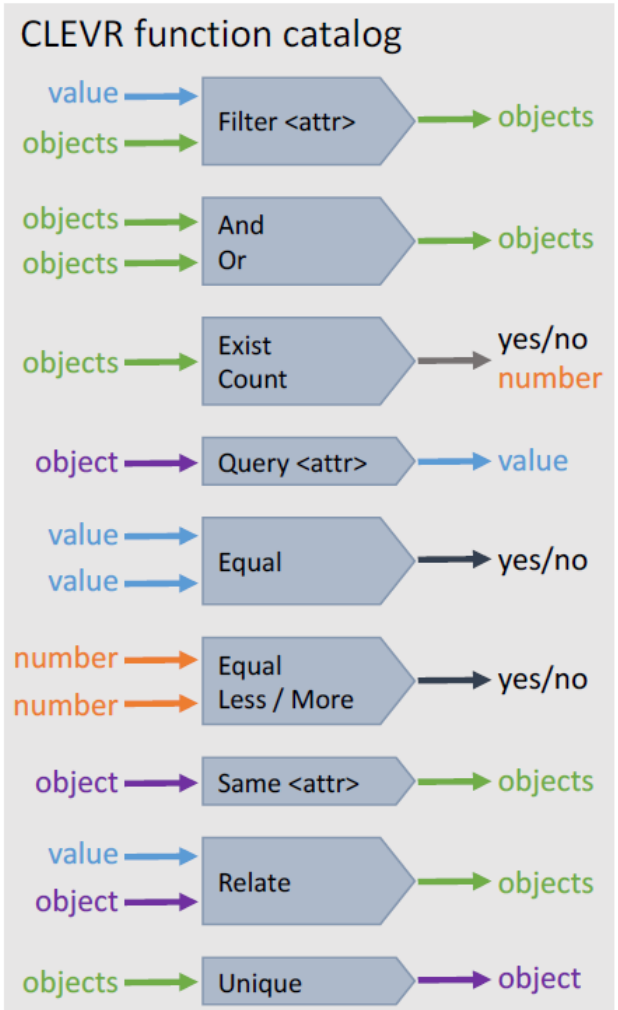
- Intuitif
- Réactif
- Distribué
- Iconique
- Statistique
- « Data driven »
- Holistique
- Opaque
- Empirique
- McCullough et Pitts
- « Deep Learning »
- System 1 (pensée rapide)
- Bottom up
- CVPR
- Vision

IA connexionniste

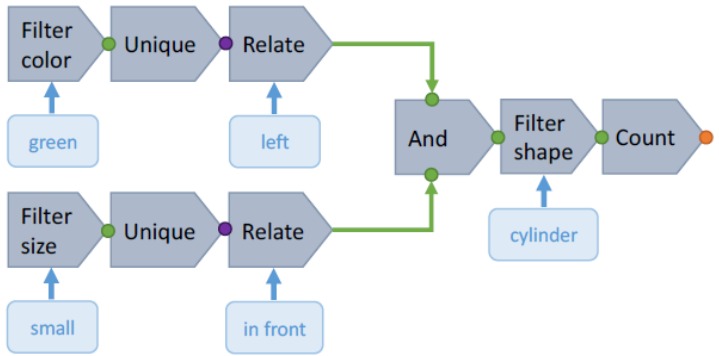
- Conceptuel
- Réfléchi
- Localisé
- Symbolique
- Logique
- « Model-based »
- Composé
- Explicite
- Rationaliste
- Turing
- GOFAI (« rule based »)
- System 2 (pensée lente)
- Top down
- AAAI
- Langage

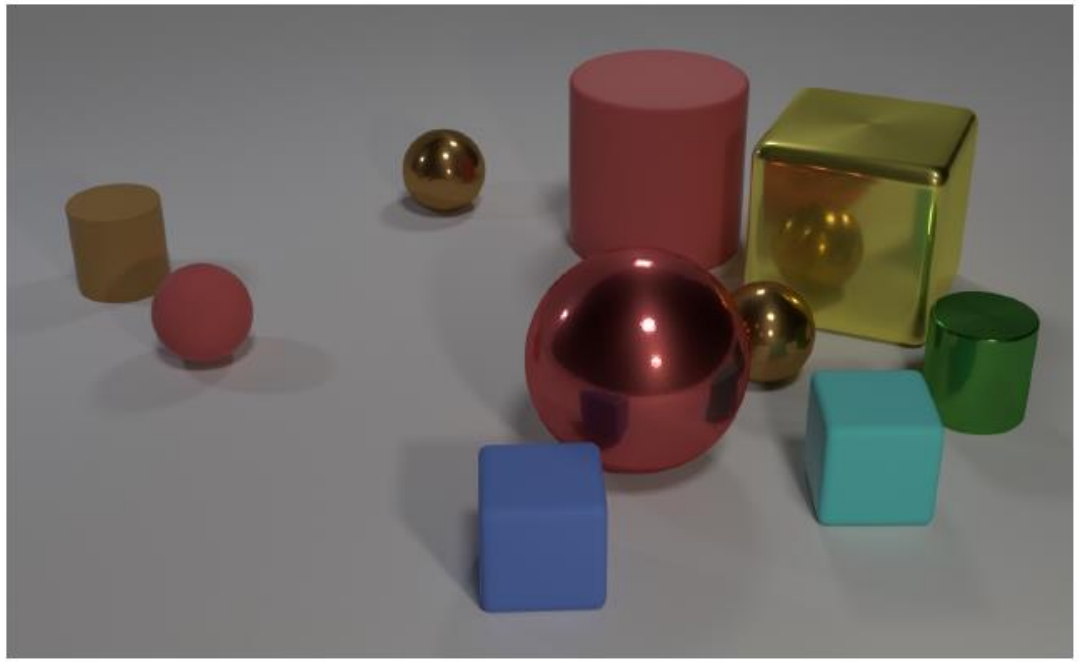
IA symbolique

ClevR: une première base pour tester le raisonnement



Sample tree-structured question:





Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

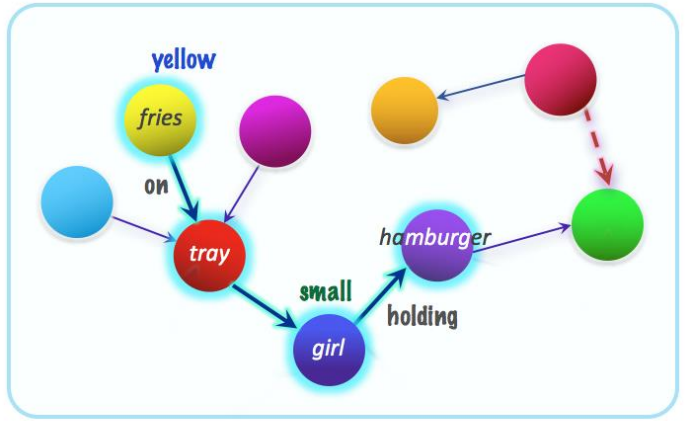
Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR.

GQA: une base plus « photo-réaliste »

Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?
Program: **Select:** <dobject> ⟶ **Choose** <type>: <attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger ⟶ **Relate:** girl,holding ⟶ **Filter** size: small ⟶ **Relate:** object, left ⟶ **Filter** color: red ⟶ **Relate:** food,on ⟶ **Choose** color: yellow | brown



Graph Normalization

- Ontology construction
- Edge Pruning
- Object Augmentation
- Global Properties

Question Generation

- Patterns Collection
- Compositional References
- Decoys Selection
- Probabilistic Generation

Sampling and Balancing

- Distribution Balancing
- Type-Based Sampling
- Deduplication

Entailments Relations

- Functional Programs
- Entailment Relations
- Recursive Reachability

New Metrics

- Consistency
- Validity & Plausibility
- Distribution
- Grounding

Hudson, D. A., & Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR.

GQA: des questions sur les relations entre objets

VQA

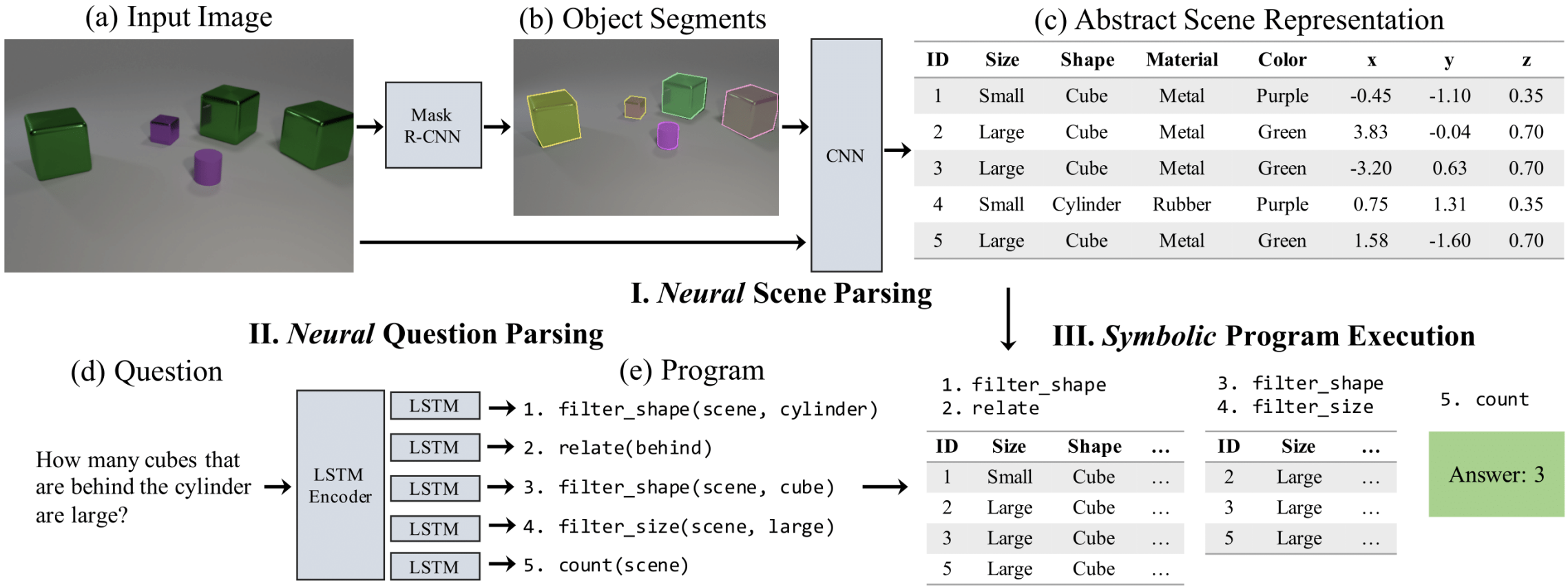
1. Does this man need a haircut?
2. What color is the guy's tie?
3. What is different about the man's suit that shows this is for a special occasion?

GQA

1. Is the person's hair long and brown?
2. What appliance is to the left of the man?
3. Who is in front of the refrigerator on the left?
4. Is there a necktie in the picture that is not red?
5. Is the color of the vest different than shirt?

Hudson, D. A., & Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR.

Approche Neuro-symbolique: du vrai hybride?



ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

ID	Size	Shape	…
1	Small	Cube	…
2	Large	Cube	…
3	Large	Cube	…
5	Large	Cube	…

ID	Size	…
2	Large	…
3	Large	…
5	Large	…

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., and Tenen-baum, J. B. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. arXiv:1810.02338, 2018

« Neural State Machine »

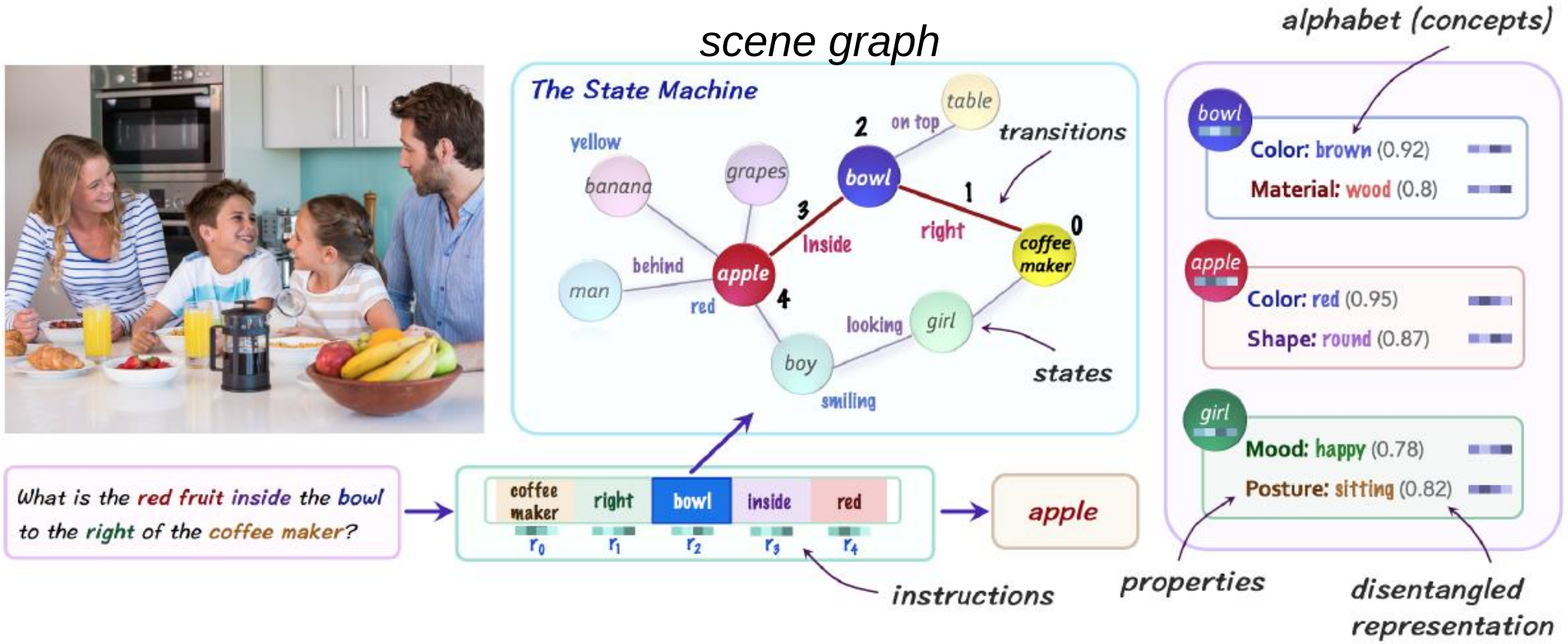


Figure 1: The Neural State Machine is a graph network that simulates the computation of an automaton. For the task of VQA, the model constructs a probabilistic scene graph to capture the semantics of a given image, which it then treats as a state machine, traversing its states as guided by the question to perform sequential reasoning.

[Hudson, 2019]

Hudson, D., & Manning, C. D. (2019). Learning by Abstraction: The Neural State Machine. Advances in Neural Information Processing Systems, 32, 5903-5916.

Algorithmes de VQA

Architectures « embedding »: Structure + sémantique → Signal

- Deux flux d'encodage séparés de chaque modalité + fusion
  - Texte: « word embedding » dans un espace vectoriel (Glove, Word2Vec) puis intégration (moyenne, LSTM, GRU)
  - Image: caractéristiques profondes de régions d'objets détectées (RCNN) puis concaténation
- Utilisation de mécanismes attentionnels dans la fusion des flux texte/image
- Réponse = classification sur un vocabulaire > 3000 catégories.
- Futur = pré-apprentissage de représentations visuo-textuelles (modèles de fondation).

Architectures « raisonnement »: Signal → Structure + sémantique

- Comment construire une représentation structurée d'une image/scène?
- Comment réaliser des inférences à partir de cette représentation?
- Comment rendre les inférences robustes aux incertitudes et erreurs?
- Que peut-on apprendre vs. modéliser?

MODÈLES DE FONDATION

Modèles de fondation: pourquoi?

- Un seul modèle pour plusieurs tâches
  - Le travail lourd (apprentissage de représentation) fait une seule fois!
  - Il n'y a plus qu'à adapter « localement » pour un contexte ou une tâche donnée.

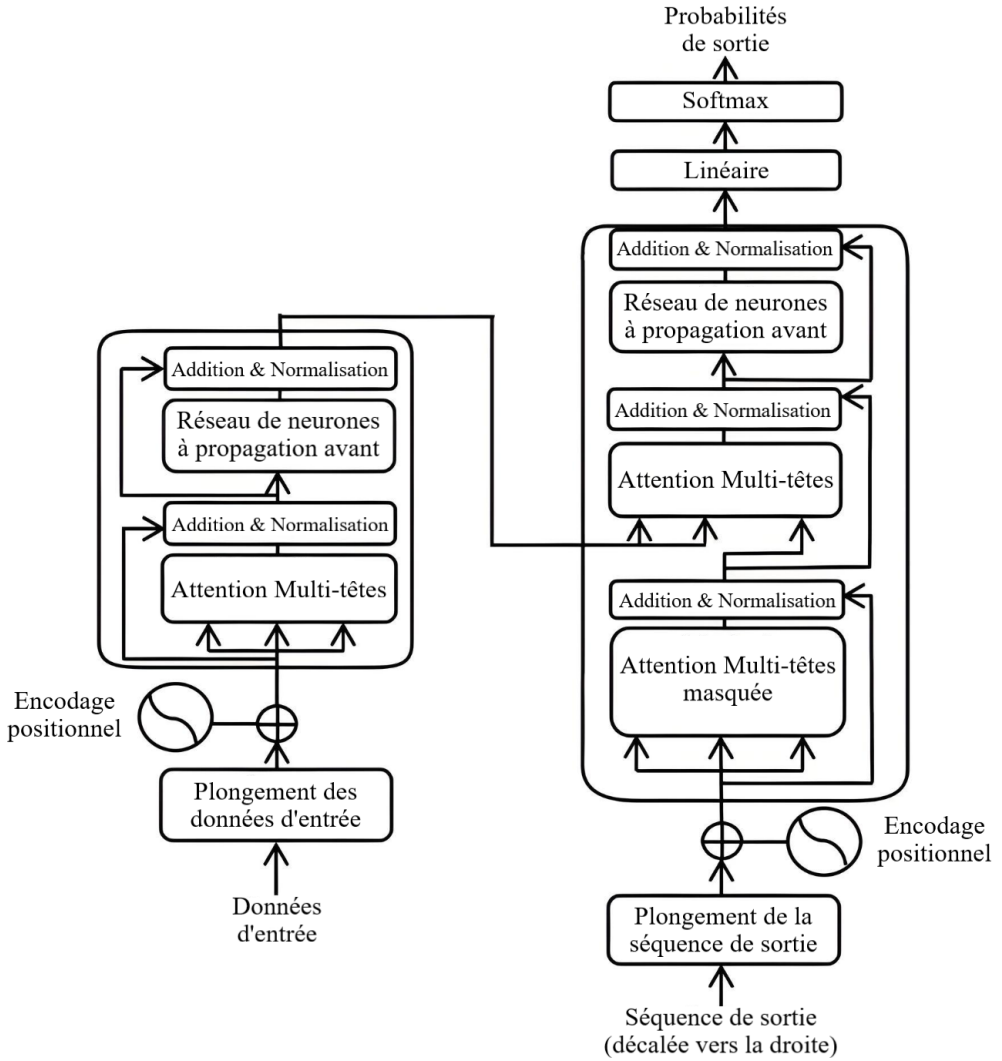
- Comment produire cette universalité?
  - Le langage comme référence, mais aussi d'autres modalités
  - Des données, encore des données…
  - Une architecture commune: les transformer (mais des doutes sur leur efficacité à long terme: Mamba?)

Transformers: « Attention is all you need »

- Architecture encodeur/décodeur
- Utilisé au départ pour de la traduction:

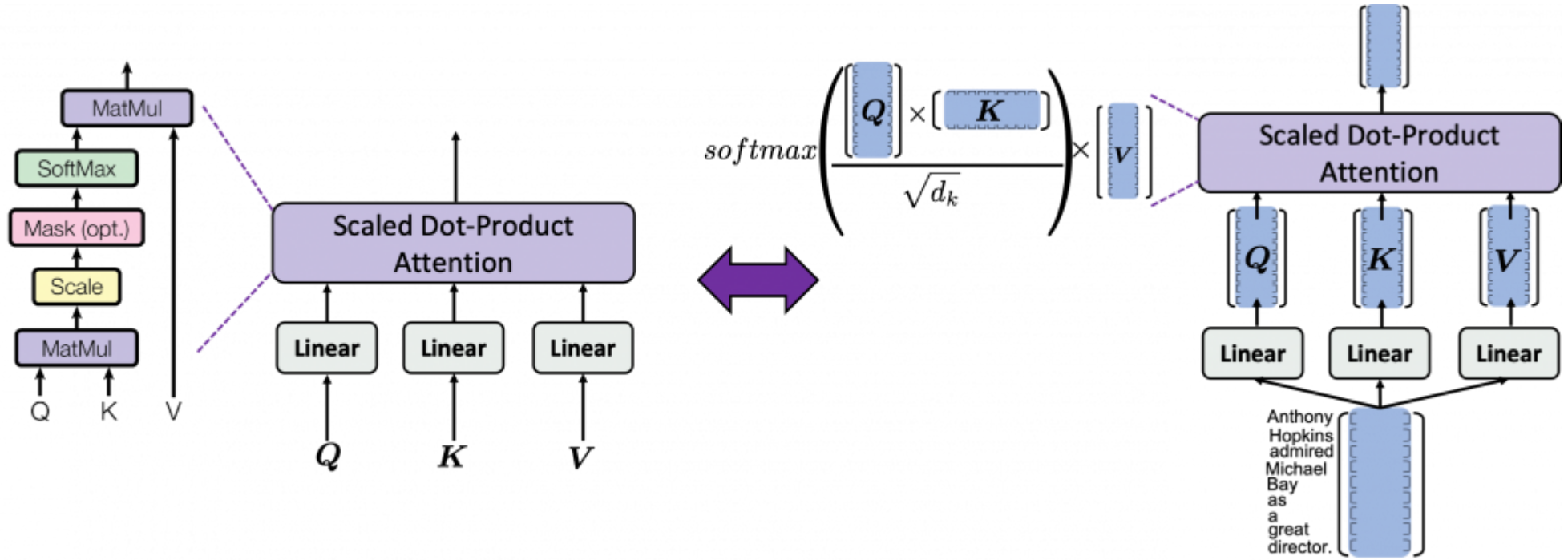
  séquence → séquence
- Justification: réussir à exploiter les interactions longue portée

  Interaction = auto-attention
- Transformer = garde la structure (la séquence) mais transforme les codes.
- Pas de réduction de dimension (= pas de *stride* ou *pooling*)
- Multi channel (convolution) → Multi head
- Position codée par ajout d'un signal (encore un code, mais connu!)



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017.

Query, Key, Value



- Query = le signal à transformer = un ensemble de vecteurs
- Key = l'index ou le vecteur du code
- Value = la transformation associée à chaque code

Query, Key, Value

- Interprétation: décomposition selon un dictionnaire « transformé » de l'entrée Q

$$Y_i = \sum_i < Q_i, K_j > V_j$$

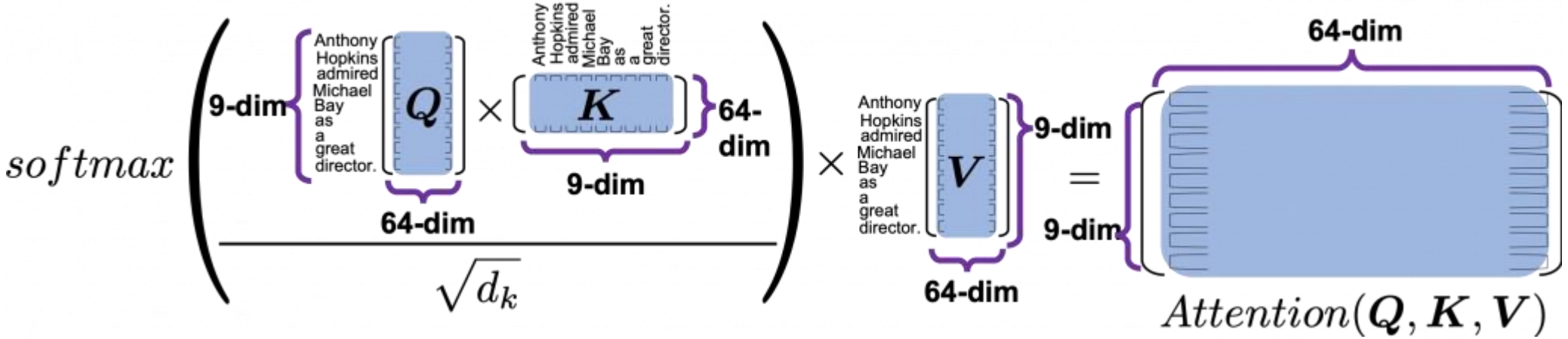
$$Y_i = \sum_i < W_q Q_i, W_k K_j > W_v V_j$$

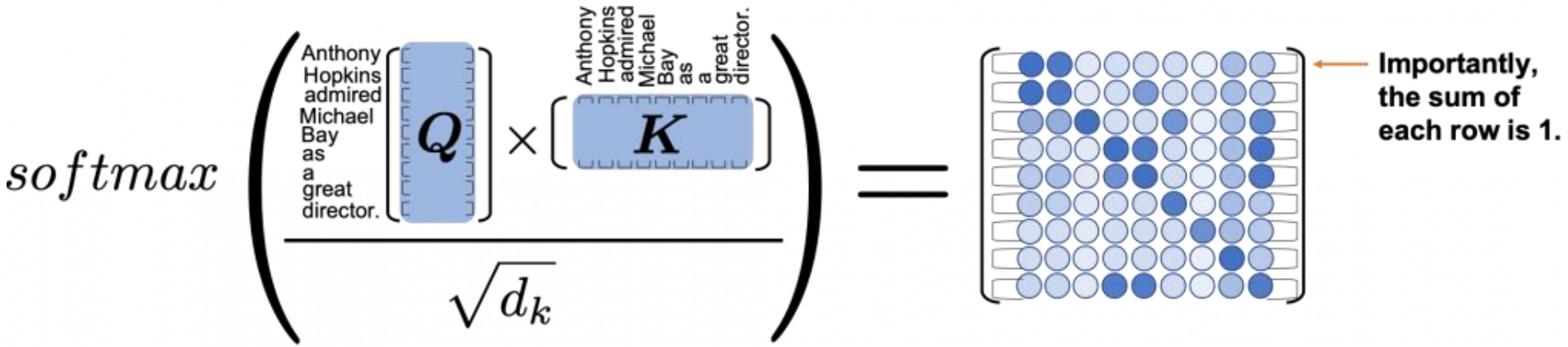
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{W_q Q [W_k K]^T}{\sqrt{d_k}}\right) W_v V$$

- Intérêt: Transforme un signal de taille quelconque en un signal de même taille.
- Adapté au texte si chaque Q_i est un vecteur

- Self-attention: $Q = K = V$, Cross attention:$Q \neq V$

Query, Key, Value





Multi head = multi canaux

1) This is our input sentence*

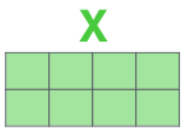
2) We embed each word*

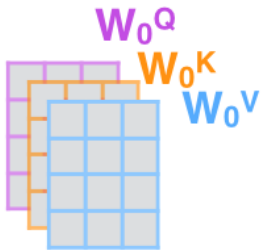
3) Split into 8 heads. We multiply X or R with weight matrices

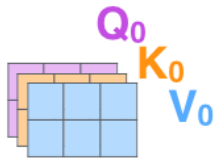
4) Calculate attention using the resulting Q/K/V matrices

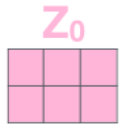
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

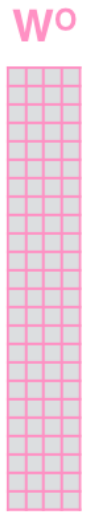
Thinking Machines

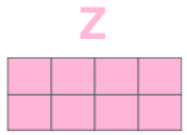


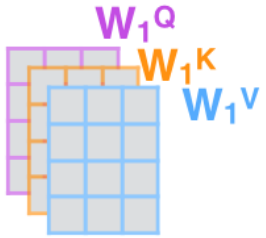


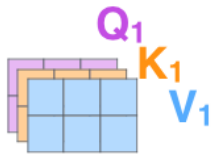


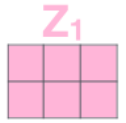




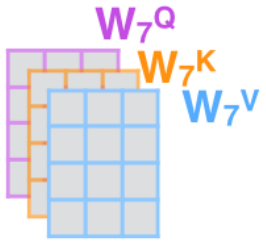


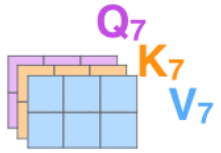


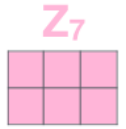




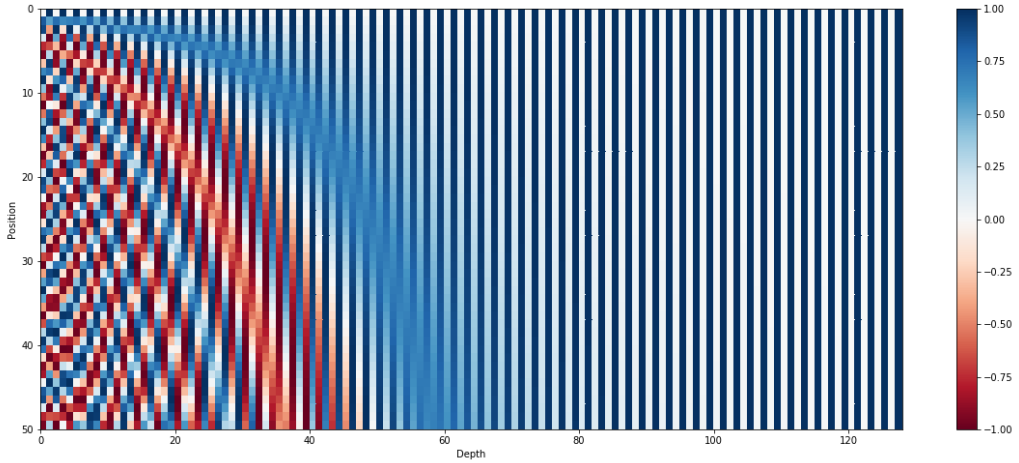
* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one







Positional encoding



- Stratégie pour coder la localisation des éléments d'entrée, leur indice: utile pour représenter les « distances » (relatives ou absolues).
- Astuce: on ajoute un signal sinusoïdal faible qui dépend de la localisation t

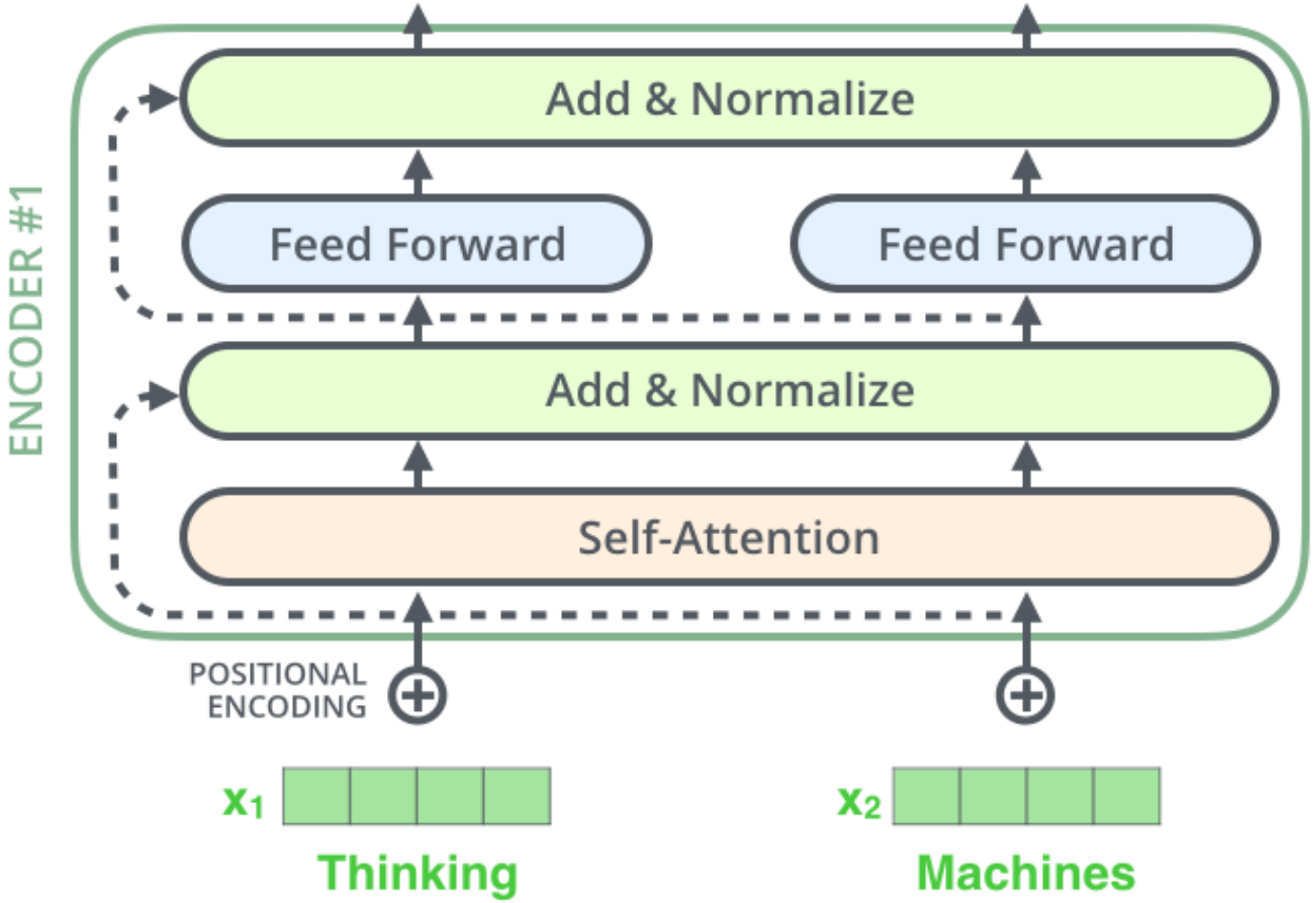
$$p(t) = [\sin(w_1.t), \cos(w_1.t), \sin(w_2.t), \cos(w_2.t), \ldots \sin(w_{d/2}.t), \cos(w_{d/2}.t)]$$

Où

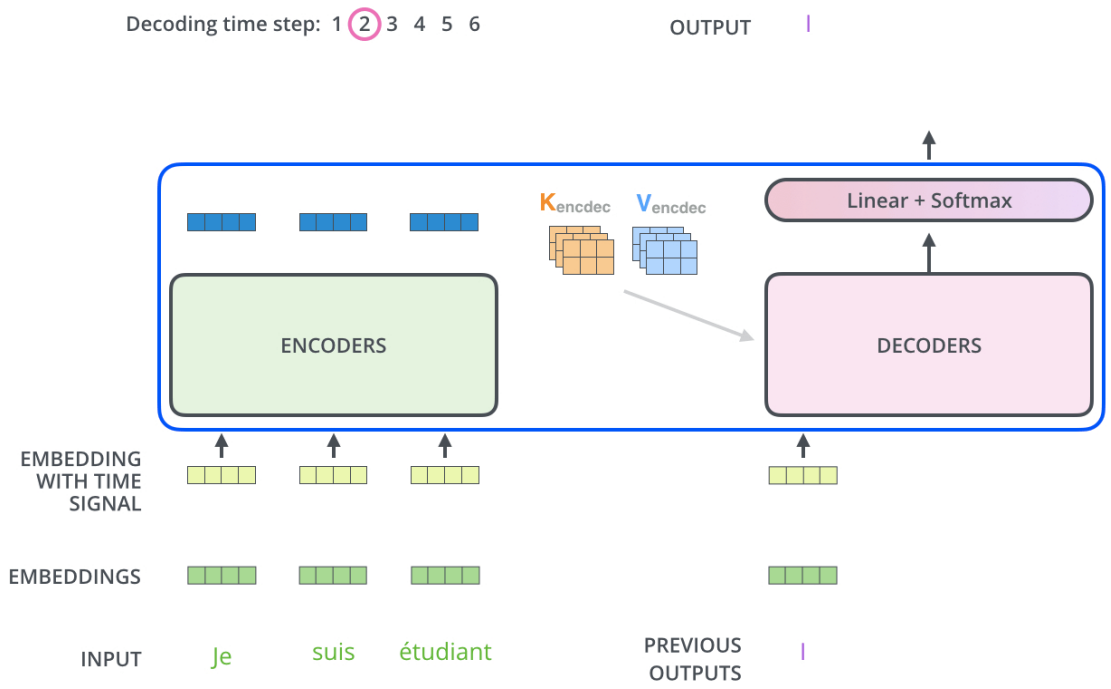
$$w_k = \frac{1}{10000^{2k/d}}$$

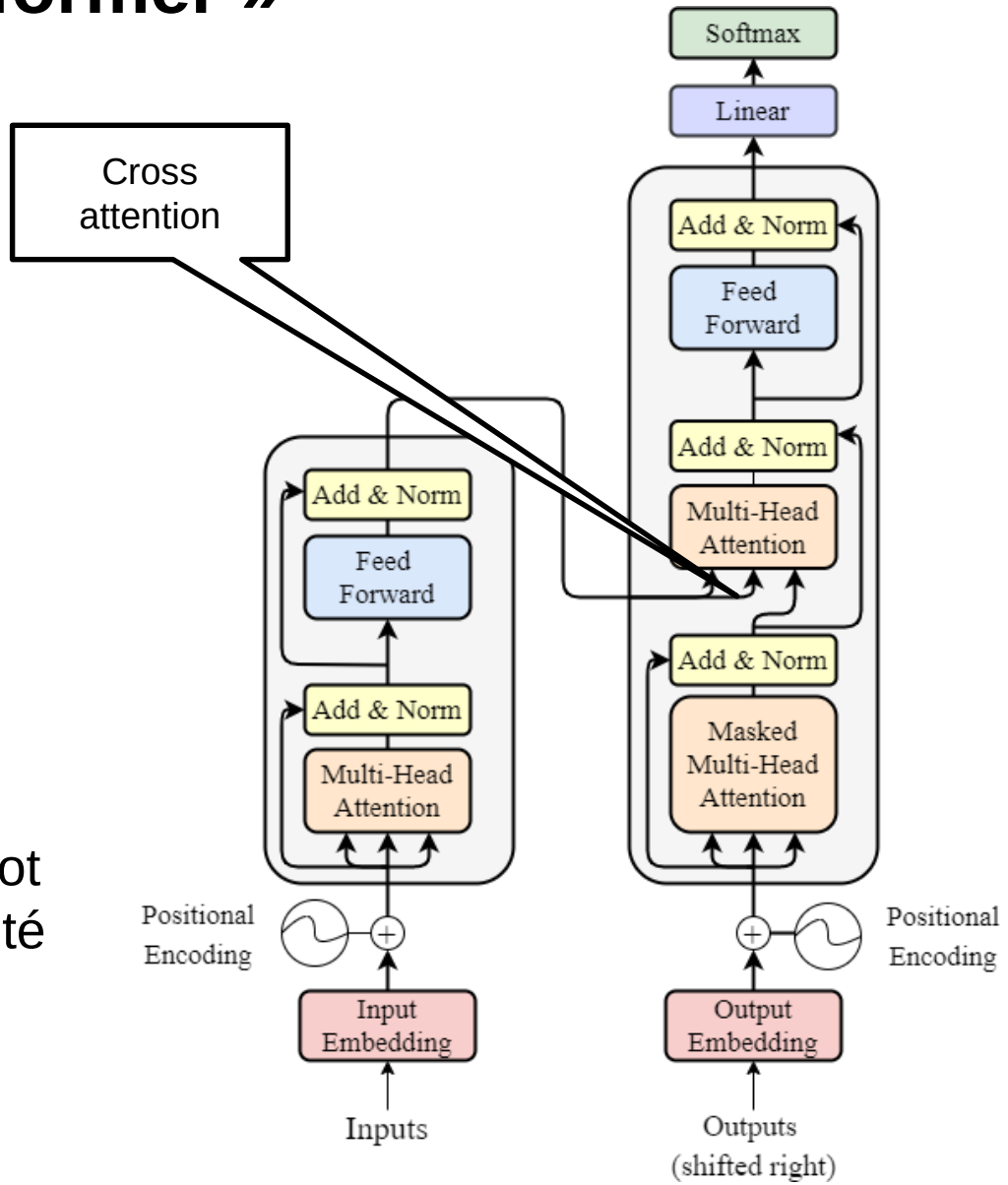
- La différence entre deux codes ne dépend pas de t

La couche « encoder » complète



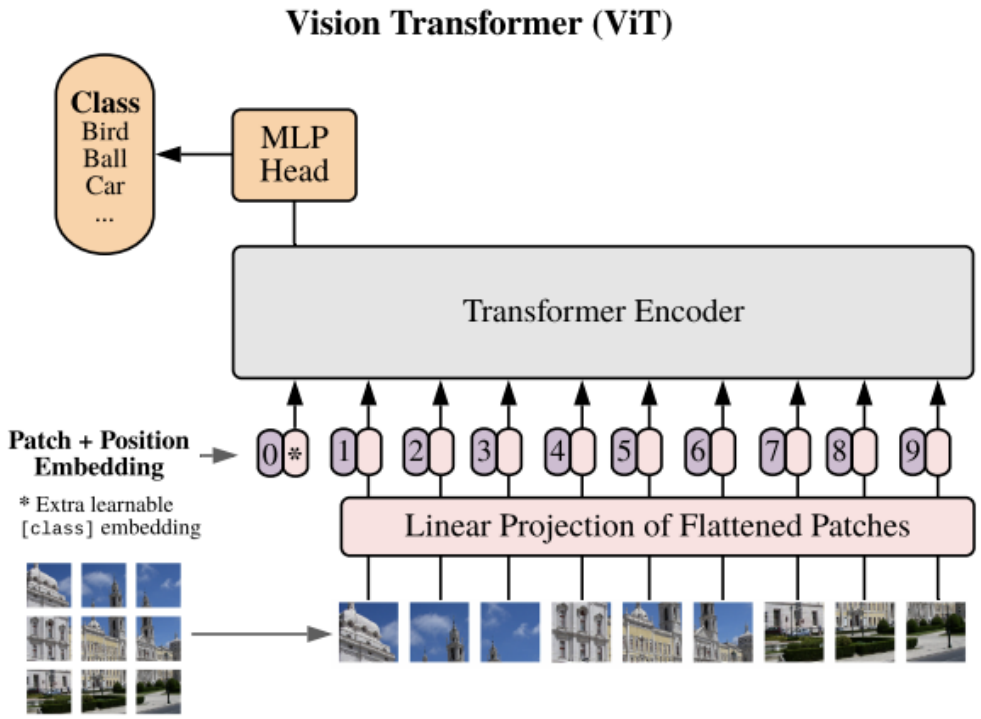
Decodeur « transformer »



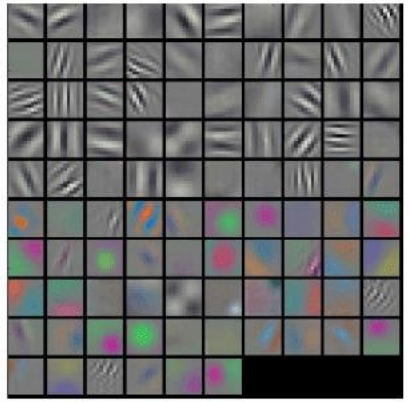


- Modèle auto-régressif: ajoute séquentiellement le mot décodé en entrée du décodeur pour générer la totalité de la séquence
- Etage de « cross attention » pour combiner code et processus de décodage

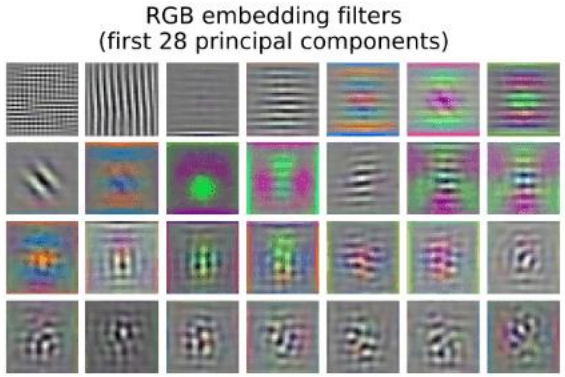
VIT : transformer pour l'image



Alexnet 1st conv filters



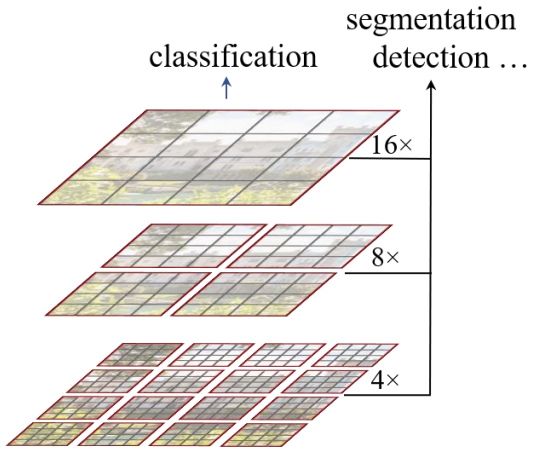
ViT 1st linear embedding filters



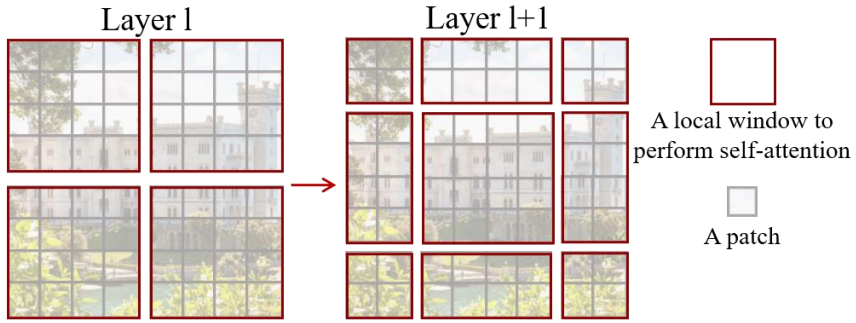
- Décompose une image en liste de patch et l'encode avec un transformeur
- Un « token » fictif encode l'ensemble de l'information

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020, October). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations.

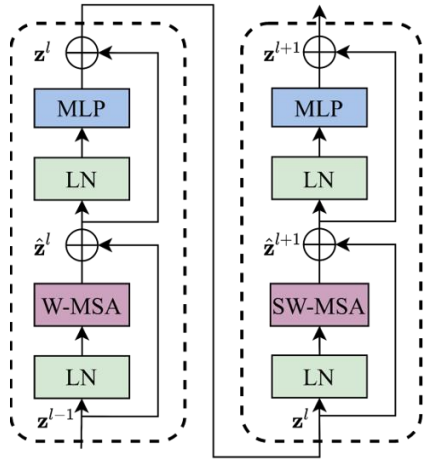
SWIN: un transformer image multi-échelle



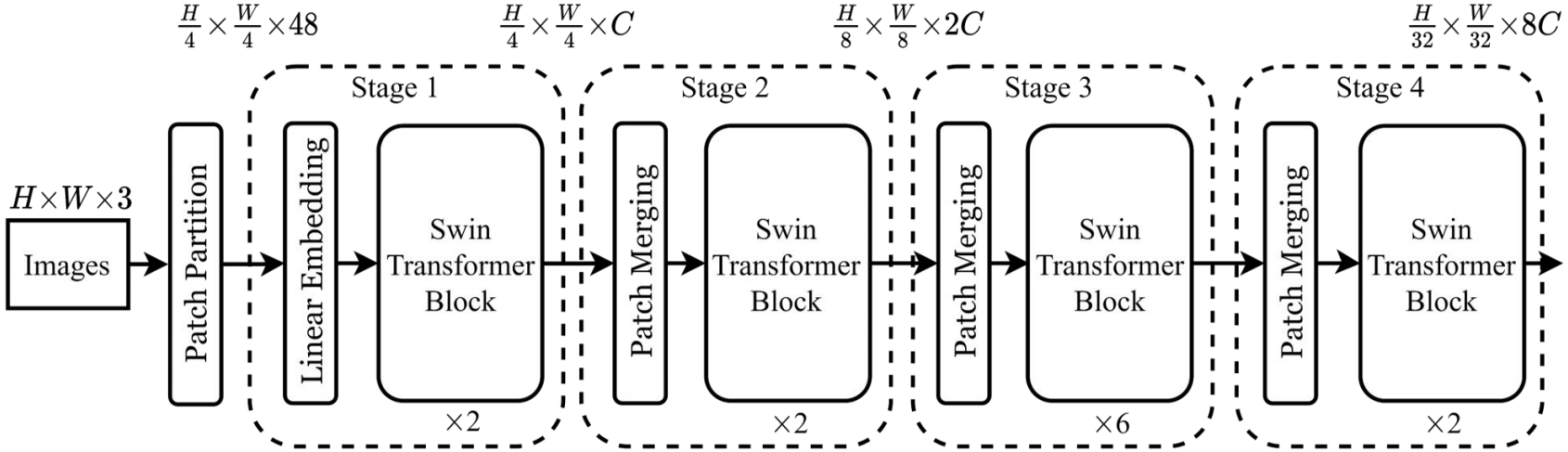
(a) Swin Transformer



(b) Shifted Window



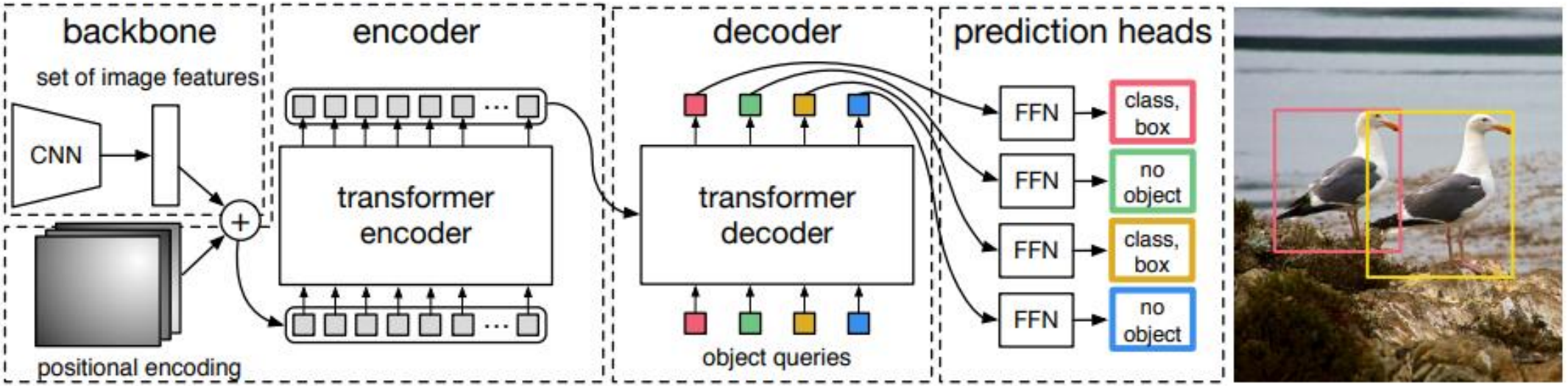
(c) Two Successive Swin Transformer Blocks



(d) Architecture

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer : Hierarchical Vision Transformer Using Shifted Windows. ICCV.

DETR: détecter = transformer des représentations



Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229).

Generative Pretrained Transformers

GPT1

Improving Language Understanding by Generative Pre-Training

Alec Radford, OpenAI, alec@openai.com
Karthik Narasimhan, OpenAI, karthikn@openai.com
Tim Salimans, OpenAI, tim@openai.com
Ilya Sutskever, OpenAI, ilyasu@openai.com

GPT2

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

GPT3

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei
OpenAI

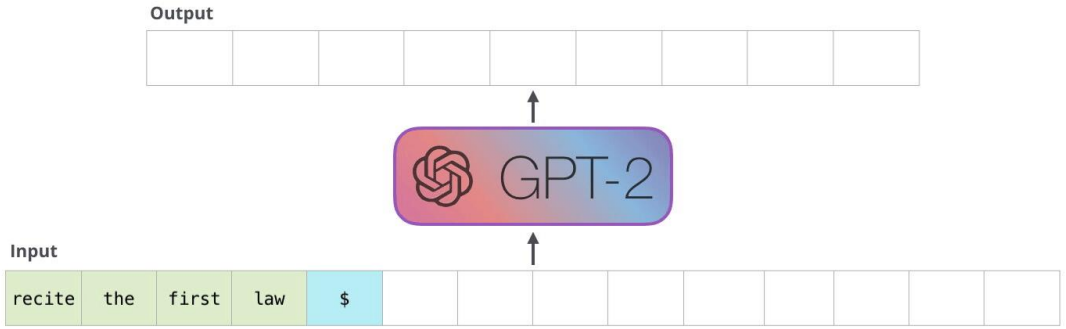
GPT4

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*
Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray
John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens
Amanda Askell† Peter Welinder Paul Christiano*†
Jan Leike* Ryan Lowe*
OpenAI

- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744.

Comment est appris GPT-1?



- Etape 1: « Self-supervision » prédiction de la séquence de tokens (= mots) $u_1, u_2 \ldots u_n$ à partir du contexte (les premiers mots) sur base $\mathcal{U}$.

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

- Etape 2: « Supervised Fine-Tuning » sur un problème de classification sur base $\mathcal{C}$

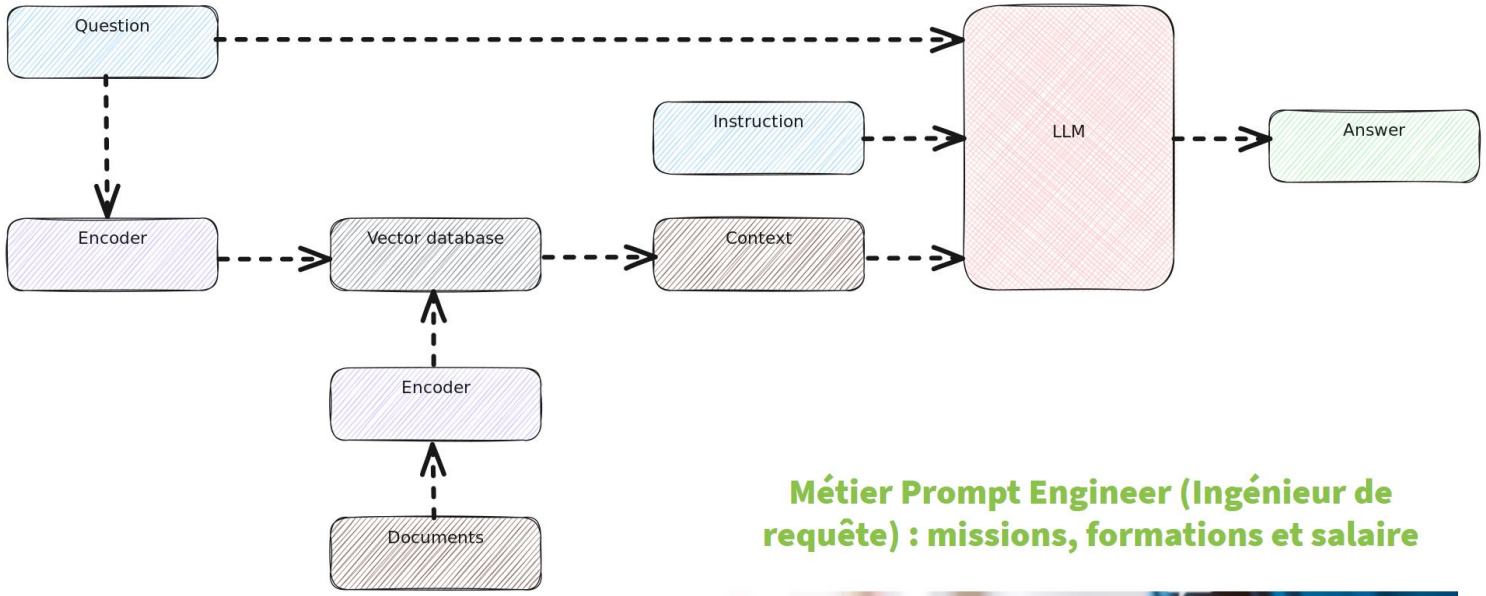
$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y | x^1, \ldots, x^m) \qquad P(y | x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y)$$

$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$

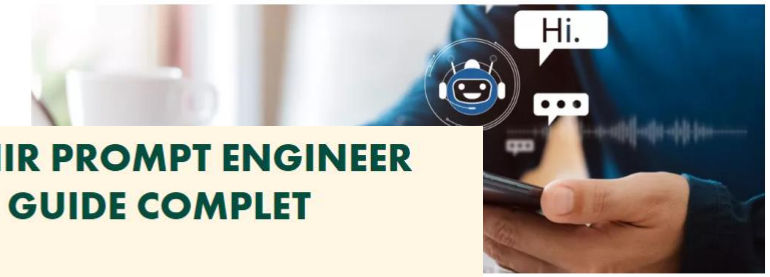
Le principe du prompt

- Types de prompt:
  - Zero-Shot Learning
  - One-Shot Learning
  - Few-Shot Learning
  - Chain-of-Thought Prompting
  - Iterative Prompting
  - Negative Prompting
  - Hybrid Prompting
  - Prompt Chaining
  - …

- In context « learning »
  - En fait pas vraiment d'apprentissage car pas d'adaptation du modèle!
  - Contextualisation par un texte concaténé à la requête



Métier Prompt Engineer (Ingénieur de requête) : missions, formations et salaire



ıre de prompt (requête, en français) pour les
vrez tout ce qu'il faut savoir sur le nouveau
formation, ses débouchés et même son

COMMENT DEVENIR PROMPT ENGINEER EN FREELANCE ? – GUIDE COMPLET

Mathias Savary • octobre 16, 2023

ONERA
THE FRENCH AEROSPACE LAB

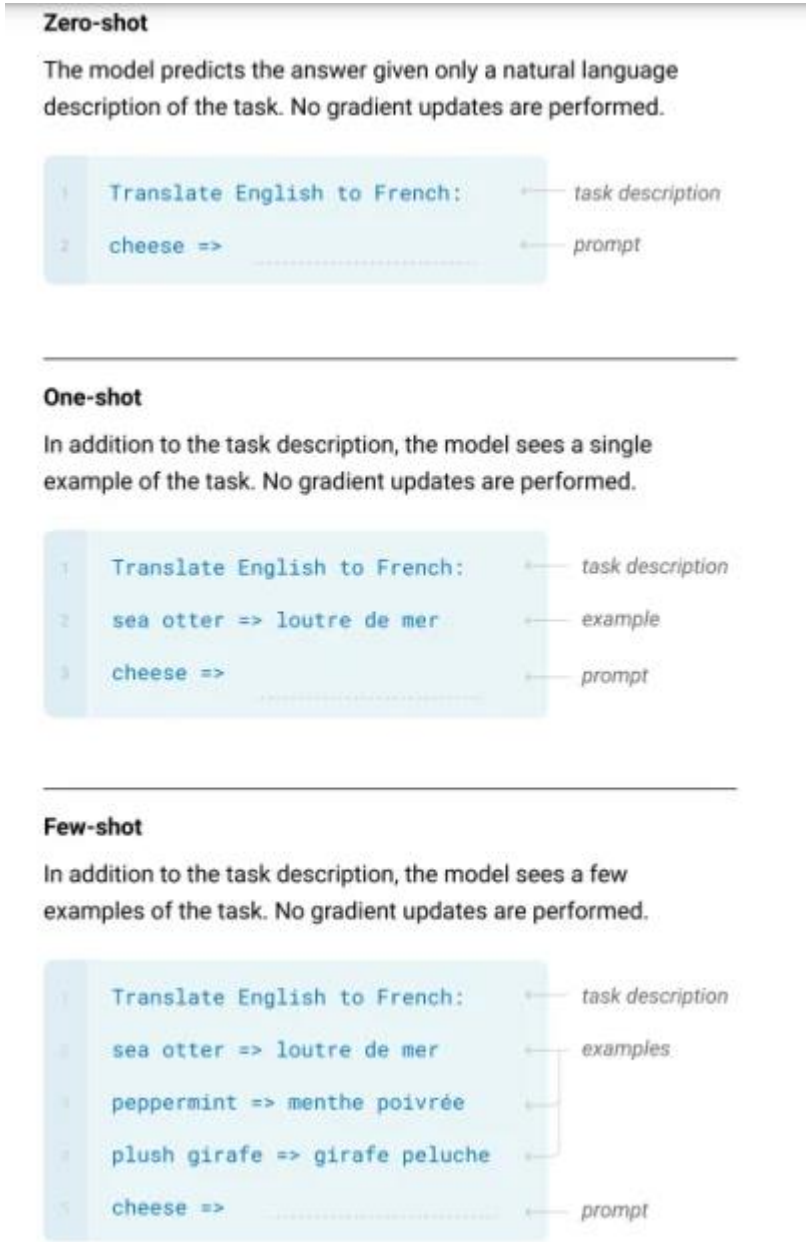
Les GPT suivants

GPT2
- Abandon du *supervised fine tuning* mais
  - Plus de données, un réseau plus grand
  - Des loss complémentaires (Perplexity, Fluency, Coherence, Diversity)
- « Decoder only »
- Evaluation sur des tâches non supervisées dans le pre-training (In context « learning »)

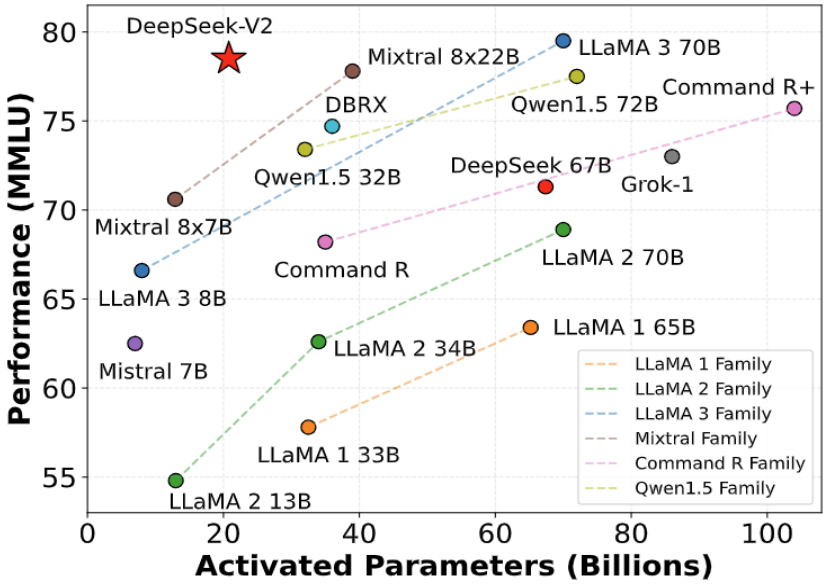
GPT3
- Abandon du « positional encoding »
- Plus grand, plus de données…

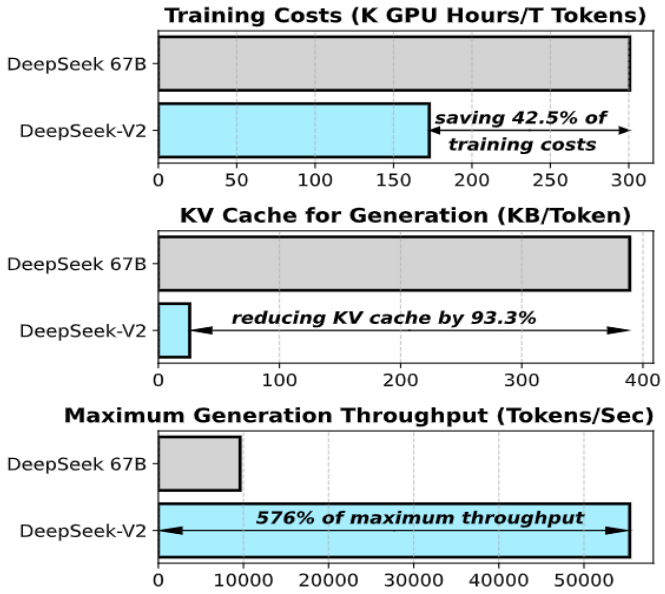
GPT4
- Multimodal (images)

ChatGPT
- Interface + Reinforcement Learning Human FineTuning

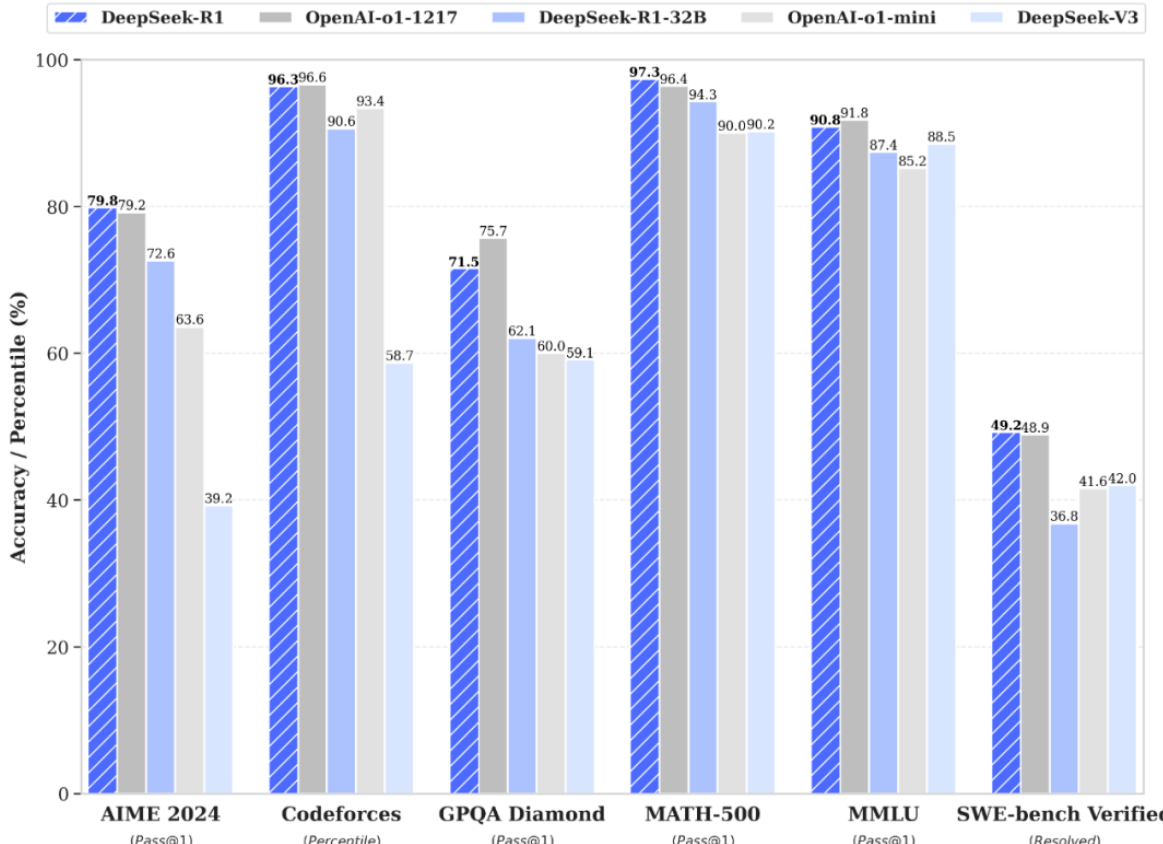


La déflagration DeepSeek





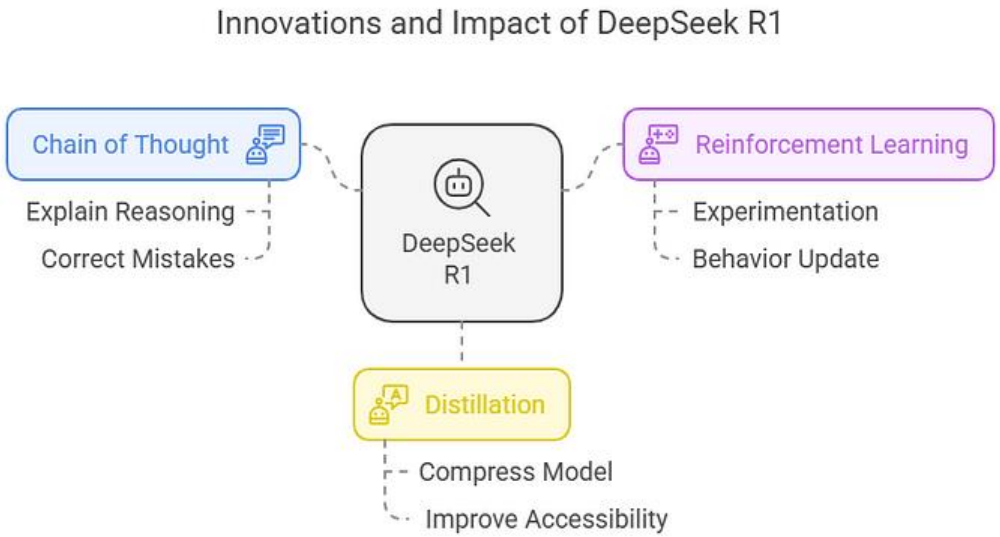
DeepSeek R1



DeepSeek saga

1. DeepSeek-LLM *(Jan '24)*: an early investigation of scaling laws and data-model tradeoffs.
2. DeepSeek-V2 *(Jun '24)*: introducing Multi-Head Latent Attention (MLA) and DeepSeekMoE to improve memory and training efficiency.
3. DeepSeek-V3 *(Dec '24)*: scaling sparse MoE networks to 671B parameters, with FP8 mixed precision training and intricate HPC co-design
4. DeepSeek-R1 *(Jan '25)*: building upon the efficiency foundations of the previous papers and using *large-scale reinforcement learning* to incentivize emergent chain-of-thought capabilities, including a “zero-SFT” variant.

https://martinfowler.com/articles/deepseek-papers.html



DeepSeek V2: évolutions architecturales → réduire les poids en inférence

$$\mathbf{h}'_t = \mathbf{u}_t + \sum_{i=1}^{N_s} \mathrm{FFN}_i^{(s)}(\mathbf{u}_t) + \sum_{i=1}^{N_r} g_{i,t}\,\mathrm{FFN}_i^{(r)}(\mathbf{u}_t),$$

$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \mathrm{Topk}(\{s_{j,t} | 1 \leqslant j \leqslant N_r\}, K_r), \\ 0, & \text{otherwise}, \end{cases}$$

$$s_{i,t} = \mathrm{Softmax}_i\left(\mathbf{u}_t{}^T \mathbf{e}_i\right),$$

$$\mathbf{q}_t = W^Q \mathbf{h}_t, \qquad [\mathbf{q}_{t,1}; \mathbf{q}_{t,2}; \ldots; \mathbf{q}_{t,n_h}] = \mathbf{q}_t,$$

$$\mathbf{k}_t = W^K \mathbf{h}_t, \qquad [\mathbf{k}_{t,1}; \mathbf{k}_{t,2}; \ldots; \mathbf{k}_{t,n_h}] = \mathbf{k}_t,$$

$$\mathbf{v}_t = W^V \mathbf{h}_t, \qquad [\mathbf{v}_{t,1}; \mathbf{v}_{t,2}; \ldots; \mathbf{v}_{t,n_h}] = \mathbf{v}_t,$$

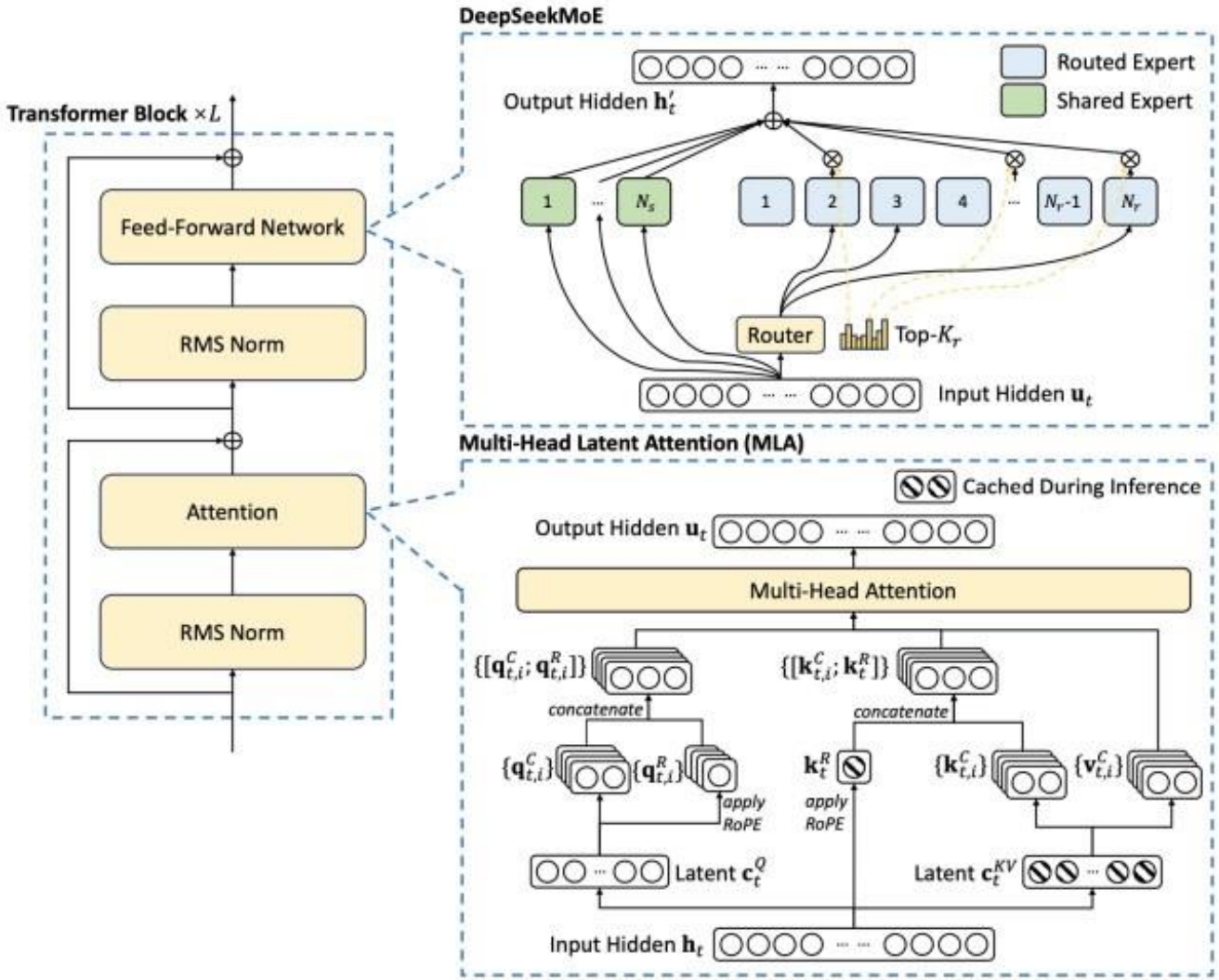
$$\mathbf{o}_{t,i} = \sum_{j=1}^{t} \mathrm{Softmax}_j\left(\frac{\mathbf{q}_{t,i}^T \mathbf{k}_{j,i}}{\sqrt{d_h}}\right)\mathbf{v}_{j,i},$$

$$\mathbf{c}_t^{KV} = W^{DKV} \mathbf{h}_t,$$

$$\mathbf{k}_t^C = W^{UK} \mathbf{c}_t^{KV},$$

$$\mathbf{v}_t^C = W^{UV} \mathbf{c}_t^{KV},$$

$$\mathbf{u}_t = W^O [\mathbf{o}_{t,1}; \mathbf{o}_{t,2}; \ldots; \mathbf{o}_{t,n_h}],$$



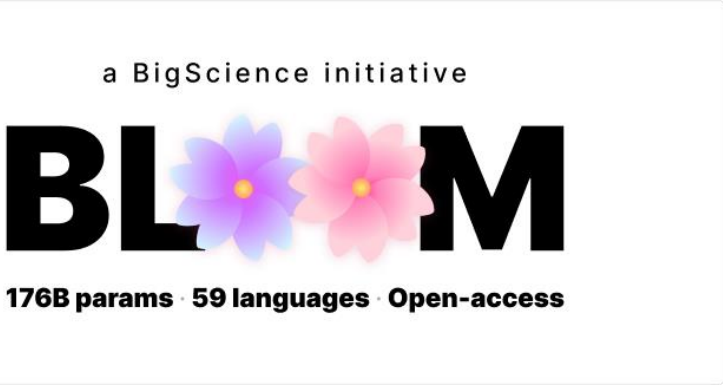
Large Language Models

open

Model		GPT-3	BLOOM	LLaMA	LLaMA-2	T5	PaLM
Developer		OpenAI	BigScience	Meta		Google	
Model Size (# parameters)		175B	175B	7B, 13B, 33B, 65B	7B, 13B, 34B, 70B	11B	540B
Training Data (# tokens)		300B	350B	1.4T	2T	34B	795B
Training Compute (FLOPs)		3.2E+23	3.7E+23	9.9E+23	1.5E+24	2.2E+21	2.6E+24
Processor	Manufacturer	Nvidia	Nvidia	Nvidia	Nvidia	Google	Google
	Type	GPU	GPU	GPU	GPU	TPU	TPU
	Model	V100	A100	A100	A100	TPU v3	TPU v4
Processor Hours		3,552,000	1,082,990	1,770,394	3,311,616	245,760	8,404,992
Grid Carbon Intensity (kgCO2e/KWh)		0.429	0.057	0.385	0.423	0.545	0.079
Data Center Efficiency (PUE)		1.1	1.2	1.1	1.1	1.12	1.08
Energy Consumption (MWh)		1,287	520	779	1,400	86	3,436
Carbon Emissions (tCO2e)		552	30	300	593	47	271

$$KWh = Hours\ to\ train \times Number\ of\ Processors \times Average\ Power\ per\ Processor \times PUE \div 1000$$

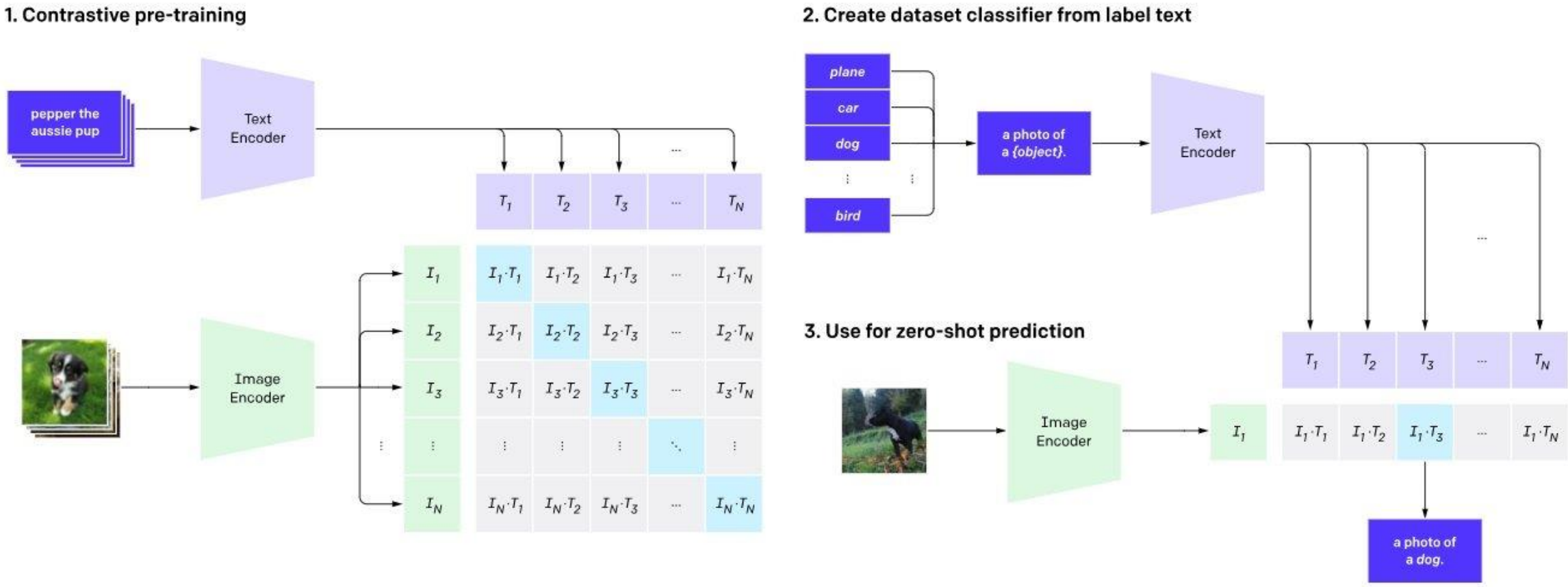
$$tCO2e = KWh \times kg\ CO2e\ per\ KWh \div 1000$$



Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., ... & Bari, M. S. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

ONERA
THE FRENCH AEROSPACE LAB

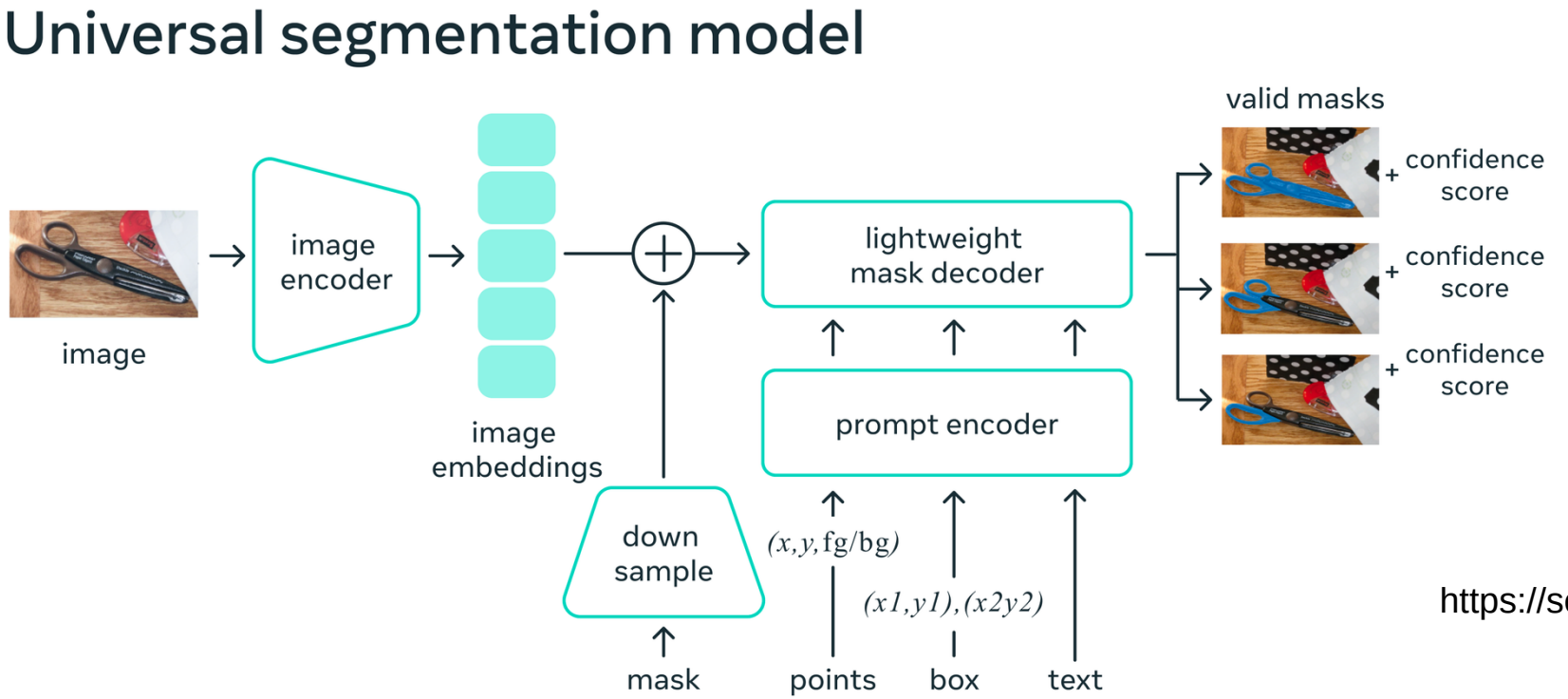
CLIP: Contrastive Language Image Pre-training



- Contrastive loss pour apprendre des codes appariés texte et image
- Zero-shot classification facile (un prompt par hypothèse + décision selon score de similarité)

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning, 8748-8763.

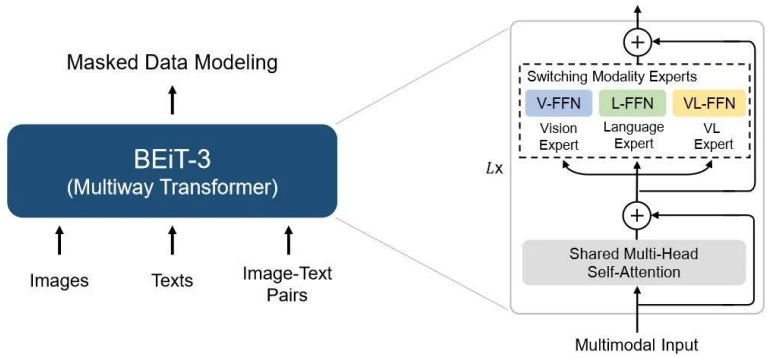
SAM: Segment Anything Model

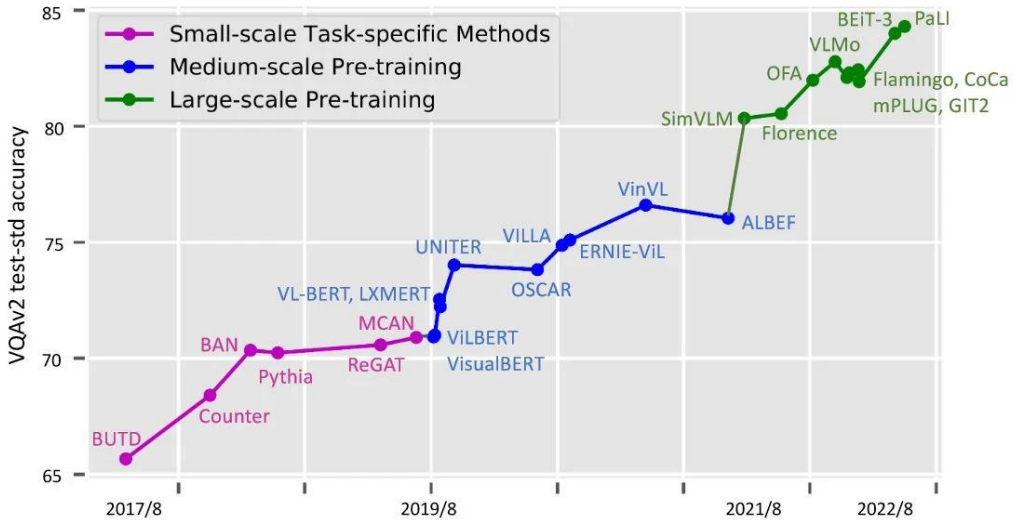


https://segment-anything.com/

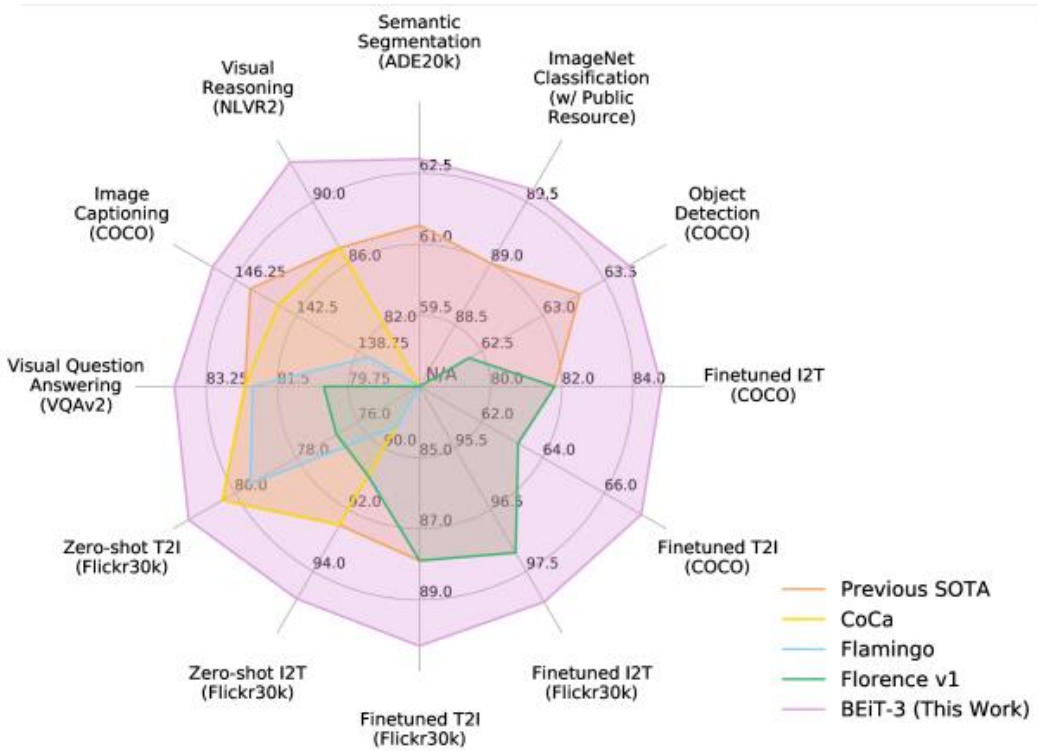
- « SAM has learned a general notion of what objects are »
- Utilise des prompt graphiques (en plus du textuel) pour cibler des objets

Modèles de fondation vision & langage





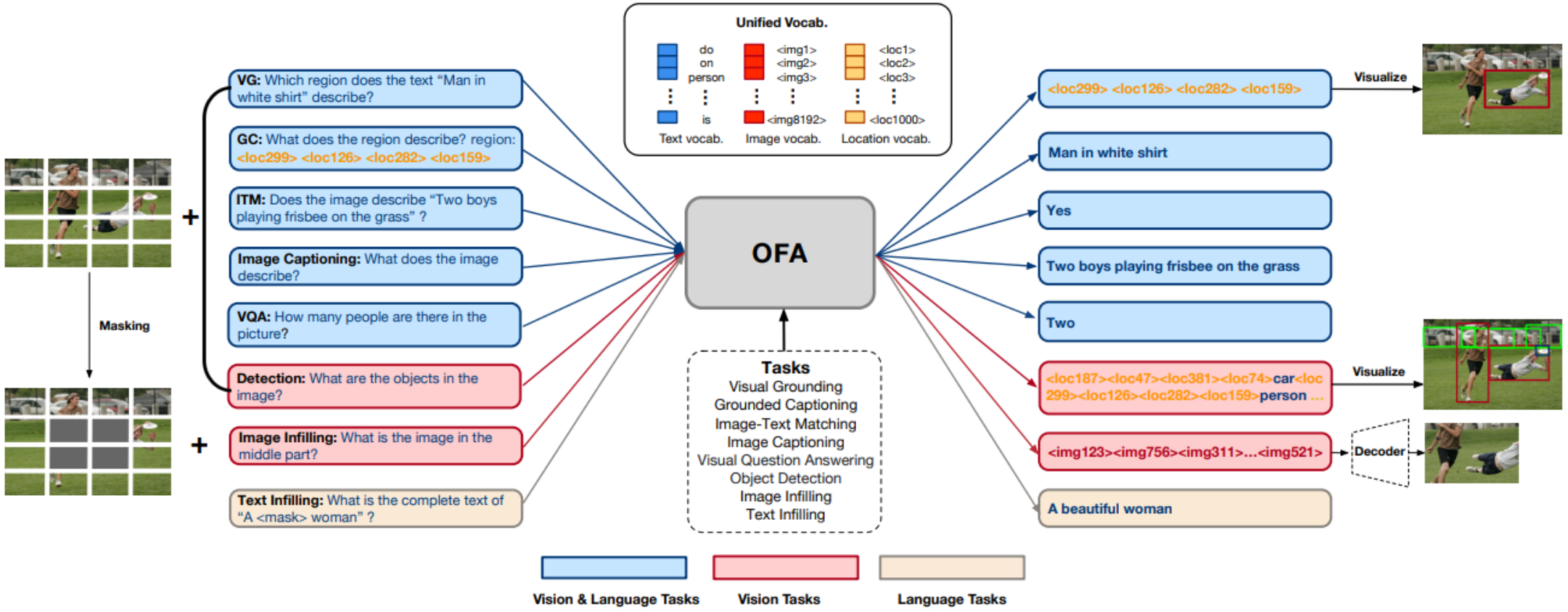
Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., & Gao, J. (2022). Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision, 14(3–4), 163-352.



Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., ... & Wei, F. (2022). Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442.

ONERA
THE FRENCH AEROSPACE LAB

Tâche de vision = un (bon) prompt et c'est tout?



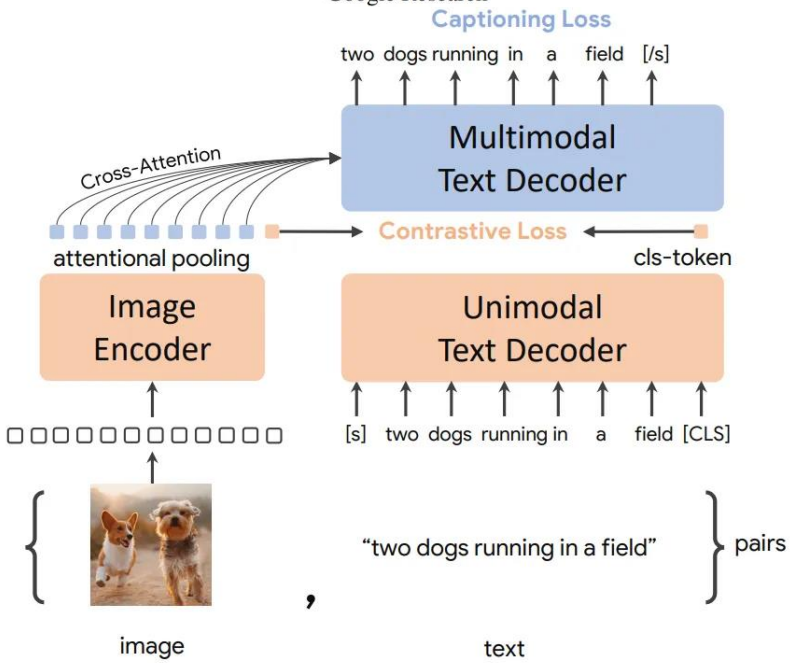
Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., ... & Yang, H. (2022). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. ICML

Ne vous trompez pas de Coca!

Co-créée avec
l'intelligence artificielle

CoCa: Contrastive Captioners are Image-Text Foundation Models

Jiahui Yu† **Zirui Wang**†
{jiahuiyu, ziruiw}@google.com
Vijay Vasudevan **Legg Yeung** **Mojtaba Seyedhosseini** **Yonghui Wu**
Google Research



Utiliser les modèles de fondation

Pour résoudre une tâche

- « In context learning »: tâche définie par le prompt.
- « Fine tuning »: adaptation des poids à un contexte. Demande d'avoir des données (idéalement peu).
- « Transfer Learning »: on ne garde qu'une partie du modèle
- Distillation de connaissance: utilisé pour régulariser l'apprentissage d'un autre modèle plus spécialisé.
- « Prompt Learning or Tuning »: on calcule un conditionnement (numérique) adapté

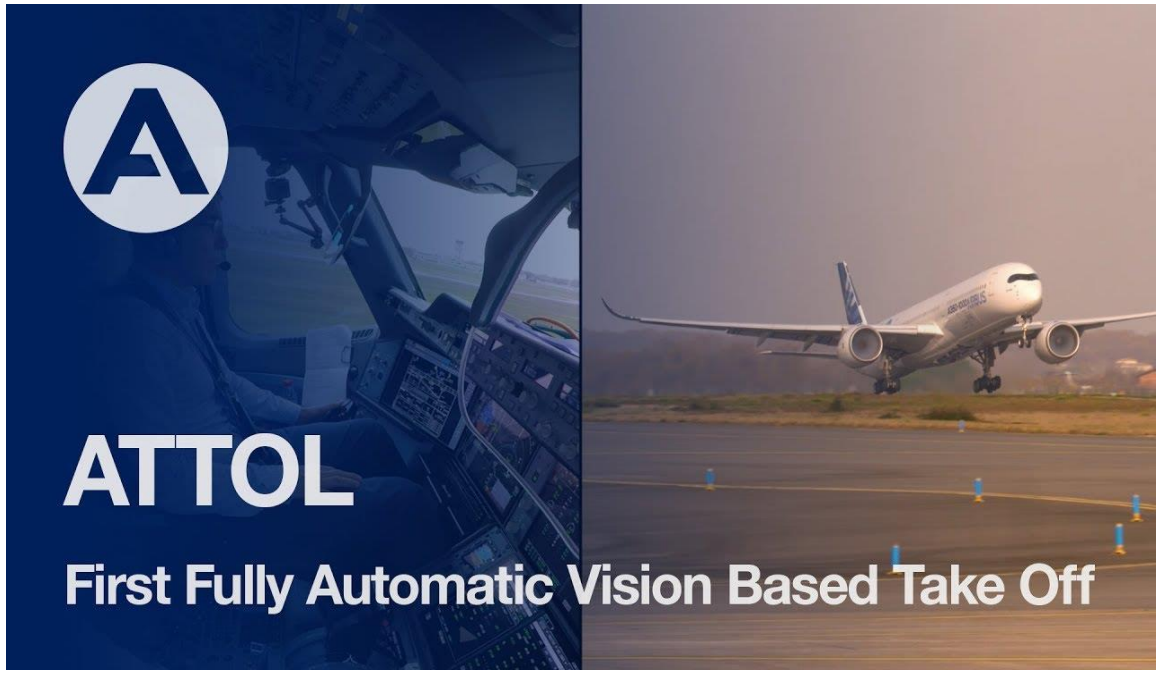
Les limitations

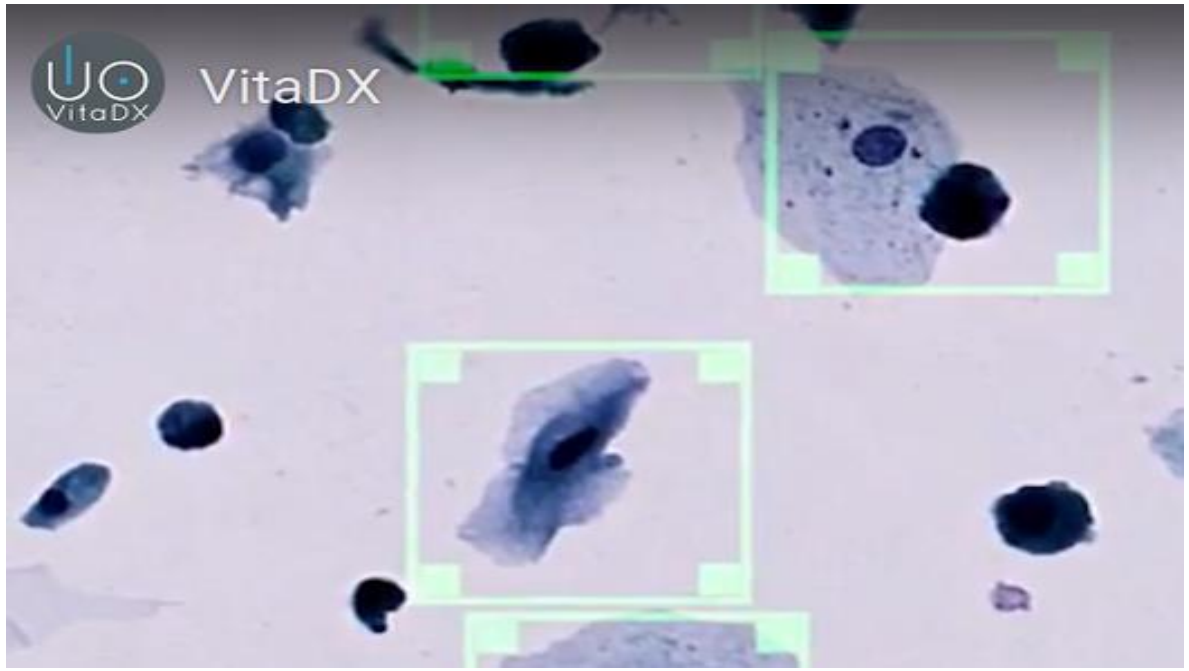
- Les modèles de fondation sont gros (plusieurs centaines de millions de paramètres) et calculatoires (0.001 - 0.002 kg CO_2 pour une requête GPT4)

  ➔ « Parameter Efficient Fine Tuning »

IA DE CONFIANCE

Des applications critiques potentielles?





Forbes Billionaires Innovation Leadership Money Business Small Business Lifestyle L

32 096 views | May 16, 2019, 12:13pm EDT

Investigators Say Tesla Model 3 Driver Killed In Florida Crash Was Using Autopilot

Alan Ohnsman Forbes Staff
Transportation

The roof of Jeremy Beren Banner's 2018 Tesla Model 3 was sheared off when the car struck a semi-truck trailer in Delray Beach, Florida, on March 1, 2019. [-] NATIONAL TRANSPORTATION SAFETY BOARD

(Updates with Tesla comments)

Safety officials investigating a deadly crash in Florida in which a Tesla Mo driver was killed in a collision with a commercial truck said the electric-c maker's semi-automated driving system was being used at the time and that neither it nor the driver took evasive action.

The National Transportation Safety Board issued preliminary findings from its investigation of the March 1 crash in Delray Beach that killed 50-year-old Jeremy Beren Banner, who died after his 2018 Model 3 hit a semi-truck that crossed his path on Florida State Highway 441 in early morning traffic. The Tesla, travelling at 68 miles per hour on a section of highway with a posted speed limit of 55 mph, slammed into the side of the truck's trailer, shearing the Tesla's roof off. The truck driver was unharmed.

Cookies on Forbes

Comment éviter ca?

The New York Times

Amazon Pauses Police Use of Its Facial Recognition Software

The company said it hoped the moratorium "might give Congress enough time to put in place appropriate rules" for the technology.

Civil liberties advocates began calling for a ban on the use of facial recognition by law enforcement in 2018. Elaine Thompson/Associated Press

By Karen Weise and Natasha Singer

June 10, 2020

SCIENCES • INTELLIGENCES ARTIFICIELLES GÉNÉRATIVES

Intelligences artificielles, les mille et une façons de les faire dérailler

Par David Larousserie

Publié hier à 06h00, modifié hier à 06h38

Lecture 11 min.

Le Monde, 13/2/24

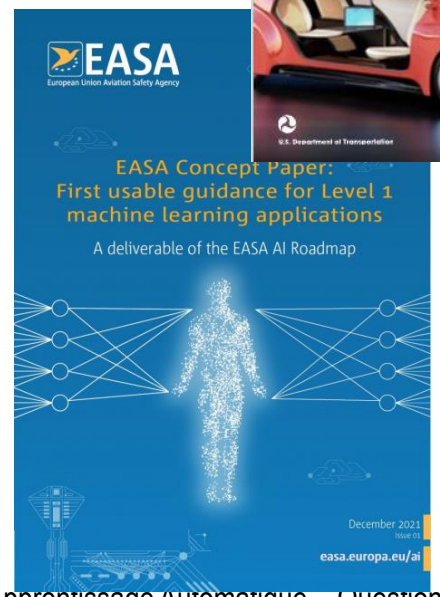
Nouvelle problématique: l'IA de confiance!

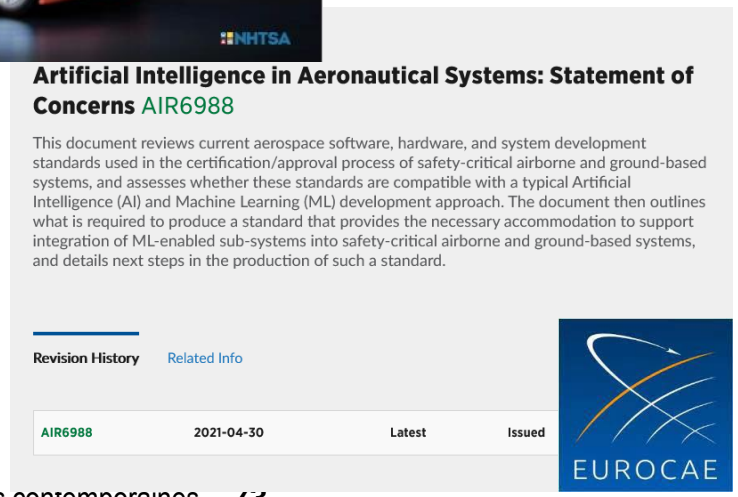
- Le qualificatif: « trustworthiness »

- Projets de recherche
  - Grand défi IA (3IA ANITI, Confiance.ai)
- « White papers »
  - UE, France, Allemagne, USA
- Standards
  - ISO, EASA, NHTSA…
- Industrie
  - IBM, Microsoft…

https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html

Grande activité normative





European AI Act

Législation sur l'intelligence artificielle: le Conseil et le Parlement parviennent à un accord sur les premières règles au monde en matière d'IA

À l'issue de trois jours de pourparlers-marathon, la présidence du Conseil et les négociateurs du Parlement européen sont parvenus à un accord provisoire sur la proposition relative à des règles harmonisées concernant l'intelligence artificielle (IA), dénommée la "**législation sur l'intelligence artificielle**". Le projet de règlement a pour objectif de veiller à ce que les systèmes d'IA mis sur le marché européen et utilisés dans l'UE soient **sûrs** et à ce qu'ils respectent les **droits fondamentaux** et les valeurs de l'UE. Cette proposition qui fait date vise également à stimuler l'investissement et l'innovation dans le domaine de l'IA en Europe.

> Il s'agit là d'une réalisation historique et d'un pas de géant vers l'avenir! L'accord conclu aujourd'hui répond efficacement à un défi mondial dans un environnement technologique en évolution rapide concernant un domaine clé pour l'avenir de nos sociétés et de nos économies. Et dans cette entreprise, nous sommes parvenus à maintenir un équilibre extrêmement délicat: encourager l'innovation et l'adoption de l'intelligence artificielle dans toute l'Europe tout en respectant pleinement les droits fondamentaux de nos citoyens.
>
> — Carme Artigas Brugal, secrétaire d'État espagnole à la numérisation et à l'intelligence artificielle

https://www.consilium.europa.eu/fr/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

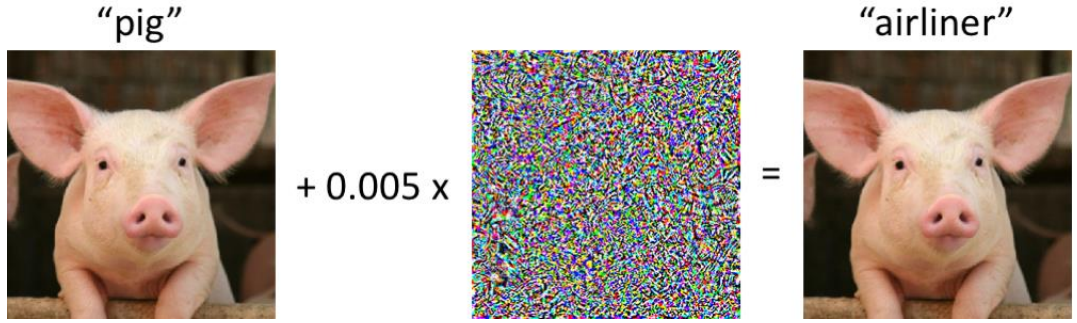
Contrôler et repérer les situations de défaillance
Adversaires

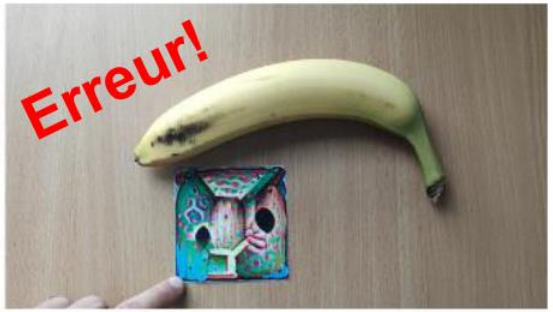
Robustesse

Plusieurs causes de défaillance

- Attaques adversaires bornées
- Patchs adversaires
- Empoisonnement de données
- Corruption

- Peut-on limiter leur effet?
- Sont-elles dangereuses?



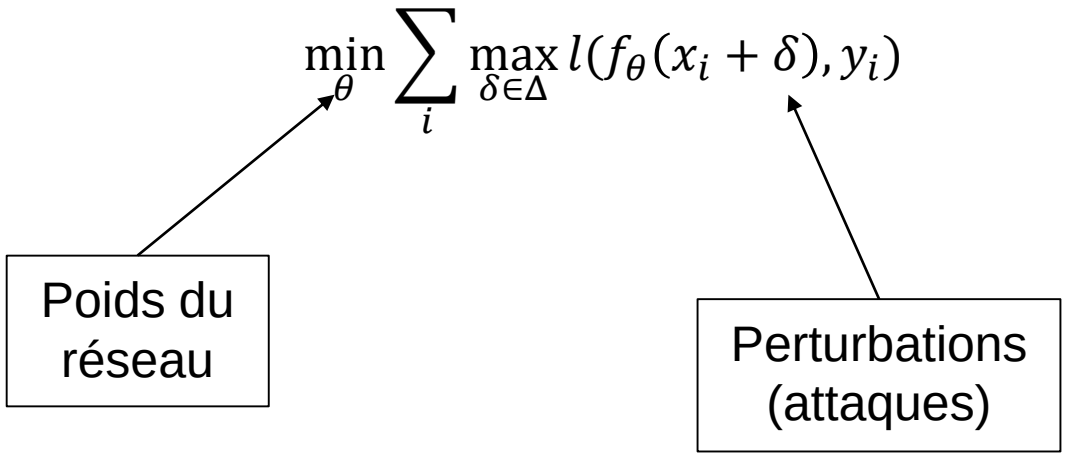


[Labarbarie et al., 2024]

[Thys et al., 2019]

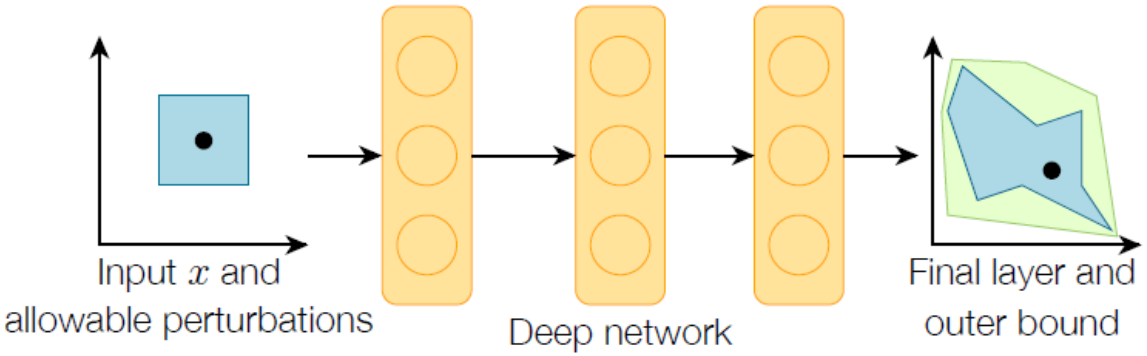
[Eykholt et al., 2018]

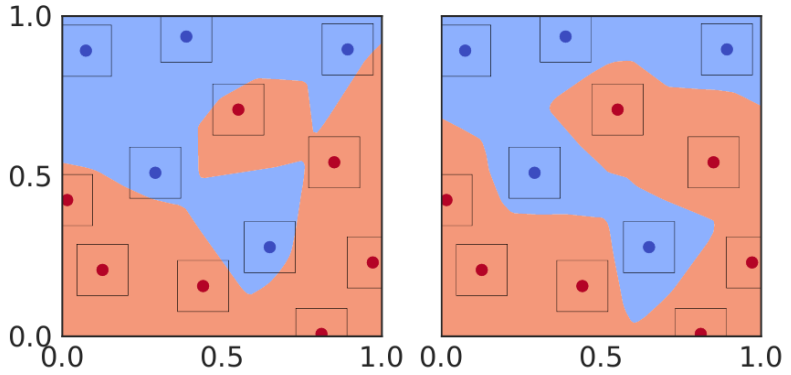
Une première solution: Apprentissage adversarial



- Objectif = minimiser l'impact de perturbations $\delta \in \Delta$ sur l'apprentissage
  = pb. Min/max
- Principe = générer des attaques (le « max ») pendant l'apprentissage (le « min »)
- C'est sous-optimal car dépendant des attaques
- En général: robustification mais baisse de performance

Approches de vérification formelle

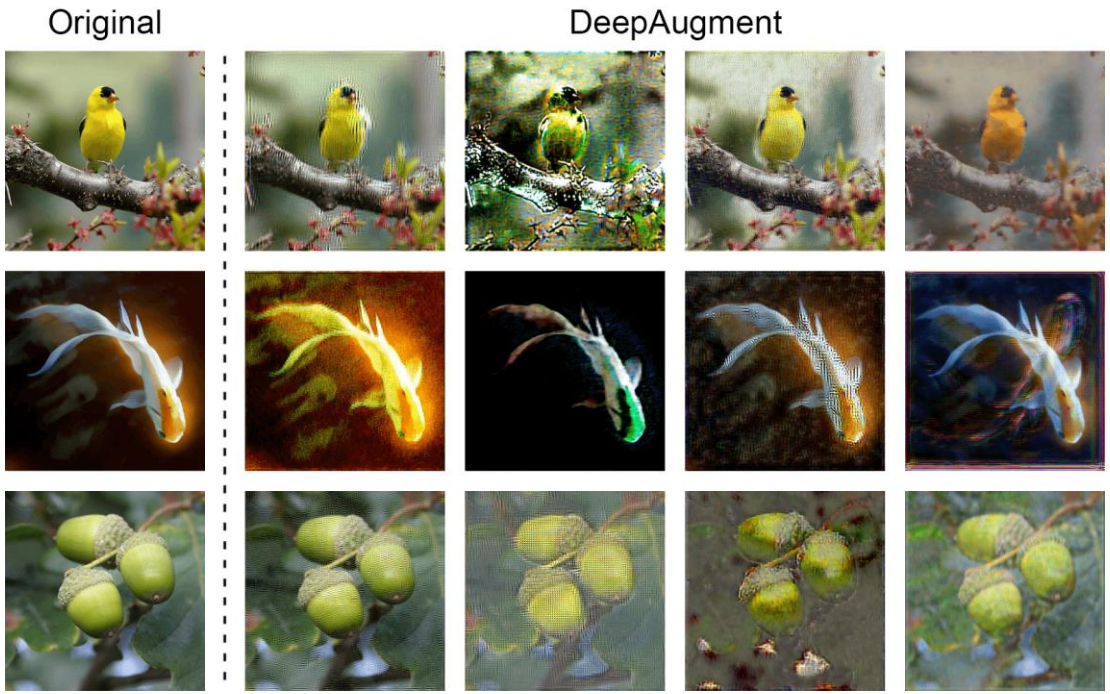




[Wong et al., 2017]

- Utilisation de méthodes d'optimisation exacte (LP + convexification)
- Certification d'exemples = garantie que la valeur de leur prédiction dans une boule est stable = ce n'est pas un adversaire
- Permet aussi d'apprendre de manière robuste (les convexifiés sont dérivables % poids du réseau)
- Mais: temps de calcul important, performances finales médiocres après apprentissage

Autre démarche: corruptions « réalistes »



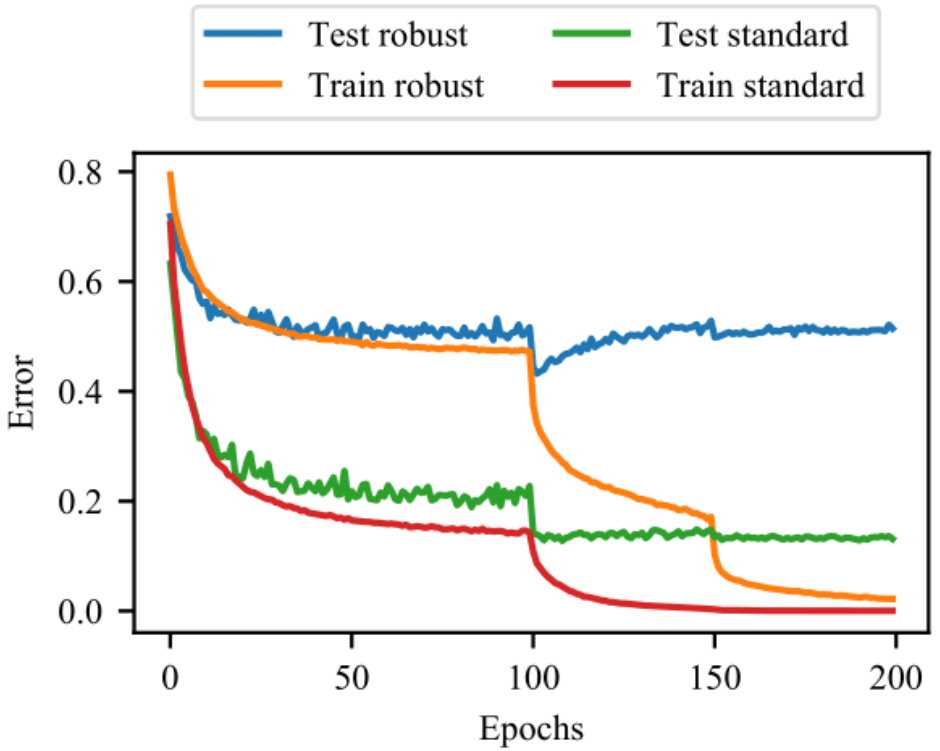
[Hendrycks, 2020]
https://github.com/hendrycks/imagenet-r

- Problèmes de classification grande échelle mais faible résolution.
- Autres contextes (scènes urbaines, télédétection) et tâches (détection, segmentation)?

“These new datasets demonstrate the importance of conducting multi-faceted evaluations of robustness as well as the general complexity of the landscape of robustness research, where it seems that so far nothing consistently helps in all settings.” (conclusion of [Hendrycks, 2020])

Bilan des approches adversaires

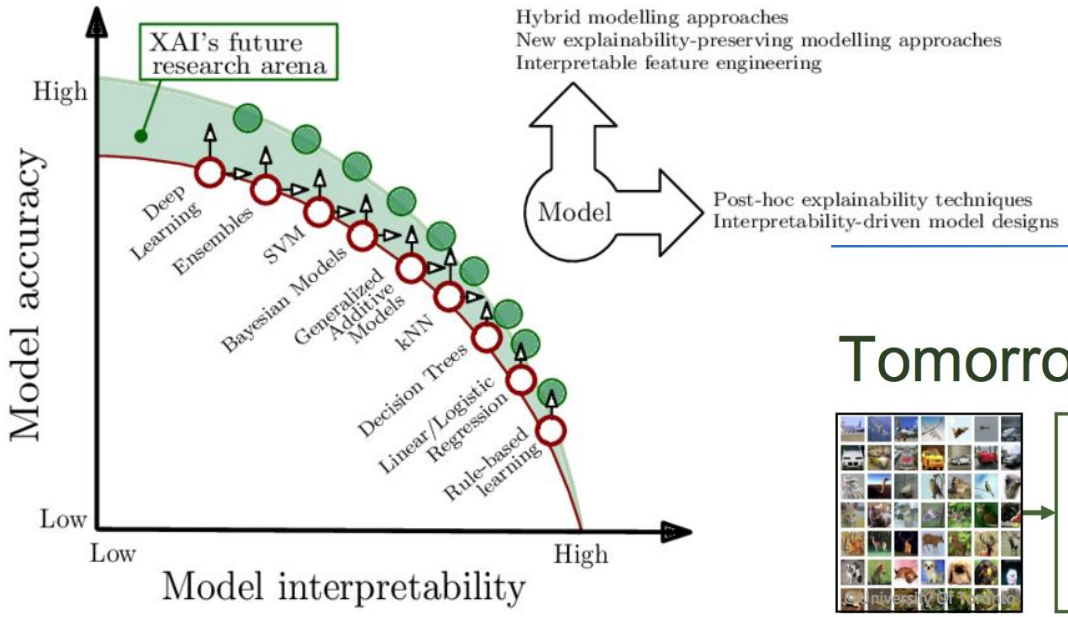
- Enjeu théorique
  - Compréhension des DNN
  - Robustification de l'apprentissage
- Limitations pratiques
  - Garanties locales (= il faut des exemples)
  - Lien avec généralisation à construire
  - Les attaques malveillantes efficaces demandent un accès au DNN (« white box »)
  - Les attaques physiques réalistes se développent, mais leur transfert sur plusieurs modèles reste difficile (tant mieux!)
- Il y a d'autres approches de robustesse
  - Randomisation, Réseaux Lipsichtz, etc.

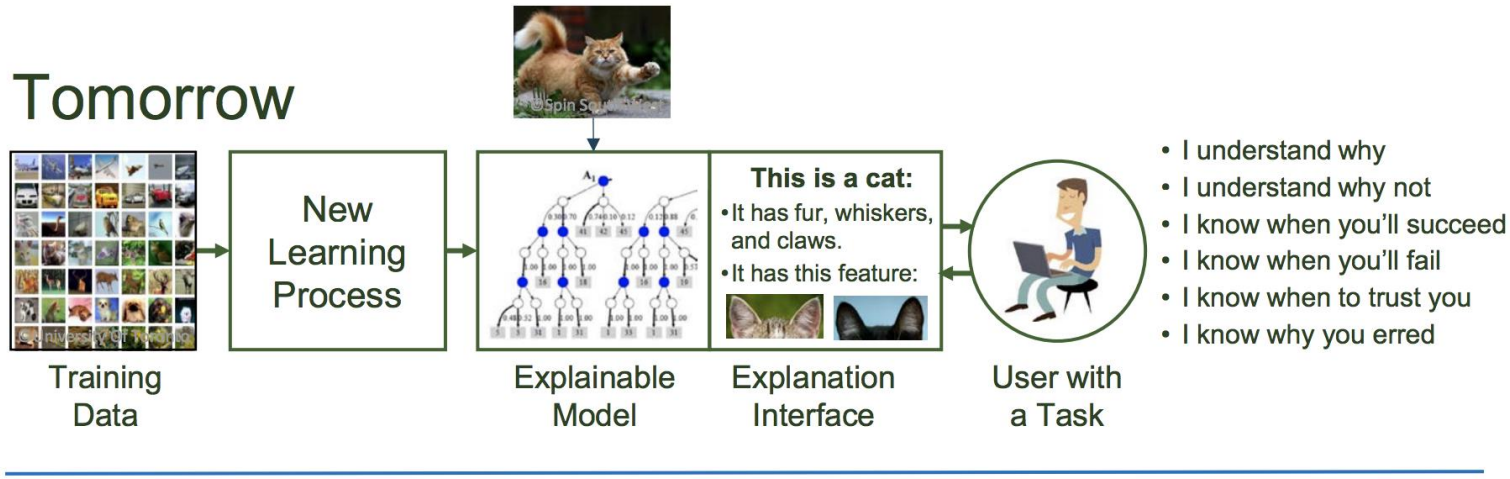


Rice, Leslie, Eric Wong, and Zico Kolter. (2020) "Overfitting in adversarially robust deep learning." ICML..

Gagner la confiance
Explicabilité

Explicabilité

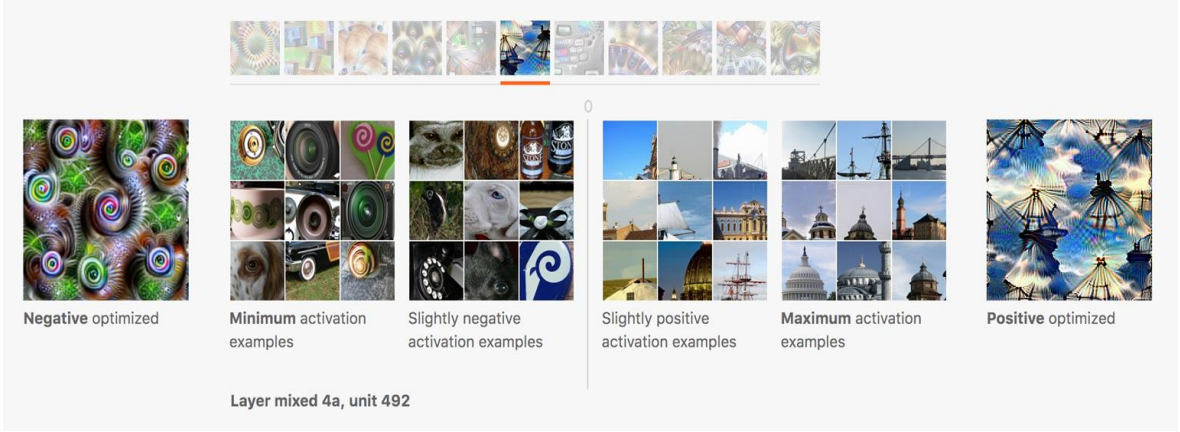




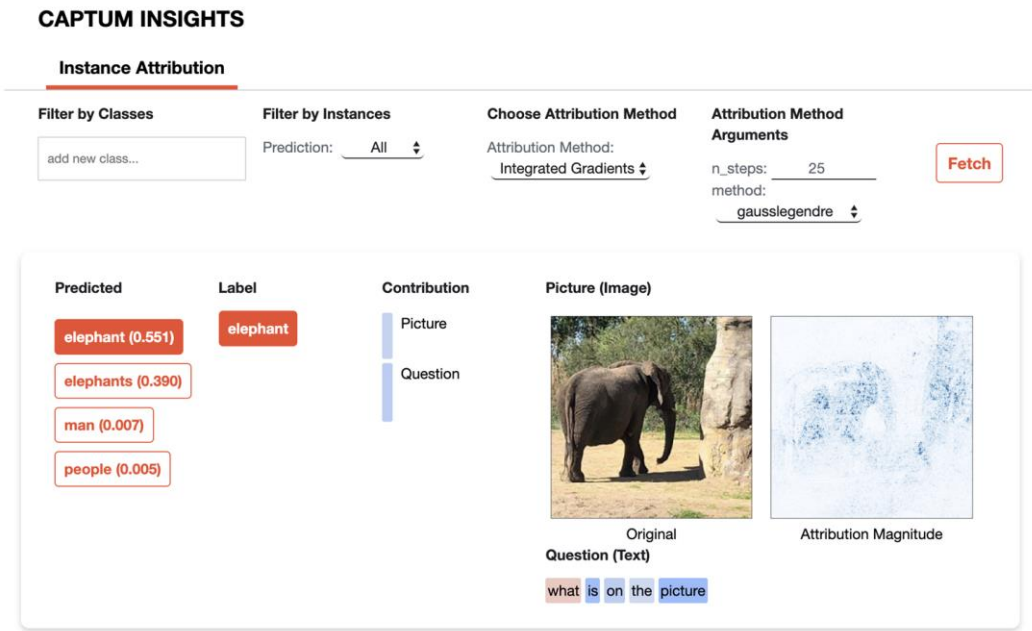
- Un ingrédient potentiel pour améliorer la confiance
- Compromis entre performance et interprétabilité: est-ce acceptable ?

Explicabilité

- De quoi?
  - Comportement, caractéristiques, propriétés
  - Prédiction (local) or prédicteur (global)
- Pour qui?
  - End-user, expert du domaine, data scientist, chercheur en IA, autorité, étudiant, auditeur
- Pour quoi faire?
  - Description de comportement, debugging, enseignement, confiance, contrôle, investigation, compréhension, identification de domaine, etc.
- Comment?
  - Post-hoc vs. par conception, model specific vs. agnostic



https://distill.pub/2017/feature-visualization/



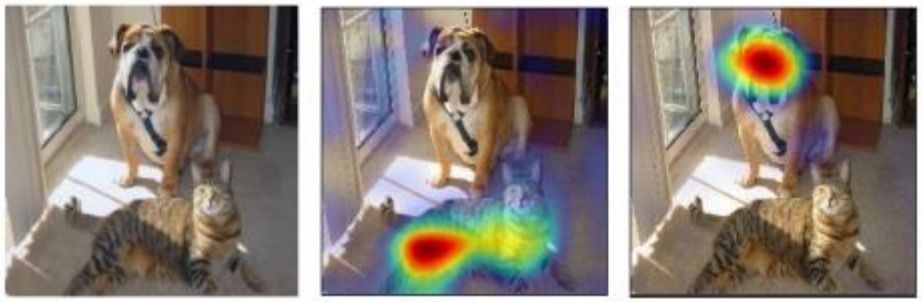
https://captum.ai/

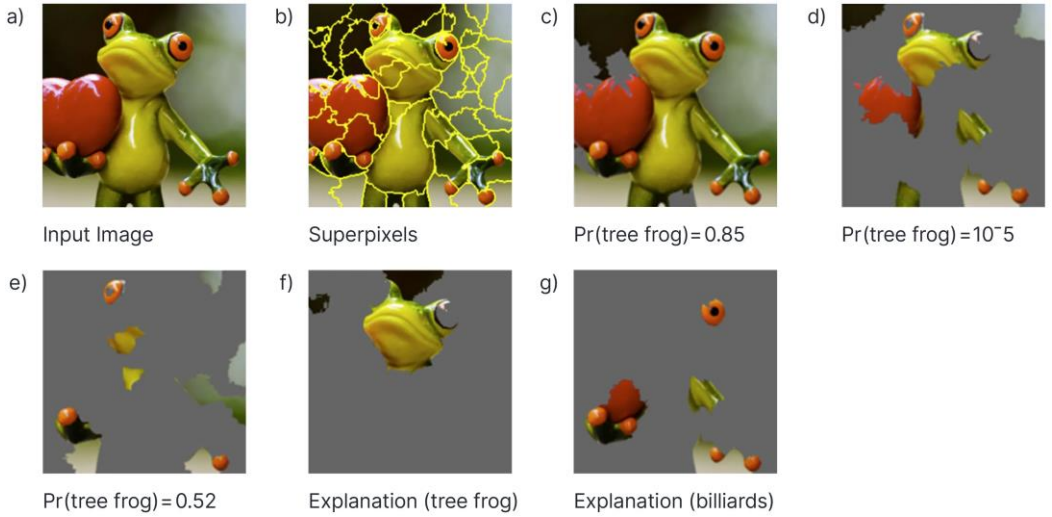
ONERA
THE FRENCH AEROSPACE LAB

Explicabilité post-hoc

Legend: ■ Negative □ Neutral ■ Positive

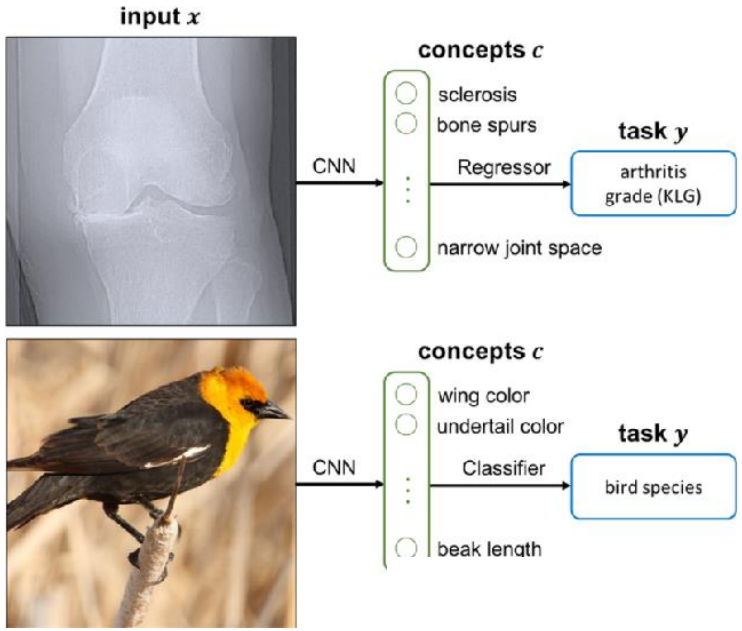
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
pos	pos (0.96)	pos	1.29	it was a fantastic performance ! #pad
pos	pos (0.87)	pos	1.56	best film ever #pad #pad #pad #pad
pos	pos (0.92)	pos	1.14	such a great show ! #pad #pad
neg	neg (0.29)	pos	-1.11	it was a horrible movie #pad #pad
neg	neg (0.22)	pos	-1.03	i 've never watched something as bad
neg	neg (0.07)	pos	-0.84	that is a terrible movie . #pad



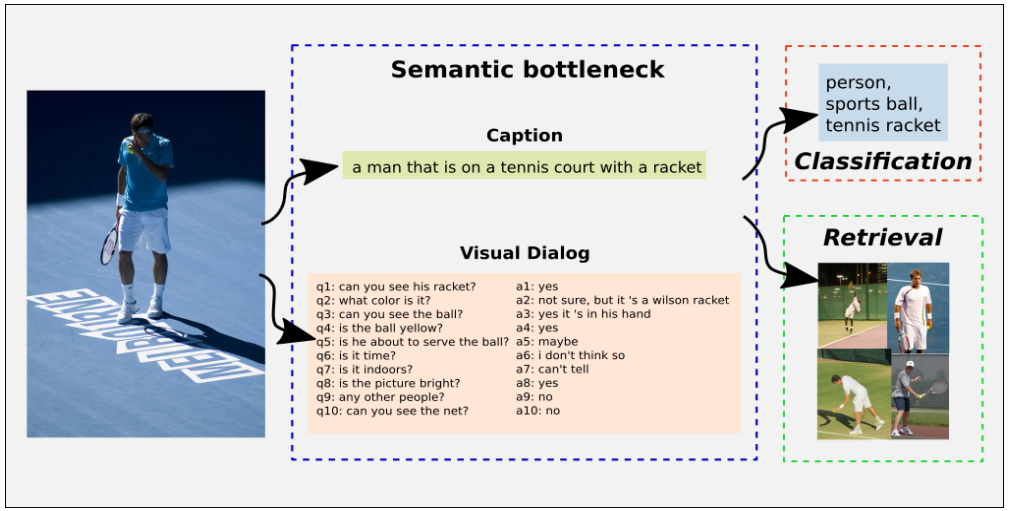


- Repère les éléments contribuant « le plus » à la décision (attribution de caractéristiques) en exploitant les gradients % entrée
- Repère les éléments sensibles par perturbation

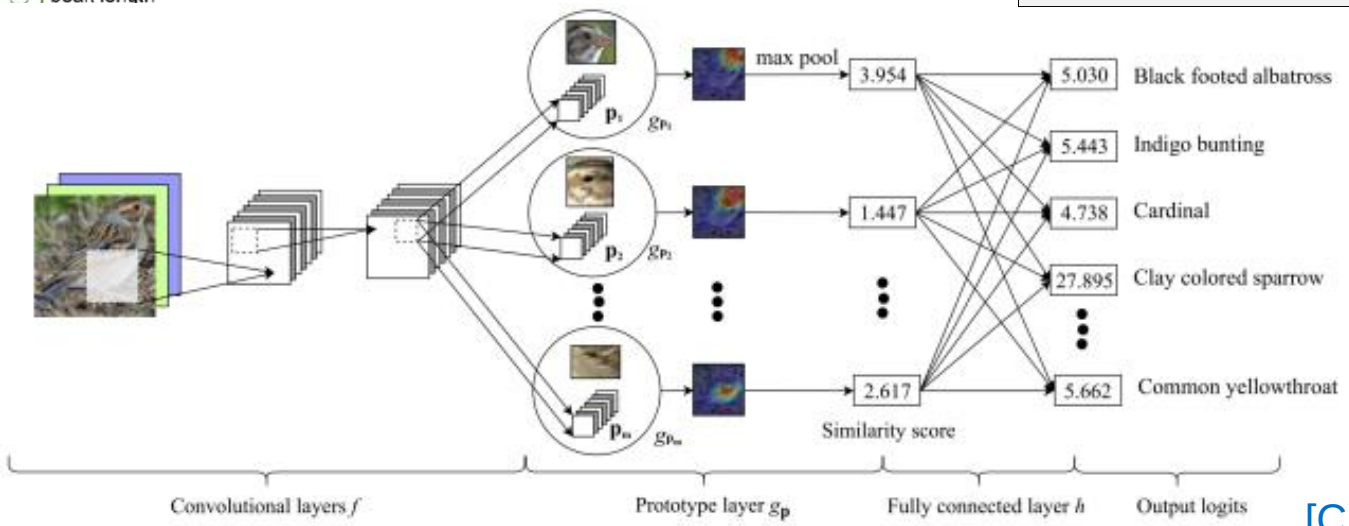
Explicabilité par conception



[Koh et al., 2020]



[Bucher, Herbin, Jurie, 2018]



[Chen et al., 2019]

Faire dépendre la prédiction d'un état latent **interprétable**

Qu’est-ce qu’une bonne explication?

- **« Explanation:** a formal representation that **causally** depends on **system features** (processing, internal states, architecture, etc.), is **interpretable** by humans, and contains **predictive** information of system **behavior »**
- Evaluation?
  - Des critères abstraits (fidélité, sensibilité, consistance…)
  - Peu de prise en compte des facteurs humains
  - Indépendant du rôle attendu de l’explication
- L’intérêt pratique de l’explicabilité est encore à démontrer.

Définir et garantir le « bon fonctionnement »
Domaine opérationnel

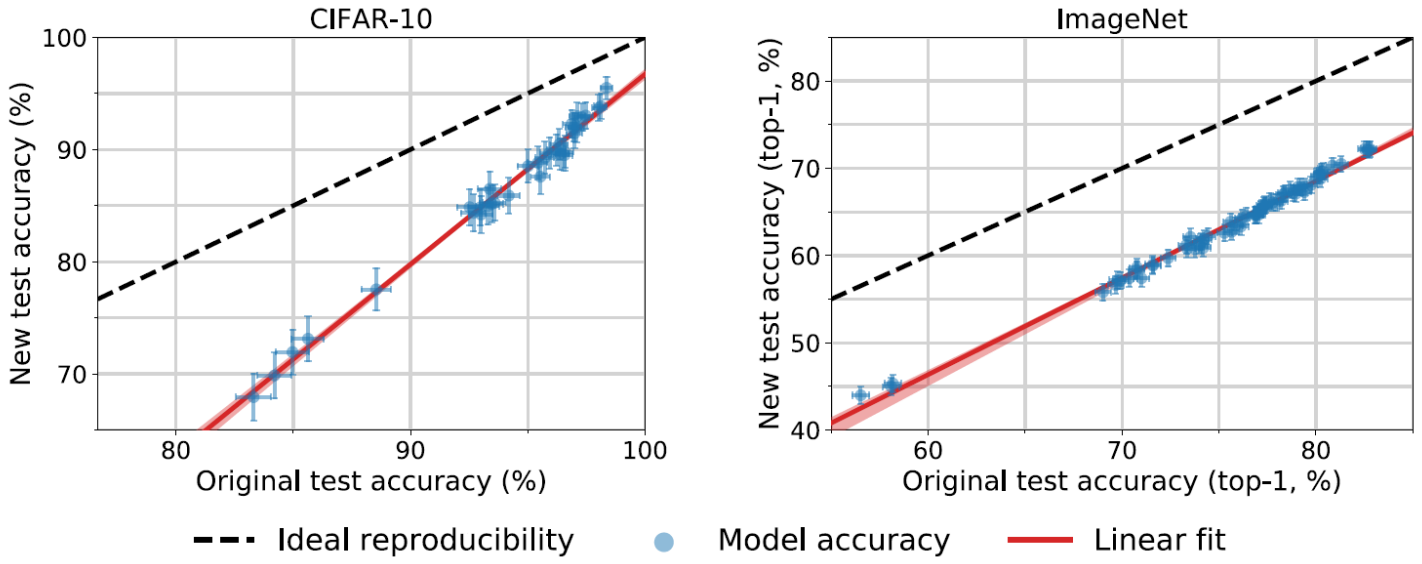
Définition

- Domaine Opérationnel = ensemble des conditions vérifiables de (bon) fonctionnement attendu

- « Attendu »
  - Spécification explicite
- « Vérifiable »
  - Doit disposer de moyens (métrique, calcul, visualisation, etc.) pour tester que les exigences sont vérifiées

Reproduction de dataset de test

- Reproduire les performances du dataset de test originel est difficile: baisse de performance, mais classement préservé
- Causes?
  - Pas de sur-apprentissage des hyper-paramètres.
  - Mais répartition exemples faciles/difficiles différente dans chaque catégorie (« selection frequency ») [Recht, 2019][Engstrom, 2020]



ImageNet Top-1

Orig. Rank	Model	Orig. Accuracy	New Accuracy	Gap	New Rank	Δ Rank
1	pnasnet_large_tf	82.9 [82.5, 83.2]	72.2 [71.3, 73.1]	10.7	3	-2
4	nasnetalarge	82.5 [82.2, 82.8]	72.2 [71.3, 73.1]	10.3	1	3
21	resnet152	78.3 [77.9, 78.7]	67.0 [66.1, 67.9]	11.3	21	0
23	inception_v3_tf	78.0 [77.6, 78.3]	66.1 [65.1, 67.0]	11.9	24	-1
30	densenet161	77.1 [76.8, 77.5]	65.3 [64.4, 66.2]	11.8	30	0
43	vgg19_bn	74.2 [73.8, 74.6]	61.9 [60.9, 62.8]	12.3	44	-1
64	alexnet	56.5 [56.1, 57.0]	44.0 [43.0, 45.0]	12.5	64	0
65	fv_64k	35.1 [34.7, 35.5]	24.1 [23.2, 24.9]	11.0	65	0

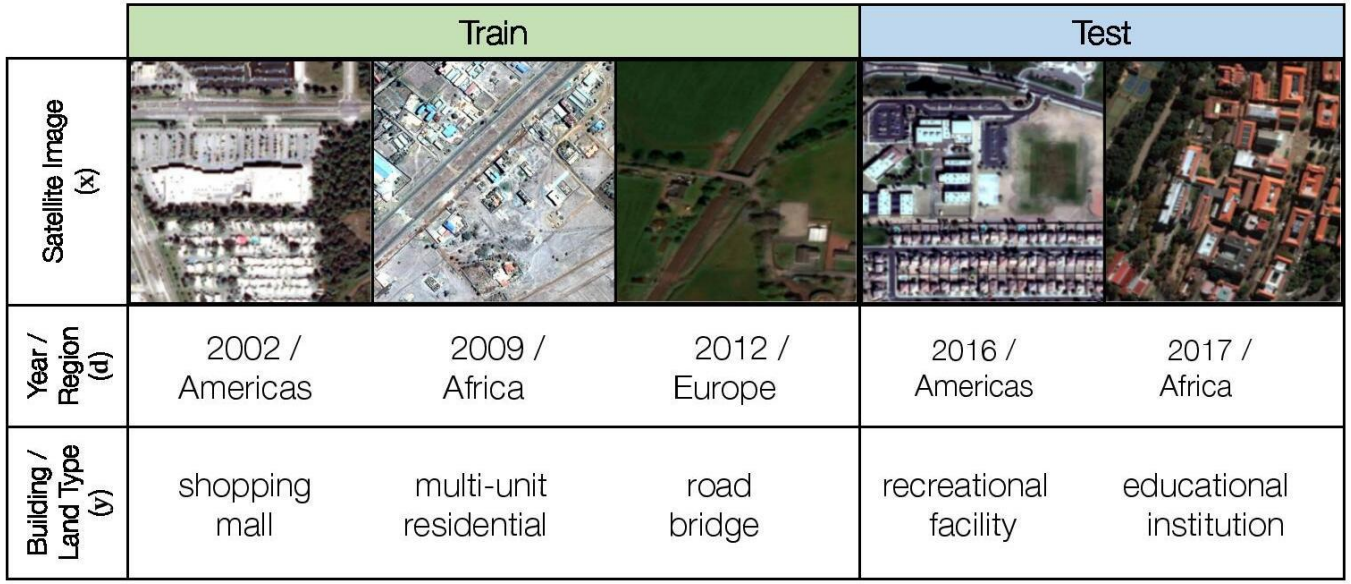
Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? ICML.

Yadav, C., & Bottou, L. (2019). Cold Case: The Lost MNIST Digits. NIPS.

Engstrom, L et al.. (2020). Identifying statistical bias in dataset replication. ICML.

« Distribution shift »

Dataset	In-dist	Out-of-dist	Gap
iWildCam2020-wilds	47.0 (1.4)	31.0 (1.3)	16.0
Camelyon17-wilds	93.2 (5.2)	70.3 (6.4)	22.9
RxRx1-wilds	39.8 (0.2)	29.9 (0.4)	9.9
OGB-MolPCBA	34.4 (0.9)	27.2 (0.3)	7.2
GlobalWheat-wilds	63.3 (1.7)	49.6 (1.9)	13.7
CivilComments-wilds	92.2 (0.1)	56.0 (3.6)	36.2
FMoW-wilds	48.6 (0.9)	32.3 (1.3)	16.3
PovertyMap-wilds	0.60 (0.06)	0.45 (0.06)	0.15
Amazon-wilds	71.9 (0.1)	53.8 (0.8)	18.1
Py150-wilds	75.4 (0.4)	67.9 (0.1)	7.5



Les données de train et de test ne sont pas issues de la même distribution

L'écart de performance peut être important

Koh, P. W., Sagawa, S., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., ... & Liang, P. (2021, July). Wilds: A benchmark of in-the-wild distribution shifts. In International Conference on Machine Learning (pp. 5637-5664). PMLR.

Echantillonner le DO?

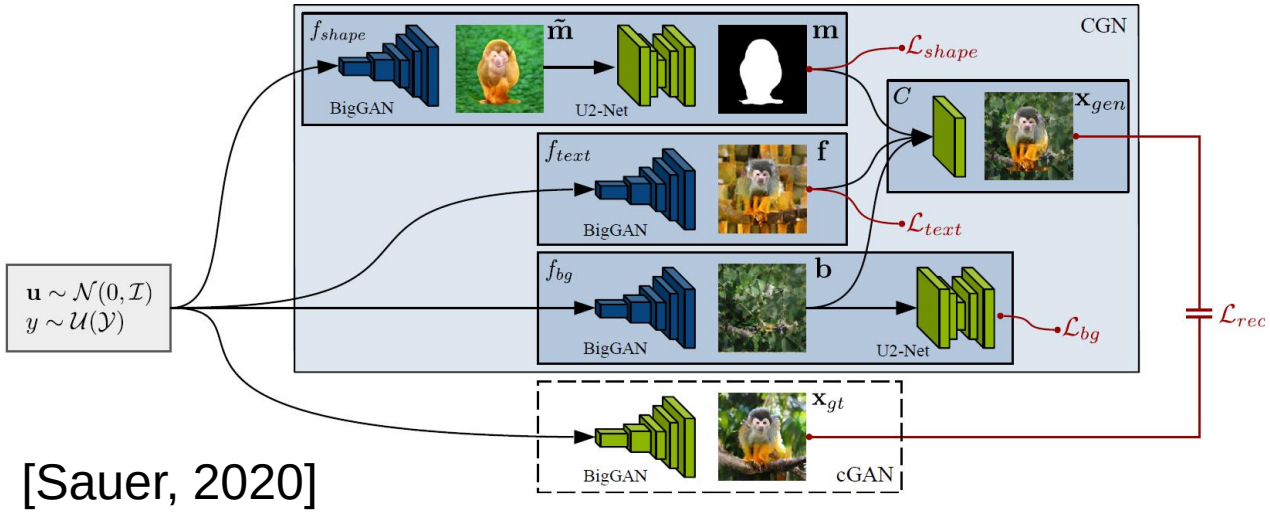
Table 1. Examples of Miles and Years Needed to Demonstrate Autonomous Vehicle Reliability

		Benchmark Failure Rate		
Statistical Question	**How many miles (years[a]) would autonomous vehicles have to be driven...**	**(A) 1.09 fatalities per 100 million miles?**	**(B) 77 reported injuries per 100 million miles?**	**(C) 190 reported crashes per 100 million miles?**
	(1) without failure to demonstrate with 95% confidence that their failure rate is at most...	275 million miles (12.5 years)	3.9 million miles (2 months)	1.6 million miles (1 month)
	(2) to demonstrate with 95% confidence their failure rate to within 20% of the true rate of...	8.8 billion miles (400 years)	125 million miles (5.7 years)	51 million miles (2.3 years)
	(3) to demonstrate with 95% confidence and 80% power that their failure rate is 20% better than the human driver failure rate of...	11 billion miles (500 years)	161 million miles (7.3 years)	65 million miles (3 years)

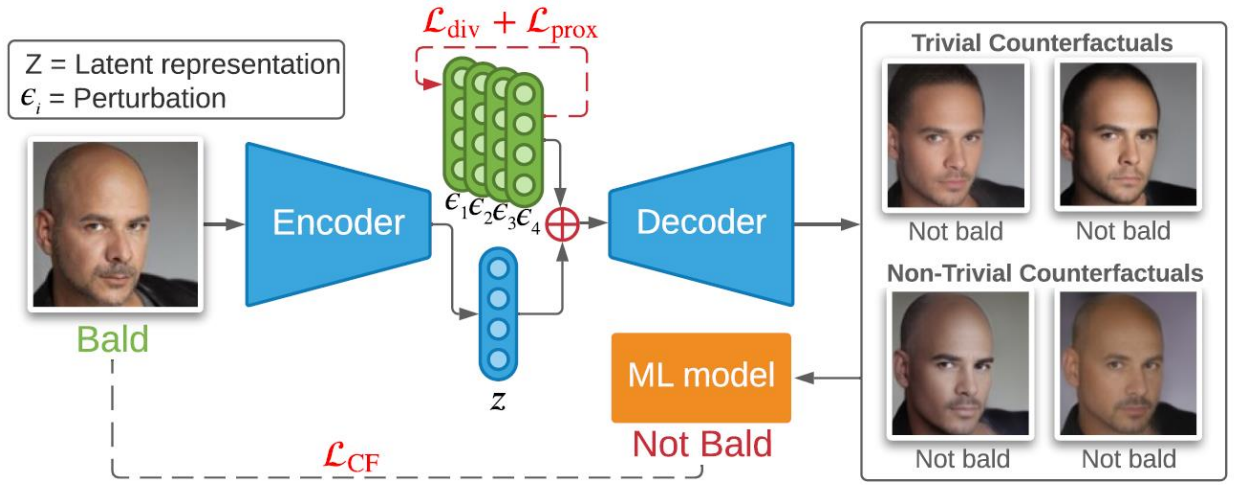
[a] We assess the time it would take to compete the requisite miles with a fleet of 100 autonomous vehicles (larger than any known existing fleet) driving 24 hours a day, 365 days a year, at an average speed of 25 miles per hour.

N. Kalra and S. M. Paddock, “Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?” Transportation Research Part A: Policy and Practice 94, pp. 182–193, 2016.

Identifier/générer des exemples limite



[Sauer, 2020]



[Rodriguez, 2021]

Counterfactual

- Exemple « proche » d'une référence mais produisant un comportement différent du prédicteur.
- Utilisé pour expliquer une prédiction (« parce que ce n'est pas ça… »)

Corner case

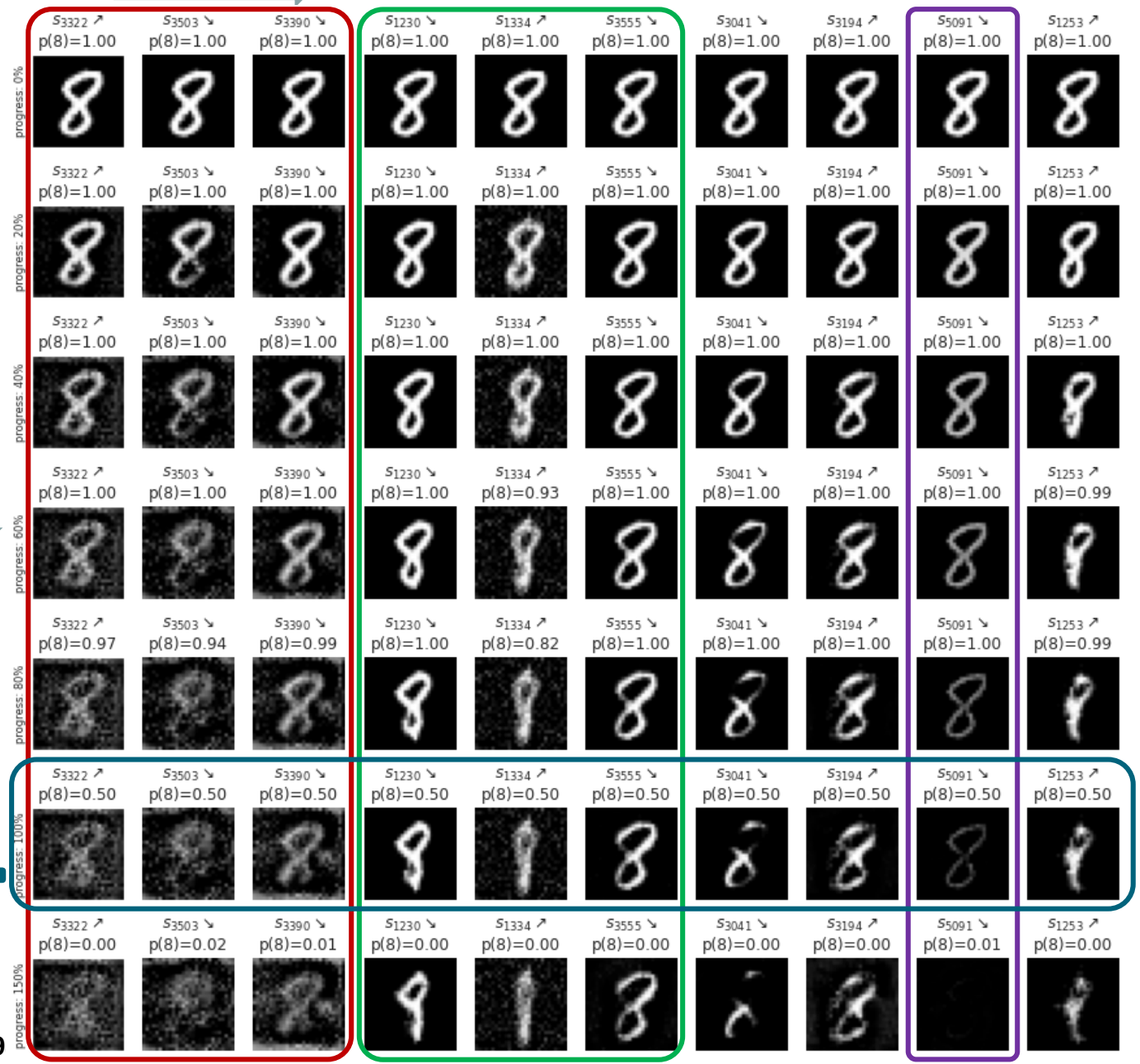
- Exemple combinant des caractéristiques ayant un impact négatif sur la prédiction
- Utilisé pour repérer les limites d'un prédicteur

Analyse du comportement d'un classifieur par modèle génératif

Attribute most impacting classifier's behavior

Editing strength

« Corner cases »



noise,
shape,
contrast

Le Coz, A., Herbin, S., & Adjed, F. (2022, July). Leveraging generative models to characterize the failure conditior of image classifiers. In The IJCA ECAI-22 Workshop on Artificial Intelligence Safety (AISafety 2022).

ONERA
THE FRENCH AEROSPACE LAB

S

Qu'attendre des modèles de fondation en IA?

Février 2024

Les attentes envers les modèles de fondation en IA sont nombreuses et variées. Voici quelques exemples :

Amélioration de la performance des tâches d'IA:

- **Meilleure précision et efficacité** dans des tâches comme la traduction automatique, la génération de texte, la reconnaissance d'image et la réponse aux questions.
- **Capacité à apprendre et à s'adapter** à de nouveaux contextes et à de nouvelles données plus rapidement.
- **Réduction du besoin de données d'apprentissage** pour chaque tâche spécifique.

Démocratisation de l'IA:

- **Rendre l'IA plus accessible** aux entreprises et aux organisations qui n'ont pas les ressources nécessaires pour développer leurs propres modèles.
- **Permettre aux développeurs de créer des applications d'IA plus facilement** et plus rapidement.

S

Qu'attendre des modèles de fondation en IA?

Encourager l'innovation dans le domaine de l'IA.

- **Résolution de problèmes complexes:**
- **Aider à résoudre des problèmes** dans des domaines comme la santé, la science et l'éducation.
- **Détecter des modèles et des insights** dans de grandes quantités de données.
- **Automatiser des tâches** qui sont actuellement effectuées par des humains.

Contribution à la société:

- **Améliorer la qualité de vie** des людей.
- **Contribuer à la croissance économique**.
- **Promouvoir l'équité et l'inclusion** dans la société.

Il est important de noter que les modèles de fondation en IA ne sont pas sans risques. Certains des défis à relever incluent :

- **Biais dans les données** utilisées pour entraîner les modèles.
- **Manque de transparence** dans le fonctionnement des modèles.
- **Risques de cybersécurité** liés à l'utilisation des modèles.

S

Qu'attendre des modèles de fondation en IA?

En conclusion, les modèles de fondation en IA ont le potentiel de révolutionner de nombreux aspects de notre vie. Il est important de continuer à investir dans la recherche et le développement de ces modèles tout en s'attaquant aux défis qu'ils présentent.

En plus des attentes générales ci-dessus, voici quelques attentes spécifiques que j'ai envers les modèles de fondation en IA :

- **Meilleure compréhension du langage naturel** : Je souhaite que les modèles de fondation en IA soient capables de comprendre et de générer le langage naturel de manière plus humaine. Cela permettrait d'améliorer la communication entre les humains et les machines et de créer des applications d'IA plus intuitives et plus faciles à utiliser.
- **Capacité à raisonner et à prendre des décisions** : Je souhaite que les modèles de fondation en IA soient capables de raisonner et de prendre des décisions de manière plus logique et plus objective. Cela permettrait d'utiliser l'IA pour résoudre des problèmes complexes et pour automatiser des tâches qui nécessitent actuellement un jugement humain.
- **Capacité à apprendre et à s'adapter en permanence** : Je souhaite que les modèles de fondation en IA soient capables d'apprendre et de s'adapter en permanence à de nouvelles informations et à de nouveaux contextes. Cela permettrait de créer des systèmes d'IA plus intelligents et plus flexibles qui peuvent s'adapter à un monde en constante évolution.

Je suis convaincu que les modèles de fondation en IA ont le potentiel de transformer notre monde en mieux. J'ai hâte de voir comment ces modèles évolueront dans les années à venir et comment ils seront utilisés pour améliorer nos vies.

Février 2025

Qu'attendre des modèles de fondation en IA?

Les modèles de fondation en IA sont de grands modèles d'apprentissage automatique qui sont entraînés sur de grandes quantités de données. Ils sont capables d'effectuer un large éventail de tâches, telles que la traduction de langues, la rédaction de différents types de contenus créatifs et la réponse à vos questions de manière informative.

Voici quelques-unes des choses que vous pouvez attendre des modèles de fondation en IA :

- **Capacité à effectuer un large éventail de tâches.** Les modèles de fondation peuvent être utilisés pour effectuer un large éventail de tâches, ce qui en fait un outil précieux pour les entreprises et les organisations.
- **Amélioration continue.** Les modèles de fondation sont en constante amélioration, ce qui signifie qu'ils deviendront encore plus puissants et polyvalents à l'avenir.
- **Disponibilité accrue.** Les modèles de fondation sont de plus en plus disponibles, ce qui les rend plus accessibles aux entreprises et aux organisations de toutes tailles.

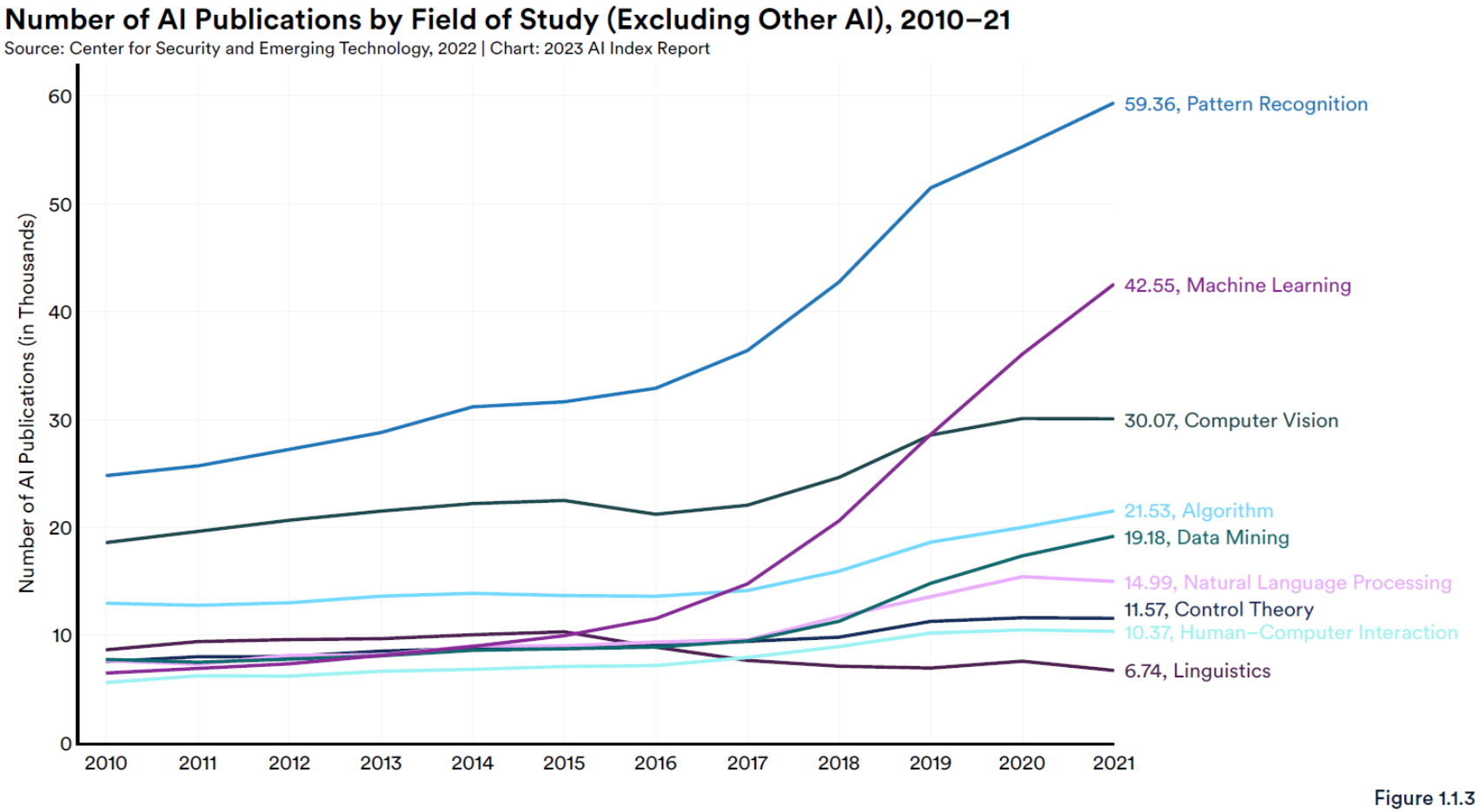
Voici quelques-unes des applications potentielles des modèles de fondation en IA :

- **Service client.** Les modèles de fondation peuvent être utilisés pour fournir un service client personnalisé et efficace.
- **Marketing.** Les modèles de fondation peuvent être utilisés pour créer des campagnes de marketing ciblées et efficaces.
- **Développement de produits.** Les modèles de fondation peuvent être utilisés pour développer de nouveaux produits et services innovants.

Si vous souhaitez en savoir plus sur les modèles de fondation en IA, veuillez consulter les ressources suivantes :

- Qu'est-ce qu'un modèle de base en IA ?
- Que sont les modèles de fondation ?
- Modèle de fondation, une notion fondamentale en intelligence artificielle

ML: un Domaine scientifique en effervescence



Stanford AI Index Report 2023

AI Index 2023

Top Ten Takeaways

1 Industry races ahead of academia.
Until 2014, most significant machine learning models were released by academia. Since then, industry has taken over. In 2022, there were 32 significant industry-produced machine learning models compared to just three produced by academia. Building state-of-the-art AI systems increasingly requires large amounts of data, computer power, and money—resources that industry actors inherently possess in greater amounts compared to nonprofits and academia.

2 Performance saturation on traditional benchmarks.
AI continued to post state-of-the-art results, but year-over-year improvement on many benchmarks continues to be marginal. Moreover, the speed at which benchmark saturation is being reached is increasing. However, new, more comprehensive benchmarking suites such as BIG-bench and HELM are being released.

3 AI is both helping and harming the environment.
New research suggests that AI systems can have serious environmental impacts. According to Luccioni et al., 2022, BLOOM's training run emitted 25 times more carbon than a single air traveler on a one-way trip from New York to San Francisco. Still, new reinforcement learning models like BCOOLER show that AI systems can be used to optimize energy usage.

4 The world's best new scientist ... AI?
AI models are starting to rapidly accelerate scientific progress and in 2022 were used to aid hydrogen fusion, improve the efficiency of matrix manipulation, and generate new antibodies.

5 The number of incidents concerning the misuse of AI is rapidly rising.
According to the AIAAIC database, which tracks incidents related to the ethical misuse of AI, the number of AI incidents and controversies has increased 26 times since 2012. Some notable incidents in 2022 included a deepfake video of Ukrainian President Volodymyr Zelenskyy surrendering and U.S. prisons using call-monitoring technology on their inmates. This growth is evidence of both greater use of AI technologies and awareness of misuse possibilities.

6 The demand for AI-related professional skills is increasing across virtually every American industrial sector.
Across every sector in the United States for which there is data (with the exception of agriculture, forestry, fishing, and hunting), the number of AI-related job postings has increased on average from 1.7% in 2021 to 1.9% in 2022. Employers in the United States are increasingly looking for workers with AI-related skills.

Top Ten Takeaways (cont'd)

7 For the first time in the last decade, year-over-year private investment in AI decreased.
Global AI private investment was $91.9 billion in 2022, which represented a 26.7% decrease since 2021. The total number of AI-related funding events as well as the number of newly funded AI companies likewise decreased. Still, during the last decade as a whole, AI investment has significantly increased. In 2022 the amount of private investment in AI was 18 times greater than it was in 2013.

8 While the proportion of companies adopting AI has plateaued, the companies that have adopted AI continue to pull ahead.
The proportion of companies adopting AI in 2022 has more than doubled since 2017, though it has plateaued in recent years between 50% and 60%, according to the results of McKinsey's annual research survey. Organizations that have adopted AI report realizing meaningful cost decreases and revenue increases.

9 Policymaker interest in AI is on the rise.
An AI Index analysis of the legislative records of 127 countries shows that the number of bills containing "artificial intelligence" that were passed into law grew from just 1 in 2016 to 37 in 2022. An analysis of the parliamentary records on AI in 81 countries likewise shows that mentions of AI in global legislative proceedings have increased nearly 6.5 times since 2016.

10 Chinese citizens are among those who feel the most positively about AI products and services. Americans ... not so much.
In a 2022 IPSOS survey, 78% of Chinese respondents (the highest proportion of surveyed countries) agreed with the statement that products and services using AI have more benefits than drawbacks. After Chinese respondents, those from Saudi Arabia (76%) and India (71%) felt the most positive about AI products. Only 35% of sampled Americans (among the lowest of surveyed countries) agreed that products and services using AI had more benefits than drawbacks.

https://aiindex.stanford.edu/

Biblio sur « Foundation Models »

Fondation models

- https://github.com/awaisrauf/Awesome-CV-Foundational-Models
- https://github.com/uncbiag/Awesome-Foundation-Models
- Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., & Khan, F. S. (2023). *Foundational Models Defining a New Era in Vision : A Survey and Outlook* (arXiv:2307.13721). arXiv. https://doi.org/10.48550/arXiv.2307.13721
- Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V., & Torr, P. (2023). *A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models* (arXiv:2307.12980). arXiv. https://doi.org/10.48550/arXiv.2307.12980
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258
- Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., & Gao, J. (2022). *Vision-Language Pre-training : Basics, Recent Advances, and Future Trends* (arXiv:2210.09263). arXiv. https://doi.org/10.48550/arXiv.2210.09263
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., & Girshick, R. (2023). *Segment Anything*. 4015-4026. https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html
- Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., & Yuan, L. (2023). *Florence-2 : Advancing a Unified Representation for a Variety of Vision Tasks* (arXiv:2311.06242). arXiv. https://doi.org/10.48550/arXiv.2311.06242
- Zhang, J., Huang, J., Jin, S., & Lu, S. (2023). *Vision-Language Models for Vision Tasks : A Survey* (arXiv:2304.00685). arXiv. https://doi.org/10.48550/arXiv.2304.00685

Few shot learning

- Parnami, A., & Lee, M. (2022). Learning from Few Examples : A Summary of Approaches to Few-Shot Learning (arXiv:2203.04291). arXiv. https://doi.org/10.48550/arXiv.2203.04291
- Wang, Y., Yao, Q., Kwok, J., & Ni, L. M. (2020). Generalizing from a Few Examples : A Survey on Few-Shot Learning (arXiv:1904.05046). arXiv. https://doi.org/10.48550/arXiv.1904.05046
- Köhler, M., Eisenbach, M., & Gross, H.-M. (2022). Few-Shot Object Detection : A Comprehensive Survey (arXiv:2112.11699). arXiv. https://doi.org/10.48550/arXiv.2112.11699
- San-Emeterio, M. G. (2022). A Survey on Few-Shot Techniques in the Context of Computer Vision Applications Based on Deep Learning. International Conference on Image Analysis and Processing, 14-25.
- Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A Comprehensive Survey of Few-shot Learning : Evolution, Applications, Challenges, and Opportunities. ACM Computing Surveys. https://doi.org/10.1145/3582688