Datalchemy

ECHOS

DE LA RECHERCHE #19

FEVRIER 2025

Causalité et IA ?

Entre fantasmes de raisonnement et réalité scientifique, retour sur le workshop Neurips2024 – CALM dédié aux derniers états de l'art de ce domaine.

TL;DR ?

Cinq mot-clés de ces échos

#Causalité, #Concept, #LLM, #Graphe, #Extraction

Si vous n’avez qu’une minute à consacrer à la lecture maintenant, voici le contenu essentiel en 7 phrases

1. La recherche de causalité en intelligence artificielle est un graal lancé en 2021 pour obtenir de nouveaux outils.

2. Un modèle IA capable d’appréhender la causalité serait beaucoup plus robuste à une nouvelle donnée et plus facilement interprétable.

3. Le domaine du Causal Representation Learning s’attaque à ce sujet avec de nombreux travaux, mais aussi de vraies difficultés, notamment sur la capacité à intervenir sur un phénomène pour modifier son état.

4. L’atelier CALM du NEURIPS 2024 était dédié à ces travaux en lien avec les Large Models. Ce sujet est encore un sujet actif de recherche.

5. Certains poussent pour réduire la complexité de cette entreprise et se concentrer sur l’identification de concepts dans un modèle.

6. D’autres proposent des approches pour utiliser correctement un LLM en génération de graphes causaux, ceci malgré les nombreuses erreurs qui seront rencontrées.

7. On observe déjà des applications de ce domaine, par exemple pour classifier correctement des comptes-rendus d’incidents très courts.

L'IA PEUT-ELLE RECHERCHER OU EXPLOITER DES CAUSALITÉS ? FONDAMENTAUX ET NEURIPS 2024

Causalité et IA : introduction

Deux travaux scientifiques sont à consulter pour l'appréhender. Le premier, *« Anchor regression: Heterogeneous data meet causality »* [1]de Rothenhausler et al, est considéré comme une base incontournable. L'auteur plaide pour remplacer l'inférence statistique (si j'observe dans une pièce que chaque fois qu'une fenêtre est ouverte, la température est plus basse, je peux considérer qu'il y a corrélation entre ces deux informations) par une inférence **causale** *(si j'ouvre la fenêtre, alors, la température va diminuer)*. L'enjeu est donc, face à un ensemble d'observations, de déterminer les facteurs sources qui impliqueront d'autres observations et de qualifier les liens de causalité entre ces éléments. On va le voir, un enjeu est surtout d'obtenir des approches beaucoup plus robustes que la statistique classique.

Mais la publication fondatrice de ce mouvement en intelligence artificielle est *« Toward causal representation learning [2]»* avec en premier auteur Schölkopf et caché derrière notre cher Yoshua Bengio. Cette publication de 2021 peut être vue comme un manifeste pour chercher de nouveaux modèles IA qui puissent justement, par apprentissage, découvrir la structure logique (les liens de causalité) d'un *dataset* représentant un phénomène. L'enjeu va bien au-delà d'une simple satisfaction intellectuelle, car de tels modèles auraient des qualités appréciables :

[1] https://arxiv.org/abs/1801.06229

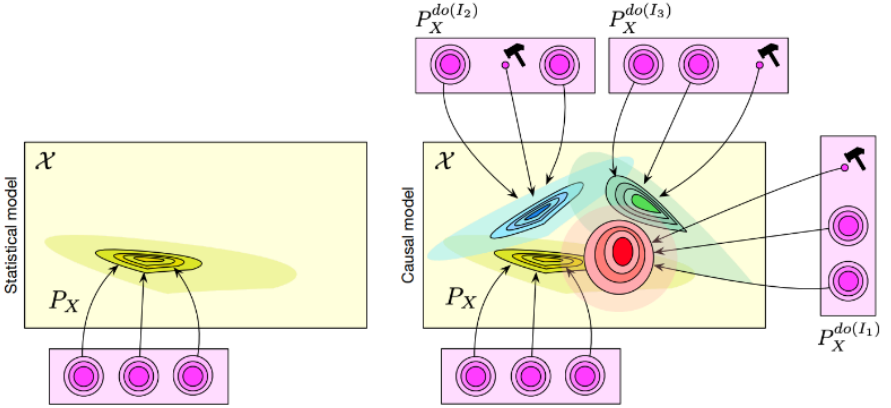
[2] https://arxiv.org/pdf/2102.11107

- **Robustesse** : contrairement à un modèle IA classique, un modèle causal dépendrait beaucoup moins des biais du *dataset* d'entraînement en ceci qu'il isolerait des mécanismes génériques qu'il utiliserait ensuite pour donner une réponse. Nous ne travaillons plus sur des corrélations observées dans la donnée, mais sur la logique qui se cache derrière.

- **Ré-utilisation de modèles** : tout pratiquant *Deep Learning* a vu le meilleur des modèles s'écrouler sur un nouveau *dataset* de travail… Un modèle causal ne souffrirait pas de ces problèmes pour les raisons exposées ci-dessus liée à sa robustesse.

- **Une qualité supérieure** : débarrassé de corrélations, un modèle causal serait beaucoup plus efficace. En *Deep Learning* classique, nous modélisons les dépendances par des probabilités conditionnelles d'observation, par exemple : *« Voir des gens avec des parapluies dans la rue suggère qu'il est en train de pleuvoir ».* Mais cette probabilité conditionnelle ne permet pas d'agir correctement, typiquement : *« Fermer les parapluies dans la rue augmenterait la probabilité qu'il ne pleuve plus »*… Au contraire, un système causal serait débarrassé de ce type d'erreur.

Se glisse ici un léger problème : si nous voulons découvrir les liens de causalité, nous aurons besoin de pouvoir connaître l'impact **d'interventions**. Une intervention vise à modifier une seule des variables fondamentales pour observer les conséquences de cette modification. Nous reparlerons de ce point plus tard, mais le schéma ci-dessous permet d'observer la différence entre un modèle IA classique et un modèle causal :

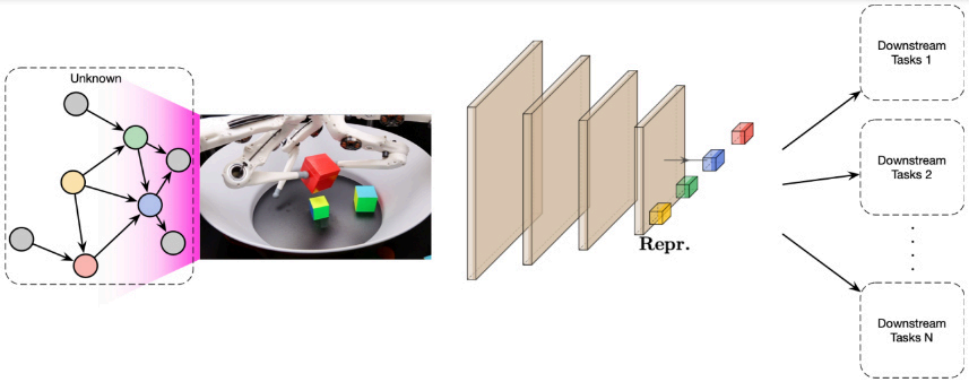


A gauche : un modèle classique. Chaque « rond » est une variable d'intérêt, et le modèle projette l'ensemble de ces variables dans une distribution cible

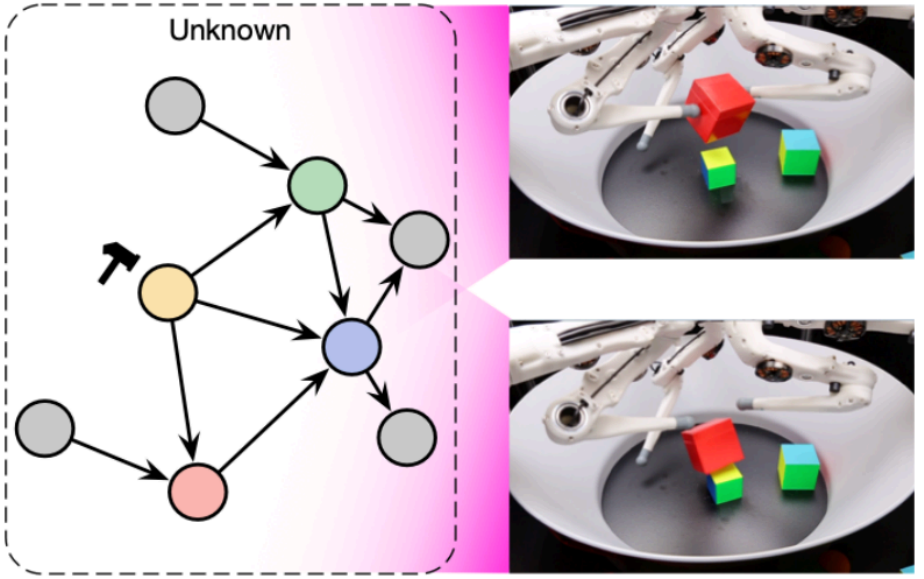
A droite : nous pouvons modéliser l'impact d'une intervention sur chaque variable d'intérêt *(les petits marteaux mignons. Si, ils sont mignons)* sur la distribution cible, et donc estimer comment chacune modifie le résultat.

Un modèle causal présente de nombreux intérêts face à un modèle classique. Rappelons déjà qu'un modèle classique n'est valable que si on l'utilise sur une donnée appartenant à la même distribution que le dataset d'entraînement, ce qui n'est plus le cas ici. Le terrible *Distribution Drift* qui fait trembler les ingénieurs IA devient ici appréhendable : ce n'est pas parce que nos images médicales viennent d'un autre praticien ou d'un autre hôpital que notre modèle s'écroulera. Et surtout, nous pouvons espérer répondre à des questions contrefactuelles, par exemple *« Ce patient aurait-il eu un accident cardiaque s'il avait fait plus d'exercice physique »*. Ce dernier point est très important, en ceci qu'un modèle causal pourrait jouer un rôle de modèle prédictif beaucoup plus efficace, rejoignant la recherche en *World Models* par exemple.

Le schéma ci-dessous représente le problème fondamental auquel nous voulons nous atteler : un bras robot manipule des objets et n'a accès qu'aux pixels de l'image d'une caméra, sans connaître le graphe causal modélisant les interactions entre les objets. S'il est peu probable que nous redécouvrions l'intégralité des relations logiques entre les éléments, un système causal pourrait au moins exposer des variables claires en séparant par exemple, dans ses représentations, les informations de position et d'apparence de chaque objet.

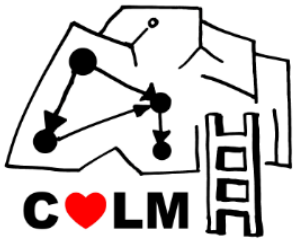


À la page précédente, on observe une intervention modifiant la position d'un des appendices du robot. Un modèle classique n'y verrait qu'une modification de pixels avec comme objectif très lourd d'identifier une corrélation. Un modèle causal serait en capacité d'exposer l'information importante ayant changé.



Prenons un peu de recul : nous sommes en train de parler de raisonnement d'un modèle IA, finalement. Et nous allons voir que les dernières sorties d'*OpenAI* ou *DeepSeek* ne peuvent prétendre réellement à une forme de raisonnement de par leur nature de modèles statistiques. Faut-il alors se précipiter sur ces modèles ? Pas encore. A l'époque, cette publication ne proposait pas de solution au problème, et ne faisait « que » lancer un courant de recherche. Modéliser l'impact d'une intervention est extrêmement difficile, et les outils du *Deep Learning* relèvent plus d'un « *bourrinage* » pour entraîner les plus gros modèles possibles que d'une approche fine de contrôle.

Cela n'a pas empêché de nombreux chercheurs d'avancer sur ce sujet. Et nous vous proposons donc de faire une revue d'un workshop récent du NeuRIPS 2024, dédié à la causalité appliquée au *Large Models*. Ce workshop nous donne une vision passionnante de l'état actuel de la recherche en causalité et des opportunités qui y sont reliées.



C❤LM: First Workshop on

Causality and Large Models

December 14 @ NeurIPS 2024 in East Hall C

Chercher des concepts plutôt que des causes

Parmi les publications à l'honneur de ce workshop se trouve un travail très intéressant issu notamment des laboratoires de *Meta*. « From Causal to Concept-Based Representation Learning[1] » de Rajendran et al, reprend les objectifs fondamentaux de l'apprentissage causal et nous offre déjà une vue accélérée des nombreux travaux qui se sont succédé depuis la publication fondatrice jusqu'à aujourd'hui. Trois ans se sont écoulés et force est de reconnaître que si la représentation causale passionne toujours autant les foules (de chercheurs), les avancées sont faibles.

Une question fondamentale, en effet, est de savoir face à un problème si un graphe causal peut exister ! Dans de nombreux cas, même un expert humain ne pourra pas affirmer avec certitude qu'il y a un lien logique entre deux affirmations, voire même s'il existe un lien logique pouvant justifier de certaines observations. Les chercheurs ont donc essayé de simplifier le problème au maximum pour au moins établir dans quels cas de figure la recherche d'un graphe de causalité était simplement possible. Une limite importante a été établie pour décider si la recherche de causalité était possible : le besoin de disposer de plusieurs *datasets* représentant le problème, ces *datasets* se distinguant par des **interventions** (on les retrouve !). Or, si les *datasets* classiques sont déjà peu courants, de tels datasets modélisant l'impact d'une intervention sur chaque variable sont extrêmement rares, voir, impossible à générer (si un lecteur est capable par exemple de modifier la gravité en temps réel, merci de nous envoyer un CV de toute urgence)…

[1] https://openreview.net/forum?id=FcVnIBYbkW

Dans cette publication, les auteurs plaident pour une approche moins ambitieuse mais plus raisonnée. Nous ne cherchons plus les causes originelles des observations, mais des projections de ces causes, projections que nous pouvons mettre en lien avec des **concepts** compréhensibles par des humains. L'idée est que pour chaque concept global, chaque déclinaison de ce concept existe dans un sous-espace affine :

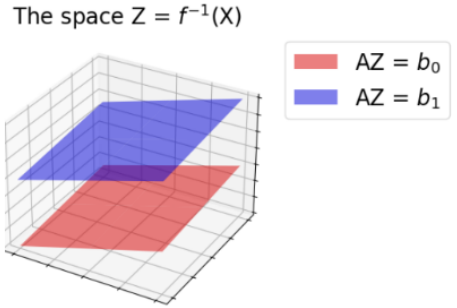


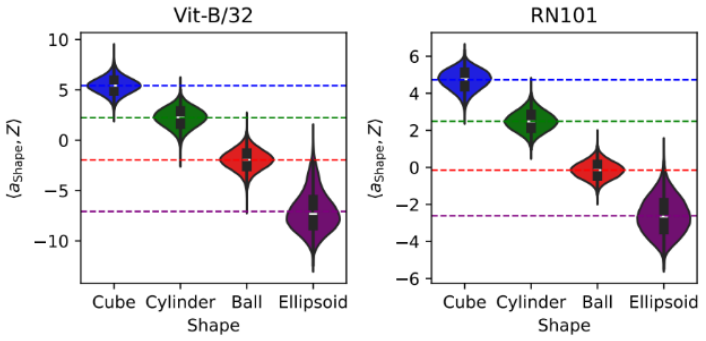
Figure 1: Concepts live in affine subspaces. The two subspaces in the figure correspond to the same concept but of different valuations.

Nous remarquons que nous parlons maintenant de concepts, plus de causalité. Nous rejoignons donc le champ de recherche de l'interprétabilité de l'intelligence artificielle. Ce déplacement vise à ne plus dépendre **d'interventions** pour découvrir les causalités, mais de **conditionnement** pour découvrir les concepts. Là où une intervention suppose d'agir sur le phénomène pour observer, le conditionnement vise plus à grouper le *dataset* en sous-espaces où un concept est identifié avec des variances.

Nous perdons l'aspect de la découverte logique, mais gagnons en efficacité. Et c'est là la plaidoirie principale des auteurs : les approches de *Causal Representation Learning* sont condamnées, et mieux vaudrait déjà travailler sur la modélisation de concepts dans l'apprentissage d'un modèle.

Un exemple rendra tout cela probablement plus digeste. Les auteurs ont travaillé avec le modèle CLIP qui, pour rappel, projette dans un même espace une image et du langage naturel. Ils ont ensuite mis en entrée des formes géométriques simples,

différentes, avec des sous-variations en couleur, position, déformation. Et ils ont ensuite cherché à identifier, par leur méthode, un unique vecteur lié au concept de forme. On observe ci-dessous que la projection sur ce vecteur concept donne des sous-espaces bien délimités en fonction de la variation de forme, ce pour deux architectures différentes :



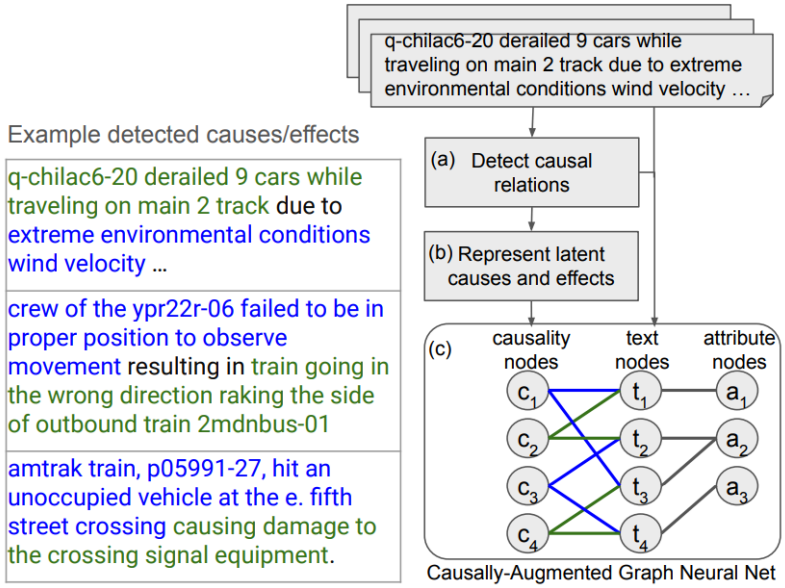
Et ? Et c'est tout. Pas d'autre application supplémentaire, le problème d'une recherche trop récente. Mais, en rejoignant les travaux en interprétabilité, le domaine de l'apprentissage causal devient plus crédible et générique, à nous de surveiller la suite…

Découvrir la causalité pour classifier des textes courts

Sortons de ces travaux très théoriques pour nous intéresser à une autre gemme du Workshop CaLM. « Using Relational and Causality Context for Tasks with Specialized Vocabularies that are Challenging for LLMs[1] », de Nakashini et al (Toyota research), s'attaque à un problème très classique et toujours assez difficile : la classification de comptes-rendus très courts.

L'enjeu ici est de qualifier des comptes-rendus d'incidents qui restent très courts, écrits rapidement par des opérationnels. Un texte court contient très peu d'informations exploitables, empêchant les approches classiques de donner des résultats pertinents (et encourageant fortement l'apparition d'hallucinations chez un de nos bons vieux LLMs). L'idée ici est de générer un graphe de causalité à partir de ces rapports. Ce graphe permet ensuite de classifier les textes en entrée :

[1] https://openreview.net/forum?id=pAWrQEdlxq



Dans le schéma ci-dessus, nous avons différents rapports textuels d'incidents ferroviaires. Les couleurs (vert et bleu) sont issues d'appels au LLM pour distinguer la cause de l'effet. Ces informations sont ensuite agrégées afin de représenter les causalités partagées par différents fragments textuels. Ce graphe permet ensuite d'apprendre des classes utilisables pour ces rapports. Ci-dessous quelques exemples de résultats :

Text reports	Predicted classes
Employees' skill gaps hinder team performance.	Lack of training
Micromanagement is making the workload worse.	Workload & Stress

Figure 2: Example predicted results.

Ici, l'approche fonctionne car nous ne cherchons pas un graphe causal « absolu », mais plus à séparer dans un domaine métier très réduit causes et conséquences. L'agrégation permet de limiter l'effet de mauvaises prédictions du LLM (qui arriveront évidemment). D'une manière pratique, classifier des textes courts de ce type présente de nombreux cas d'application dans différents domaines métier où l'on veut exploiter un retour textuel limité.

Chercher l'ordre causal plus que le graphe causal

Dernière publication phare du workshop, « Causal Order: The Key to Leveraging Imperfect Experts in Causal Inference[1] » de Vashishtha et al, est aussi à l'honneur de l'ICLR 2025 et laissera sa marque. Revenons à notre objectif théorique (et quelque peu désespéré) : découvrir un graphe causal à partir d'observations dans une donnée. Même avec les meilleurs LLMs du monde, l'entreprise est vite insurmontable. Un exemple d'application testée par les chercheurs a été, face à un ensemble d'observations (A,B,C,D…) de soumettre à un LLM chaque couple (AB, AC, AD… BC, BD, etc.) en demandant au LLM si une des variables pouvait causer l'autre variable, et ensuite d'agréger les résultats.

[1] https://openreview.net/forum?id=3fzCBL6ar7

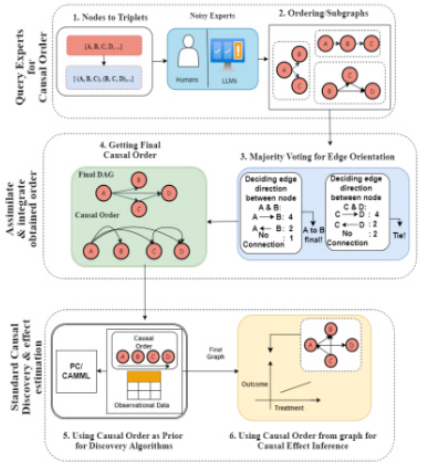
Le problème de cette approche naïve est qu'il sera impossible, en travaillant de la sorte, de distinguer si une variable cause **directement** ou **indirectement** une autre variable. Par exemple, si nous avons le graphe causal *« Cancer du poumon », « Consultation médicale »* et *« Rayons X positifs »*, mais que nous n'interrogeons le LLM que sur *« Cancer du poumon »* et *« Rayons X positifs »*, ce modèle affirmera que le premier cause le second, mais cela ne nous permettra pas de construire un graphe global de causalité acceptable ! Or, au-delà de la réponse unitaire du LLM, c'est ce graphe de causalité qui nous intéresse particulièrement.

Les auteurs proposent donc une approche différente sur deux points, le premier fondamental, le second plus pratique :

- Nous ne voulons pas demander au LLM d'extraire un graphe de causalité, mais **un ordre de causalité**. On parle d'ordre de causalité quand on distingue entre deux observations si l'une est parente de l'autre en termes de causalité. Dans l'exemple ci-dessus, *« Cancer du poumon »* est parent de *« Rayons X positifs »*. L'information d'ordre est certes moins directe, mais les auteurs démontrent qu'elle est beaucoup plus efficace pour espérer converger vers un graph causal acceptable. En effet, il devient plus simple d'agréger toutes les réponses du *LLM* entre elles avec cette information.

- Nous ne voulons pas interroger notre *LLM* sur des couples d'informations, mais sur des triplets, en demandant au *LLM* de donner l'ordre causal sur les éléments donnés.v

Les auteurs démontrent que cette approche est généralement meilleure, que l'on utilise des LLMs ou des experts humains pour obtenir l'information d'ordre causal. Le schéma ci-dessous récapitule l'approche :



De haut en bas :

1 & 2 : Nous générons des triplets à partir de l'ensemble des informations que nous voulons traiter, et demandons, soit à des *LLMs*, soit à des humains, de donner l'ordre causal pour les éléments de chaque triplet. Nous utilisons ensuite ces informations pour générer des graphes sur chaque triplet.

3 & 4 : En analysant l'ensemble des résultats, nous générons un graphe final de causalité

5 : Il est ensuite possible d'utiliser ce graphe comme une condition pour les algorithme de recherche/découverte en orientant / qualifiant les hypothèses

6 : Il est aussi possible d'utiliser ce graphe à l'inférence pour identifier des causes dans un cas spécifique.

Les auteurs proposent aussi de nouvelles métriques pour qualifier un graphe. Ils comparent ensuite différentes approches, notamment (ci-dessous) entre les graphes générés par couples d'éléments (avec plusieurs variantes) et ceux générés par triplets :

Dataset	Metric	Pairwise (CoT)	Triplet
Using LLM			
Earthquake	D_{top}	**0**	**0**
	SHD	**4**	**4**
	Cycles	**0**	**0**
	IN/TN	**0/5**	**0/5**
Asia-M	D_{top}	-	**1**
	SHD	13	**11**
	Cycles	1	**0**
	IN/TN	**0/7**	**0/7**
Child	D_{top}	-	**1**
	SHD	138	**28**
	Cycles	»500	**0**
	IN/TN	**0/20**	10/20
Using Human Annotators			
Earthquake	D_{top}	**0**	**0**
	SHD	4.67	**1.67**
	Cycles	**0**	**0**
	IN	**0**	0.33
Asia-M	D_{top}	-	**1.33**
	SHD	11.67	**11.33**
	Cycles	3	**0**
	IN	**0**	**0**

Les métriques sont le *Dtop* (une métrique des auteurs sur la qualité topologique du graphe), le SHD qui était la métrique classique pour estimer la qualité d'un graphe (plus c'est bas mieux c'est), le nombre de cycles dans le graphe (on en veut le moins possible si on travaille sur de la causalité), et v*IN/TN* qui donne le nombre de nœuds isolés dans le graphe et le nombre total de nœuds (on veut le moins de nœuds isolés possibles aussi)

Conclusions

Que retenir de ce petit parcours ?

Déjà, nous pouvons observer que le but « ultime », celui de générer un graphe causal via un apprentissage de type Deep Learning relève actuellement de l'impossible. Cela pourrait surprendre quand on croise la valse des communiqués de presse sur le « raisonnement » des IAs, mais nous constatons juste le grand écart entre la réalité de la recherche académique et la communication débordante autour de ces outils.

Néanmoins, force est de reconnaître que les choses avancent. Que ce soit pour déplacer le combat vers un domaine (un peu) moins complexe comme l'identification de concepts et l'interprétabilité, où en développant des méthodes plus intelligentes comme l'utilisation d'un ordre causal, nous avons déjà des outils intéressants pour travailler l'information et en extraire une structure pertinente. Evidemment, ces systèmes ne sont pas parfaits et doivent être considérés comme des assistances à experts humains avec leurs avantages et leurs limites 😊

Datalchemy

contact@datalchemy.net

Images extraites des articles respectifs ou générées par IA