



Exploration de l'explicabilité de l'intelligence
artificielle et application au traitement
automatique de la langue naturelle
Soutenance du Projet Immersion

Marine DELVALLEZ

Master 2 Informatique GPEx Minerve

7 Janvier 2026

Sommaire

- ① Intelligence Artificielle et Modèles de Langue
 - Définitions
 - Le modèle Transformer
 - Les modèles boîte-noire
- ② L'eXplicabilité de l'Intelligence Artificielle
 - Définitions et enjeux
 - Taxonomie des méthodes d'explication
- ③ Évaluer les méthodes d'explications
 - Les attendus
 - Les approches
 - Retour sur les enjeux de l'évaluation des méthodes d'explication
- ④ Conclusion

Sommaire

- 1 Intelligence Artificielle et Modèles de Langue
 - Définitions
 - Le modèle Transformer
 - Les modèles boîte-noire
- 2 L'eXplicabilité de l'Intelligence Artificielle
 - Définitions et enjeux
 - Taxonomie des méthodes d'explication
- 3 Évaluer les méthodes d'explications
 - Les attendus
 - Les approches
 - Retour sur les enjeux de l'évaluation des méthodes d'explication
- 4 Conclusion

Intelligence Artificielle

Intelligence Artificielle

Comportement humain	Comportement rationnel
Raisonnement humain	Raisonnement rationnel

Intelligence Artificielle

Intelligence Artificielle

Comportement humain	Comportement rationnel
Raisonnement humain	Raisonnement rationnel

Données : Images, Sons, Vidéo, Texte, Tables Attribut-Valeur, Entiers, Classes,...

Intelligence Artificielle

Intelligence Artificielle

Comportement humain	Comportement rationnel
Raisonnement humain	Raisonnement rationnel

Données : Images, Sons, Vidéo, Texte, Tables Attribut-Valeur, Entiers, Classes,...

Tâche

Nature de donnée d'entrée + Nature de sortie + spécification

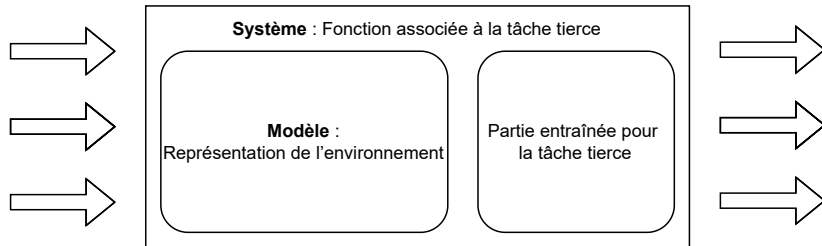
Classification : Texte $\rightarrow$ Classe, Condition

Sous-titrage d'images : Image $\rightarrow$ Texte, Décrire

Traduction : Texte $\rightarrow$ Texte, Changer la langue

...

Grands Modèles de Langue

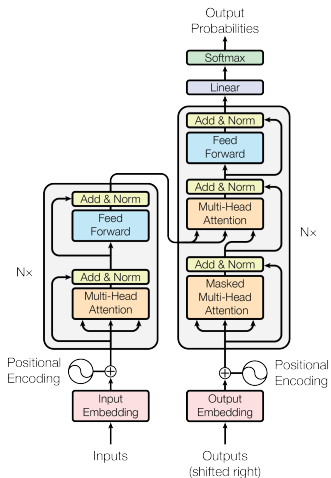


Exemple : BERT pour la classification de texte

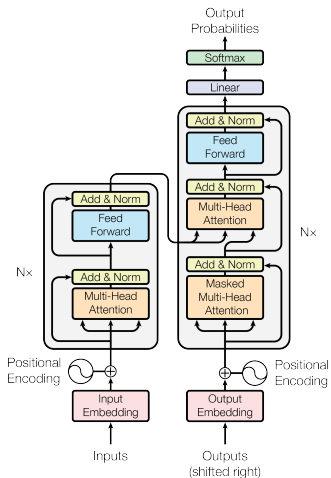
Architecture du Transformer [Vaswani et al., 2017]

Composants Techniques :

- Mécanisme d'attention multi-tête
- Réseau de neurones
- Connexion résiduelle & Normalisation
- Embedding et Encoding positionnel



Architecture du Transformer [Vaswani et al., 2017]



Composants Techniques :

- Mécanisme d'attention multi-tête
- Réseau de neurones
- Connexion résiduelle & Normalisation
- Embedding et Encoding positionnel

Modèles dérivés du Transformer :

- BERT [Devlin et al., 2019]
- GPT [Radford et al., 2018]
- T5 [Raffel et al., 2020]
- ...

Mécanisme d'attention

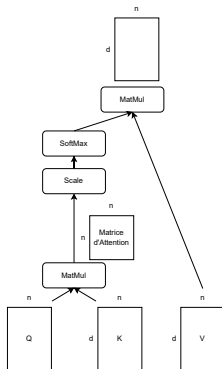


Figure – Scaled Dot Product Attention [Vaswani et al., 2017]

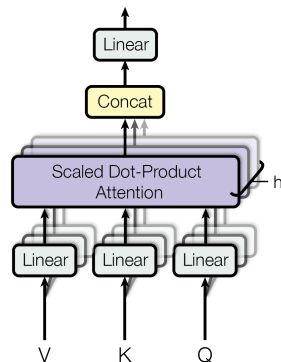


Figure – Mécanisme d'attention multi-tête [Vaswani et al., 2017]

Modèle Boite-noire

Modèle Boite noire

Modèles pour lesquels les processus de décision ne sont pas transparents pour l'humain

Problème : Perte de confiance pour certains usages

Sommaire

- 1 Intelligence Artificielle et Modèles de Langue
 - Définitions
 - Le modèle Transformer
 - Les modèles boîte-noire
- 2 L'eXplicabilité de l'Intelligence Artificielle
 - Définitions et enjeux
 - Taxonomie des méthodes d'explication
- 3 Évaluer les méthodes d'explications
 - Les attendus
 - Les approches
 - Retour sur les enjeux de l'évaluation des méthodes d'explication
- 4 Conclusion

Définitions

Explicabilité de l'IA

Rendre humainement compréhensible les raisons pour lesquelles un modèle effectue une certaine prédiction

[Garouani et al., 2024, Bell et al., 2022]

Définitions

Explicabilité de l'IA

Rendre **humainement compréhensible** les raisons pour lesquelles un modèle effectue une certaine prédiction

[Garouani et al., 2024, Bell et al., 2022]

Définitions

Explicabilité de l'IA

Rendre humainement compréhensible les raisons pour lesquelles un **modèle effectue** une certaine prédiction
[Garouani et al., 2024, Bell et al., 2022]

Définitions

Explicabilité de l'IA

Rendre humainement compréhensible les raisons pour lesquelles un modèle effectue une certaine prédiction

[Garouani et al., 2024, Bell et al., 2022]

Comprendre

Être capable d'anticiper et/ou justifier le comportement du modèle

[Bell et al., 2022]

Définitions

Explicabilité de l'IA

Rendre humainement compréhensible les raisons pour lesquelles un modèle effectue une certaine prédiction

[Garouani et al., 2024, Bell et al., 2022]

Comprendre

Être capable d'anticiper et/ou justifier le comportement du modèle

[Bell et al., 2022]

Explication

Tout support rendant le processus de décision d'un modèle humainement intelligible.

Formes d'explication



Formes d'explication

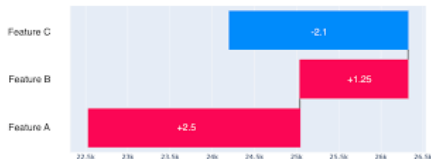


[CLS] there may not be a critic alive who harbors as much affection for shlock monster movies as i do . i delighted in the sneaky - smart entertainment of ron underwood ' s big - underground - worm yarn tremors ; i even giggled at last year ' s critically - savaged big - underwater - snake yarn anaconda . something about these films causes me to lower my inhibitions and return to the saturday afternoons of my youth , spent in the company of ghidrah , the creature from the black lagoon and the blob . deep rising , a big - undersea - serpent yarn , does n ' t quite pass the test . sure enough , all the modern monster movie ingredients are in place ; a conspicuously multi - ethnic / multi - national collection of bait . . . excuse me , characters ; an isolated location , here a derelict cruise ship in the south china sea ; some comic relief ; a few cgi - enhanced gross - outs ; and at least one big explosion . there are too - cheesy - to - be - accidental elements , like a sleazy shipping magnate (anthony heald) who also appears to have a doctorate in marine biology , or a slinky international jewel thief (famke janssen) whose white cotton tank top hides a heart of gold . as it happens , deep rising is noteworthy primarily for the mechanical manner in which it spits out all those ingredients . a terrorist crew , led by squinty - eyed mercenary hanover (wes studi) and piloted by squinty - eyed boat captain finnegan (treat williams) , shows up to loot the cruise ship ; the sea monsters show up to eat the mercenary crew ; a few survivors make it to the closing credits . and up go the lights . it ' s hard to work up much enthusiasm for this sort of joyless film - making , especially when a monster movie should make you laugh every time it makes you scream . here , the laughs are provided almost entirely by kevin j . o ' connor , generally amusing as the crew ' s fraidy - cat mechanic . writer / director stephen sommers seems most concerned with creating a tone of action - horror menace -- something over - populated with gore - drenched skeletons , something where the gunfire and special effects are taken a bit too seriously . deep rising is missing that one unmistakable cue that we ' re expected to have a ridiculous good time , not hide our eyes . case it point , comparing deep rising to its recent cousin anaconda . in deep [SEP]

Formes d'explication



Base: \$22,533, Prediction: \$24,203

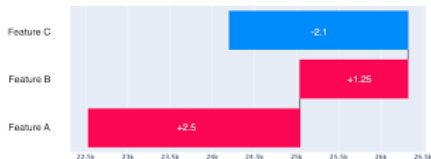


[CLS] there may not be a critic alive who harbors as much affection for shlock monster movies as i do . i delighted in the sneaky - smart entertainment of ron underwood ' s big - underground - worm yarn tremors ; i even giggled at last year ' s critically - savaged big - underwater - snake yarn anaconda . something about these films causes me to lower my inhibitions and return to the saturday afternoons of my youth , spent in the company of ghidrah , the creature from the black lagoon and the blob . deep rising , a big - undersea - serpent yarn , does n ' t quite pass the test . sure enough , all the modern monster movie ingredients are in place ; a conspicuously multi - ethnic / multi - national collection of bait . . . excuse me , characters ; an isolated location , here a derelict cruise ship in the south china sea ; some comic relief ; a few cgi - enhanced gross - outs ; and at least one big explosion . there are too - cheesy - to - be - accidental elements , like a sleazy shipping magnate (anthony heald) who also appears to have a doctorate in marine biology , or a slinky international jewel thief (famke janssen) whose white cotton tank top hides a heart of gold . as it happens , deep rising is noteworthy primarily for the mechanical manner in which it spits out all those ingredients . a terrorist crew , led by squinty - eyed mercenary hanover (wes studi) and piloted by squinty - eyed boat captain finnegan (treat williams) , shows up to loot the cruise ship ; the sea monsters show up to eat the mercenary crew ; a few survivors make it to the closing credits . and up go the lights . it ' s hard to work up much enthusiasm for this sort of joyless film - making , especially when a monster moview should make you laugh every time it makes you scream . here , the laughs are provided almost entirely by kevin j . o ' connor , generally amusing as the crew ' s fraidy - cat mechanic . writer / director stephen sommers seems most concerned with creating a tone of action - horror menace -- something over - populated with gore - drenched skeletons , something where the gunfire and special effects are taken a bit too seriously . deep rising is missing that one unmistakable cue that we ' re expected to have a ridiculous good time , not hide our eyes . case it point , comparing deep rising to its recent cousin anaconda . in deep [SEP]

Formes d'explication



Base: \$22,533, Prediction: \$24,203

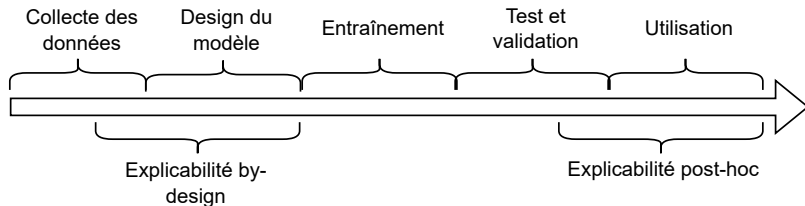


[CLS] there may not be a critic alive who harbors as much affection for shlock monster movies as i do . i delighted in the sneaky - smart entertainment of ron underwood ' s big - underground - worm yarn tremors ; i even giggled at last year ' s critically - savaged big - underwater - snake yarn anaconda . something about these films causes me to lower my inhibitions and return to the saturday afternoons of my youth , spent in the company of ghidrah , the creature from the black lagoon and the blob . deep rising , a big - undersea - serpent yarn , does n ' t quite pass the test . sure enough , all the modern monster movie ingredients are in place ; a conspicuously multi - ethnic / multi - national collection of bait . . . excuse me , characters ; an isolated location , here a derelict cruise ship in the south china sea ; some comic relief ; a few cgi - enhanced gross - outs ; and at least one big explosion . there are too - cheesy - to - be - accidental elements , like a sleazy shipping magnate (anthony heald) who also appears to have a doctorate in marine biology , or a slinky international jewel thief (famke janssen) whose white cotton tank top hides a heart of gold . as it happens , deep rising is noteworthy primarily for the mechanical manner in which it spits out all those ingredients . a terrorist crew , led by squinty - eyed mercenary hanover (wes studi) and piloted by squinty - eyed boat captain finegan (treat williams) , shows up to loot the cruise ship ; the sea monsters show up to eat the mercenary crew ; a few survivors make it to the closing credits . and up go the lights . it ' s hard to work up much enthusiasm for this sort of joyless film - making , especially when a monster moview should make you laugh every time it makes you scream . here , the laughs are provided almost entirely by kevin j . o ' connor , generally amusing as the crew ' s fraidy - cat mechanic . writer / director stephen sommers seems most concerned with creating a tone of action - horror menace -- something over - populated with gore - drenched skeletons , something where the gunfire and special effects are taken a bit too seriously . deep rising is missing that one unmistakable cue that we ' re expected to have a ridiculous good time , not hide our eyes . case it point , comparing deep rising to its recent cousin anaconda . in deep [SEP]

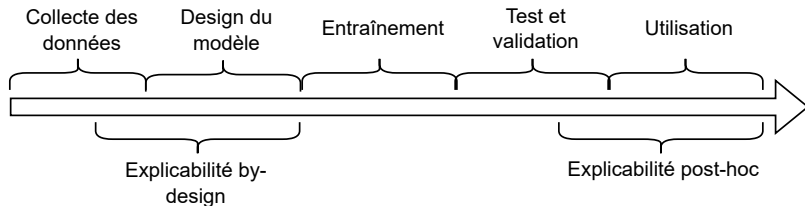
$$C_1 \iff A > B \wedge C \in \mathcal{S}$$

$$C_2 \iff A < B \vee C \notin \mathcal{S}$$

Temps de l'explicabilité



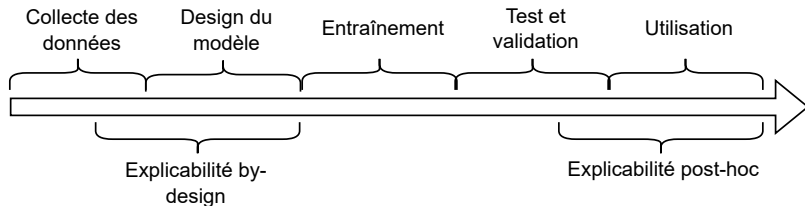
Temps de l'explicabilité



Interprétabilité

Comment ?

Temps de l'explicabilité



Interprétabilité
Comment ?

Explicabilité
Pourquoi ?

Fidélité vs Accessibilité

Des attendus contradictoires avec des conséquences

- sur-simplification $\implies$ confiance induite [Gilpin et al., 2019]
- sous-simplification $\implies$ inutilisable [Mersha et al., 2025]

Fidélité vs Accessibilité

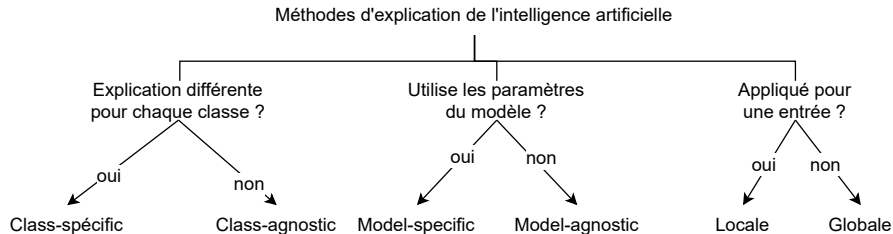
Des attendus contradictoires avec des conséquences

- sur-simplification $\implies$ confiance induite [Gilpin et al., 2019]
- sous-simplification $\implies$ inutilisable [Mersha et al., 2025]

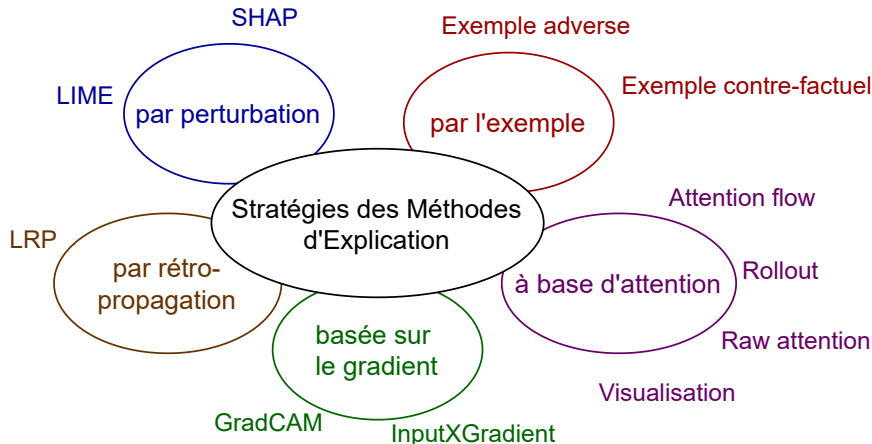
Et un rapport avec le grand public compliqué

- Les explications créent une surcharge d'informations [Bell et al., 2022]
- Créé une crainte [Kästner et al., 2021]

Caractéristiques



Stratégies



Sommaire

- 1 Intelligence Artificielle et Modèles de Langue
 - Définitions
 - Le modèle Transformer
 - Les modèles boîte-noire
- 2 L'eXplicabilité de l'Intelligence Artificielle
 - Définitions et enjeux
 - Taxonomie des méthodes d'explication
- 3 Évaluer les méthodes d'explications
 - Les attendus
 - Les approches
 - Retour sur les enjeux de l'évaluation des méthodes d'explication
- 4 Conclusion

Les attendus

Attendus d'une explication [Gilpin et al., 2019]

- Fidèle au raisonnement du modèle
- Intelligible pour le public visé

Les attendus

Attendus d'une explication [Gilpin et al., 2019]

- Fidèle au raisonnement du modèle
- Intelligible pour le public visé

Attendus d'une méthode d'explication [Mersha et al., 2025]

- Robustesse
- Consistance
- Contrastivité

Les approches

- Comparer à des explications canoniques
- Comparer les méthodes d'explication entre elles
- Évaluer individuellement
- Analyser théoriquement

Comparer les explications à des explications canoniques

Idée

- ① Définir une explication attendue
 - ② Calculer la performance de la méthode d'explication
- La *Ground Truth humaine* n'est pas adaptée
 - Construire une Ground Truth à partir du modèle utilisé ?
 - Explication *idéale* ? [Neely et al., 2022]

Comparer les méthodes d'explications entre elles

Idée

- ① Prendre une méthode d'explication comme référence
 - ② Comparer ses explications avec celles de la méthode d'explication évaluée
- *Approches différentes* d'une méthode à l'autre
 - Suppose la méthode de référence *idéale*
 - Quel sens ? [Neely et al., 2022]

Évaluer individuellement

Idée

Utiliser des métriques indépendantes qui évaluent les attendus

- Définir les attendus (exhaustivité)
- Trouver des métriques qui correspondent
- Besoin de métriques *universelles* ou *comparables*

Aparté : Cas du jugement humain

Idée

Demander à des personnes d'évaluer la méthode à l'aide de critères définis

Mais

- Évaluation qualitative
- Risque de biais [Gilpin et al., 2019]
- Difficile à réaliser
- Problèmes de comparabilité

Analyse théorique des méthodes d'explication

Idée

- ① Reprendre les définitions mathématiques du modèle et de la méthode d'explication
- ② Juger les attendus d'un point de vue théorique
 - Modèles boîte-noire
 - Définir mathématiquement les attendus
 - *Ex : [Lopardo et al., 2024]*

Existence des idéaux

- Explication *idéale*
- Méthode d'explication idéale
 - Pas de consensus [Garouani et al., 2024]
 - Une méthode *universelle* est-elle possible ? (ex : [Mersha et al., 2025])

[Retour sur les enjeux de l'évaluation des méthodes d'explication](#)

Comparaison

- Les explications sont-elles comparables ?
- Les méthodes d'explication sont-elles comparables ?
- Les métriques donnent-elles des résultats comparables ?
- Quel sens ?

Universalité

- Les méthodes d'explication sont-elles adaptées ...
 - ... à tous les modèles ?
 - ... à toutes les tâches ?
- Les métriques sont-elles adaptées à toutes les méthodes d'explication ?

Sommaire

- 1 Intelligence Artificielle et Modèles de Langue
 - Définitions
 - Le modèle Transformer
 - Les modèles boîte-noire
- 2 L'eXplicabilité de l'Intelligence Artificielle
 - Définitions et enjeux
 - Taxonomie des méthodes d'explication
- 3 Évaluer les méthodes d'explications
 - Les attendus
 - Les approches
 - Retour sur les enjeux de l'évaluation des méthodes d'explication
- 4 Conclusion

Perspectives

- Faire un bilan des attendus et des métriques existantes
- Comparer les méthodes d'explications en fonction de la tâche traitée par le système
 - Déterminer des besoins spécifiques aux tâches impliquant du texte
 - Évaluer la pertinence d'une métrique *universelle*



Bell, A., Solano-Kamaiko, I., Nov, O., and Stoyanovich, J. (2022).

It's Just Not That Simple : An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy.

In 2022 ACM Conference on Fairness Accountability and Transparency, pages 248–266, Seoul Republic of Korea. ACM.



Bento, V., Kohler, M., Diaz, P., Mendoza, L., and Pacheco, M. A. (2021).

Improving deep learning performance by using Explainable Artificial Intelligence (XAI) approaches.

Discover Artificial Intelligence, 1(1) :9.



Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).

BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding.

[arXiv :1810.04805.](#)



Garouani, M., Mothe, J., Barhrhouj, A., and Aligon, J. (2024). Investigating the Duality of Interpretability and Explainability in Machine Learning.

In 2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI), pages 861–867.

[arXiv :2503.21356 \[cs\].](#)



Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2019).

Explaining Explanations : An Overview of Interpretability of Machine Learning.

[arXiv :1806.00069 \[cs\].](#)



Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., and Sterz, S. (2021).

On the Relation of Trust and Explainability : Why to Engineer for Trustworthiness.

In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 169–175.



Lopardo, G., Precioso, F., and Garreau, D. (2024).

Attention Meets Post-hoc Interpretability : A Mathematical Perspective.

arXiv :2402.03485.



Mersha, M. A., Yigezu, M. G., and Kalita, J. (2025).

Evaluating the Effectiveness of XAI Techniques for Encoder-Based Language Models.

Knowledge-Based Systems, 310 :113042.

arXiv :2501.15374 [cs].



Montavon, G., Binder, A., Lapuschkin, S., Samek, W., and Müller, K.-R. (2019).

Layer-Wise Relevance Propagation : An Overview.

In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R., editors, *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, pages 193–209. Springer International Publishing, Cham.



Neely, M., Schouten, S. F., Bleeker, M., and Lucic, A. (2022).
A Song of (Dis)agreement : Evaluating the Evaluation of
Explainable Artificial Intelligence in Natural Language Processing.

In *HHAi2022 : Augmenting Human Intellect*, pages 60–78. IOS Press.

-  Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., and Baralis, E. (2023).
Concept-based Explainable Artificial Intelligence : A Survey.
[arXiv :2312.12936](#).
-  Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018).
Improving Language Understanding by Generative Pre-Training.
-  Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020).
Exploring the limits of transfer learning with a unified text-to-text transformer.
[J. Mach. Learn. Res.](#), 21(1).



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017).

Attention Is All You Need.

[arXiv :1706.03762](https://arxiv.org/abs/1706.03762).



Vig, J. (2019).

A Multiscale Visualization of Attention in the Transformer Model.

In Costa-jussà, M. R. and Alfonseca, E., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Sommaire

5 Annexes

- Architecture de BERT
- Architecture de GPT
- Architecture de T5
- Méthode d'explication à base d'attention
- Visualisation de l'attention
- Méthode d'explication par propagation
- Méthode d'explication par perturbation
- Méthodes d'explication par l'exemple
- Modèles à base de concepts

Architecture de BERT

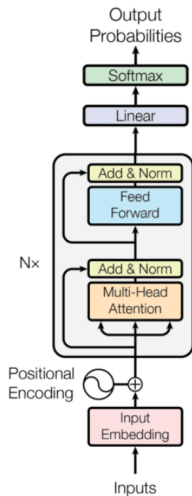


Figure – Architecture du modèle BERT

Architecture de GPT

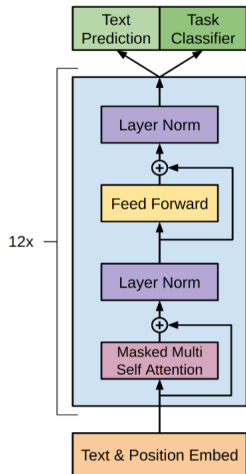
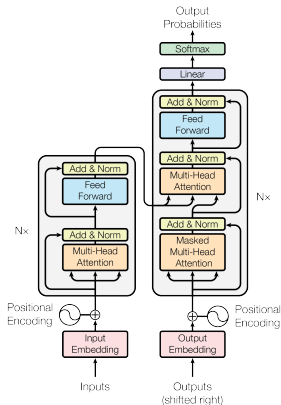


Figure – Architecture du modèle GPT
[Radford et al., 2018]

Architecture de T5



Différences :

- Normalisation sans biais
- Connexion résiduelle après la normalisation
- Encodding positionnel relatif

Figure – Architecture du Transformer [Vaswani et al., 2017]

Méthode d'explication à base d'attention

Matrice d'attention

Mesure l'importance de la relation entre chaque mot de l'entrée

Approches :

- Visualisation
- Exploitation directe d'une seule matrice
- Combinaison des matrices pour obtenir des scores

Méthode d'explication à base d'attention

Matrice d'attention

Mesure l'importance de la relation entre chaque mot de l'entrée

Approches :

- Visualisation
- Exploitation directe d'une seule matrice
- Combinaison des matrices pour obtenir des scores

Remarques :

- Exploitation d'une forme de représentation de l'entrée
- mais prise en compte partielle du modèle
- [Lopardo et al., 2024] remet en question cette approche d'un point de vue théorique

Visualisation de l'attention

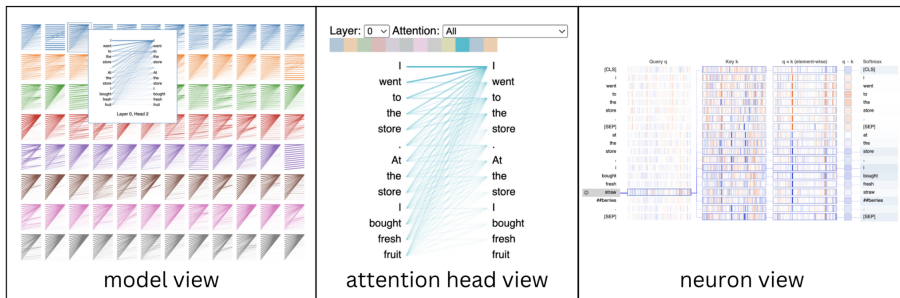


Figure – BertViz [Vig, 2019] permet de visualiser le mécanisme d'attention à différentes échelles

Méthode d'explication par propagation

Intuition

Propager de la sortie vers l'entrée pour retrouver l'importance de chaque partie de l'entrée sur une sortie

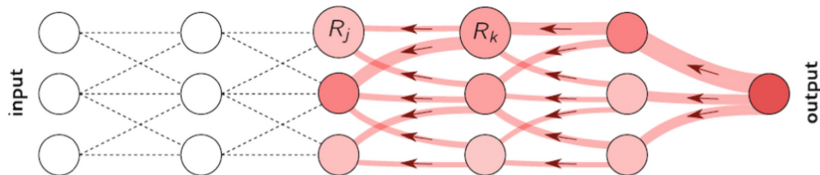


Figure – Intuition de Layer-wise Relevance Propagation [Bento et al., 2021]

$$R_j^{(n)} = \sum_i \frac{x_j^+ w_{ji}^+}{\sum_{j'} x_{j'}^+ w_{j'i}^+} R_i^{(n+1)} \text{ où } v^+ = \max(0, v) \text{ [Montavon et al., 2019]}$$

Méthode d'explication par perturbation

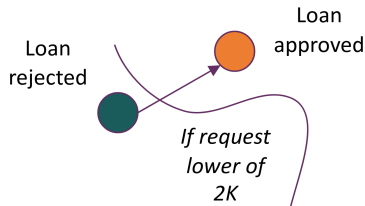
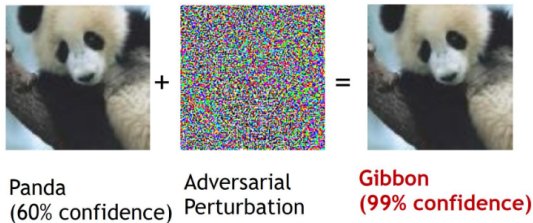
Intuition

Observer le comportement du modèle au voisinage d'une entrée et l'impact des variations sur la prédiction.

Local Interpretable Model-agnostic Explanation (LIME) : Soient M un modèle et x une entrée

- 1 Construire un ensemble $\mathcal{X}$ de versions perturbées de x
- 2 Entraîner un modèle interprétable M' sur $\mathcal{X} \times M(\mathcal{X})$
- 3 L'interprétation de M' est une approximation de M au voisinage de x

Méthodes d'explication par l'exemple



Modèles à base de concepts

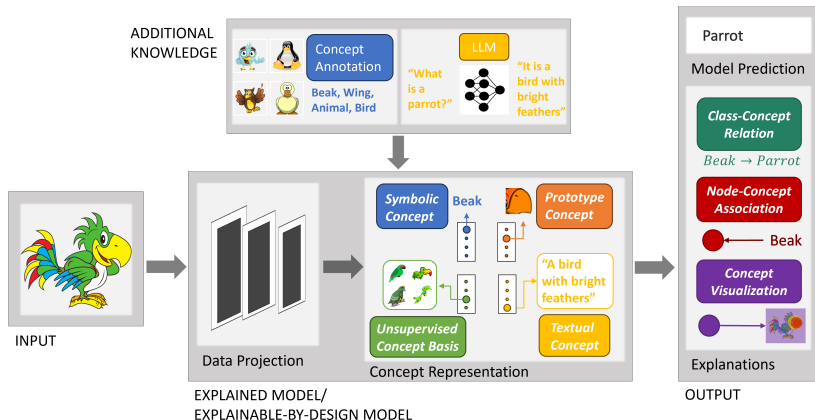


Figure – Types de concepts et d'explications à base de concepts
[Poeta et al., 2023]