Représentation des poids d'auto-attention sous forme de graphe pour l'évaluation des Transformers

Rebecca Leygonie*, Sylvain Lobry*, Laurent Wendling*

*Université Paris Cité, LIPADE, F-75006 Paris, France

Résumé. Les Transformers sont devenus la référence pour le traitement de données séquentielles, avec des applications allant de la traduction au traitement des dossiers de santé électroniques. Cependant, leur complexité pose des défis d'explicabilité, notamment dans des domaines soumis à des exigences éthiques et légales strictes, comme la santé. Pour répondre à ce besoin, nous proposons une approche qui représente l'apprentissage du mécanisme d'attention sous forme de graphe, révélant les connexions d'auto-attention entre les tokens. Nous introduisons une métrique pour évaluer la pertinence des connexions apprises par rapport à une référence établie. Nous appliquons notre approche au modèle Behrt, conçu pour prédire les diagnostics futurs à partir de diagnostics passés. Nos expérimentations montrent que notre méthode facilite la compréhension de l'apprentissage des modèles et permet une meilleure appréciation de l'influence des diagnostics entre eux, ainsi que des biais présents dans les données.

1 Introduction

Depuis leur introduction par Vaswani et al. (2017), les architectures Transformers sont devenues l'état de l'art pour le traitement des données séquentielles (Wen et al. (2023)), principalement grâce à leur mécanisme d'auto-attention qui capture les relations entre les éléments d'une séquence tout en minimisant le problème de la disparition du gradient (Bengio et al. (1994)). Les modèles basés sur les Transformers, comme BERT (Devlin et al. (2019)), ont révolutionné le traitement du langage naturel et sont maintenant utilisés pour traiter tous types de données séquentielles, y compris les dossiers de santé électroniques (EHR) (Nerella et al. (2023)). Cependant, le manque d'explicabilité des Transformers soulève des questions juridiques et éthiques qui entravent leur déploiement dans le domaine de la santé (Shortliffe et al. (2018)).

Alors qu'il est possible d'expliquer l'apprentissage d'un modèle en visualisant les poids d'auto-attention appris, cela se fait généralement sur quelques exemples, souvent pour montrer ce que le modèle a bien appris, sans comparaison avec une vérité terrain (Siebra et al. (2024)).

Pour répondre à ce problème, nous proposons une approche permettant de valider l'apprentissage d'un modèle basé sur des mécanismes d'attention en représentant, sous forme de graphe, les liens entre les tokens d'entrée, pondérés par les poids d'auto-attention appris par le modèle. Nous évaluons ensuite le graphe par rapport à un graphe de référence établi auprès d'experts afin d'obtenir un score de pertinence des liens d'auto-attention appris par le modèle.

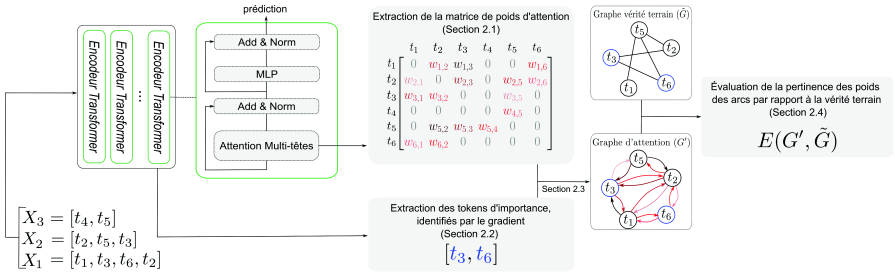

FIG. 1 – *Vue d'ensemble de la méthodologie proposée.*

Nous appliquons notre approche au modèle Behrt (Li et al. (2020)) entraîné sur les données du *Système National des Données de Santé* (SNDS) pour prédire la prochaine visite à l'hôpital d'un patient à partir d'une séquence de visites précédentes. Les résultats montrent que l'utilisation de la représentation par graphe permet de comprendre plus directement les liens que le modèle a appris, ce qui est crucial pour accroître la confiance des professionnels de la santé dans l'utilisation des prédictions du modèle dans des situations cliniques concrètes.

Ce travail comporte quatre contributions. **1** : une méthode qui représente sous forme de graphe les interactions entre les tokens apprises par le modèle. **2** : une approche pour modéliser l'expertise médicale sous forme de graphe. **3** : une nouvelle métrique pour évaluer la pertinence des connexions apprises par le modèle en les comparant à une référence établie. **4** : la validation de la méthode à travers deux cas d'usage sur des données de santé.

2 Méthodologie

2.1 Création de la matrice d'attention globale

Soit $T = \{t_1, t_2, \ldots, t_V\}$ un ensemble (ou vocabulaire) de V *tokens* distincts. Nous considérons un ensemble de données étiquetées $X = (X_i, y_i)_{i \in [\![1;N]\!]}$, composé de N séquences X_i, chacune associée à une étiquette y_i. Plus précisément, chaque X_i est une séquence de z_i tokens $(x_1, \ldots, x_{z_i})$, où $z_i \in \mathbb{N}$ et $\forall j \in [\![1, z_i]\!], x_j \in T$. Nous passons chaque séquence X_i à un Transformer entraîné pour prédire $\hat{y}_i$. Si la prédiction du modèle $\hat{y}_i$ correspond à l'étiquette réelle y_i, nous récupérons la matrice d'attention (de la dernière couche) A_i de taille $(z_i \times z_i)$ pendant le traitement de X_i par le modèle, où chaque élément $a_{m,n}$ de la matrice A_i représente l'auto-attention que le token x_m dans la séquence X_i donne au token x_n dans la même séquence. Cette valeur est calculée comme la moyenne de l'auto-attention entre ces tokens dans toutes les têtes d'attention de la couche sélectionnée.

Une fois les poids d'auto-attention récupérés pour chaque paire dans chaque séquence, nous agrégeons ces poids dans une matrice d'attention globale G de taille $(V \times V)$. G peut être interprétée comme une matrice d'adjacence pondérée, représentant les relations d'auto-attention entre tous les tokens du vocabulaire. Elle synthétise l'attention que chaque token t_m accorde à chaque autre token t_n à travers toutes les séquences analysées. Pour chaque paire de tokens (t_m, t_n), nous identifions toutes les occurrences de ces tokens dans les différentes

séquences et accumulons les valeurs d'auto-attention correspondantes à partir des matrices A_i associées à chaque séquence. Nous collectons ces valeurs dans un ensemble S_{mn} :

$S_{mn} = \{a_{m',n'} \mid x_{m'} = t_m, x_{n'} = t_n, m, n \in \{1, \ldots, V\}^2, m', n' \in \{1 \ldots, z_i\}^2\}$

Chaque élément $G[m, n]$ de la matrice est ensuite calculé en prenant la médiane des valeurs de l'ensemble S_{mn}. Le choix de l'agrégation par la médiane a pour objectif d'éviter les valeurs aberrantes potentielles. Finalement, G capture la tendance centrale de l'intensité d'interaction entre chaque paire de tokens dans l'ensemble de données.

2.2 Identification des tokens d'importance

Pour chaque séquence X_i où le modèle prédit correctement l'étiquette $\hat{y}_i$, nous récupérons le gradient $\Delta(x_j)$ de chaque token x_j dans X_i en effectuant un passage de rétropropagation. Les gradients nous permettent de mesurer l'influence de chaque token sur la prédiction. Simonyan et al. (2014) affirment que lorsque le modèle fait une bonne prédiction, des gradients de valeurs plus élevées indiquent une contribution plus significative à la décision du modèle. Pour chaque token du vocabulaire T qui apparaît dans au moins une séquence correctement prédite, nous calculons l'importance médiane des gradients associés à ce token. Enfin, nous sélectionnons un nombre prédéfini g des tokens les plus importants selon ces mesures de gradients médians, identifiant ainsi les tokens qui influencent de manière cohérente et significative les prédictions correctes du modèle.

2.3 Génération du graphe d'auto-attention

Nous souhaitons construire un graphe d'auto-attention G' qui représente les interactions entre les tokens telles qu'apprises par le modèle. Pour cela, nous extrayons un sous-graphe de la matrice d'attention globale en sélectionnant les relations entre un ensemble choisi de tokens (nœuds). Pour créer un graphe interprétable et comparable, nous utilisons les g tokens ayant les gradients les plus significatifs comme nœuds initiaux de G'. Ensuite, nous élargissons cet ensemble initial en ajoutant tout token t_n du vocabulaire global T pour lequel l'auto-attention $G[m, n]$ est non nulle, où le token t_m ou t_n appartient à l'ensemble initial des g tokens d'importance. Ainsi, l'ensemble des nœuds dans G' comprend les g tokens initiaux et tous les tokens qui leur sont directement liés, avec leurs poids d'auto-attention médians associés.

Nous construisons un graphe dirigé où chaque paire (t_m, t_n) parmi les nœuds de G' est connectée par un arc de t_m à t_n si $G[m, n]$ est non nul. Les arcs sont pondérés par les valeurs correspondantes de G, qui quantifient l'intensité de l'auto-attention que t_m accorde à t_n.

2.4 Évaluation des liens d'auto-attention appris par le modèle

Nous souhaitons évaluer la pertinence des liens d'auto-attention entre les tokens appris par un Transformer. Pour ce faire, nous comparons ces liens à une référence établie, représentée par le graphe $\tilde{G}$. L'évaluation est réalisée en calculant la différence entre la proportion pondérée des arcs de G' communs à $\tilde{G}$ et la proportion pondérée des arcs de G' non communs à $\tilde{G}$. Concernant $\tilde{G}$, le seul prérequis est qu'il soit représenté sous forme de graphe. Il peut être orienté ou non et nous considérons qu'il n'est pas pondéré car il est difficile de transposer la pondération par l'auto-attention à un cas d'usage réel. Dans la Section 3.4, nous détaillons le protocole que nous avons développé pour générer le graphe de référence utilisé dans nos

expérimentations. La fonction d'évaluation, que nous appelons $E \in [-1, 1]$, est calculée par : $E(G', \tilde{G}) = w_{in} - w_{out}$ où $w_{in} = \frac{\sum_{(m,n)\in G' \cap \tilde{G}} w_{mn}}{\sum_{(m,n)\in G'} w_{mn}} \in [0, 1]$, est la proportion pondérée des arcs dans le graphe d'auto-attention G' qui sont également présents dans le graphe de référence $\tilde{G}$, c'est-à-dire le nombre d'arcs de G' inclus dans $\tilde{G}$ pondérés par leur poids et normalisé par le nombre pondéré d'arcs dans G'. Et $w_{out} = \frac{\sum_{(m,n)\in G', (m,n)\notin \tilde{G}} w_{mn}}{\sum_{(m,n)\in G'} w_{mn}} \in [0, 1]$, est la proportion pondérée des arcs dans G' qui ne sont pas confirmés par $\tilde{G}$.

La fonction d'évaluation E mesure l'alignement entre le graphe G' et le graphe de référence $\tilde{G}$. Elle varie de -1 à 1, où un score de 1 est atteint lorsque tous les arcs de G' sont inclus dans $\tilde{G}$, et -1 lorsqu'aucun arc n'est inclus. Les scores entre 0 et 1 indiquent que les arcs inclus dans $\tilde{G}$ sont plus fortement pondérés que ceux qui ne le sont pas, reflétant une prédominance de correspondances. À l'inverse, les scores entre -1 et 0 indiquent que les arcs non inclus dans $\tilde{G}$ sont plus fortement pondérés, reflétant une prédominance de non-correspondances.

3 Expérimentations

3.1 Modèle

Nous souhaitons étudier l'apprentissage d'un modèle basé sur des mécanismes d'attention appliqué à l'analyse des données médicales. Pour nos expériences, nous utilisons le modèle Behrt, entraîné par le Lab Santé de la Drees [1] sur la tâche de prédiction des diagnostics des patients lors de futures visites à l'hôpital, en se basant sur une séquence historique de visites.

Les données sur lesquelles le modèle est entraîné proviennent des tables MCO (*Médecine, Chirurgie, Obstétrique*) du PMSI (*Programme de Médicalisation des Systèmes d'Information*) du SNDS. Chaque visite à l'hôpital est caractérisée par un ensemble de diagnostics, comprenant un diagnostic principal et, le cas échéant, un diagnostic associé, ainsi que plusieurs diagnostics complémentaires, qui enrichissent le contexte du diagnostic principal.

Le modèle est entraîné sur une tâche de classification multi-classe et multi-étiquette. Les classes sont représentées par 2 053 diagnostics codés selon la 10e révision de la CIM [2]. L'entraînement se divise en deux phases : d'abord la prédiction de mots masqués sur $14,59M$ exemples (patients avec > 2 visites et ≥ 3 codes diagnostiques), puis la prédiction des diagnostics futurs sur $5,94M$ parcours (patients avec ≥ 4 visites). Les données proviennent d'un échantillon de 4% du SNDS 2008-2017 et des données complètes 2018-2021.

3.2 Cas d'usage

Nous appliquons notre méthode lors de la phase d'inférence du modèle, en récupérant les historiques médicaux des individus pour lesquels le modèle a correctement prédit la prochaine visite, c'est-à-dire lorsque le diagnostic à prédire figure parmi le top-2 des prédictions. Nous travaillons sur deux cas d'usage distincts qui concernent la prédiction de diagnostics incidents,

1. Nous remercions Milena Suarez Castillo et Javier Nicolau pour cette collaboration, l'équipe du Lab Santé de la Drees (https ://drees.solidarites-sante.gouv.fr/drees) pour le développement du modèle et les experts médicaux Diane Naouri, Albert Vuagnat, et Constance Prieur pour leur précieuse expertise et le temps passé sur ces travaux. Leur participation active a largement contribué à la qualité de nos travaux.

2. https ://icd.who.int/browse10/2019/en

ce qui signifie que le diagnostic à prédire n'apparaît pas dans la séquence d'entrée, permettant d'analyser précisément l'influence des diagnostics précédents sur les prédictions.
Cas d'usage 1 La classe *accouchement* inclut les codes CIM-10 suivants : (**O80**) accouchement spontané unique ; (**O81**) accouchement unique par forceps et ventouse ; (**O82**) accouchement unique par césarienne ; (**O83**) autres accouchements assistés uniques ; et (**O84**) accouchements multiples. Nous sélectionnons 2 000 individus pour lesquels la visite à prédire contient exactement un code de la classe *accouchement*, qui n'est pas dans la séquence que nous donnons comme entrée au modèle. Sur l'échantillon utilisé, le modèle fait une bonne prédiction pour 190 individus.
Cas d'usage 2 La classe *maladies hypertensives* est définie par les codes CIM-10 suivants : (**I10**) hypertension essentielle ; (**I11**) cardiopathie hypertensive ; (**I12**) néphropathie hypertensive ; (**I13**) cardionéphropathie hypertensive ; et (**I15**) hypertension secondaire. Comme dans le cas de l'accouchement, nous sélectionnons 2 000 individus pour lesquels la visite à prédire contient le diagnostic I10 et aucun code de la classe des *maladies hypertensives* n'est dans la séquence correspondant à l'historique médical. Sur l'échantillon utilisé, le modèle fait une bonne prédiction pour 514 individus.

3.3 Génération du graphe G' à partir de Behrt en inférence

Pour chaque cas d'usage, la création du graphe commence par la récupération des matrices d'attention pour chaque individu, extraites de la dernière couche du modèle Behrt. Ensuite, nous construisons une matrice d'attention globale de dimensions 2053×2053, où 2053 représente le nombre de diagnostics possibles.

Nous extrayons également le gradient associé à chaque diagnostic à partir de la séquence représentant l'historique médical d'un individu, donnée en entrée du modèle. Cela nous permet d'identifier, pour chaque individu, quels diagnostics ont le plus influencé la prédiction. Nous calculons ensuite la médiane des gradients par diagnostic.

Pour les deux cas d'usage et afin de préserver la confidentialité, nous nous limitons à l'étude des diagnostics ou des paires de diagnostics présents dans au moins cinq parcours médicaux distincts. Cette méthode nous permet de maintenir l'anonymat tout en préservant la pertinence analytique. En conséquence, certains diagnostics sélectionnés parmi les g diagnostics d'importance peuvent ne pas apparaître dans notre graphe si leurs connexions avec d'autres diagnostics ne se retrouvent pas dans le minimum requis de cinq parcours médicaux distincts. De même, une connexion entre deux diagnostics ne sera pas visible dans le graphe si la paire de diagnostics n'apparaît pas dans au moins cinq séquences distinctes. Ainsi, pour les deux cas d'usage, le nombre de diagnostics significatifs identifiés par la valeur médiane des gradients est choisi arbitrairement de manière à ce que le graphe généré contienne suffisamment d'arcs.

Nous sélectionnons respectivement les 10 et 46 diagnostics les plus influents en fonction de la valeur médiane de leur gradient pour les cas d'usage de l'accouchement et de l'hypertensivité. À partir de ces diagnostics, nous extrayons le sous-graphe d'auto-attention G' qui relie les nœuds importants à d'autres nœuds associés dans la matrice d'attention globale, ainsi que les liens entre les nœuds ajoutés.

3.4 Création du graphe de référence

Pour interpréter les liens entre les diagnostics appris par le modèle Behrt à travers les mécanismes d'attention, nous souhaitons les comparer à une vérité terrain, qui se traduit par l'expertise médicale. Bien que l'auto-attention soit significative dans notre modèle, elle n'a pas de correspondance directe et évidente dans le contexte médical. Afin de valider ou d'invalider ces liens, nous avons conçu un protocole impliquant des professionnels de santé, permettant la création d'un graphe de relations entre les diagnostics. Ce graphe est conçu pour être non dirigé et non pondéré.

Le protocole repose sur deux listes de diagnostics : "gradients", qui contient les diagnostics que nous avons identifiés comme importants pour la prédiction et "autres", qui comprend les diagnostics qui sont liés aux diagnostics "gradient" selon la matrice d'attention globale. Pour produire un graphe qui reflète la réalité clinique des deux cas d'usage présentés, nous avons fait appel à l'expertise de deux experts médicaux de la Drees pour former des clusters. Nous leur avons spécifiquement demandé de relier chaque diagnostic de la liste "autres" à un ou plusieurs diagnostics de la liste "gradients", en tenant compte de l'existence d'une corrélation contextuelle entre eux. Cette corrélation peut concerner des éléments tels que la comorbidité, la causalité, l'impact sur le traitement, l'implication clinique ou la fréquence de codage.

3.5 Évaluation de G' par rapport à $\tilde{G}$

Nous évaluons les poids d'auto-attention du graphe G' par rapport au graphe de référence $\tilde{G}$ en calculant $E(G', \tilde{G})$. Étant donné que G' est dirigé et que $\tilde{G}$ ne l'est pas, nous considérons les arêtes de $\tilde{G}$ comme des arcs bidirectionnels. L'objectif est d'évaluer G' en fonction de différents seuils qui déterminent quels arcs sont pris en compte. Plus précisément, nous cherchons à déterminer s'il existe un seuil où les arcs de G' qui sont communs à $\tilde{G}$ sont correctement identifiés par le modèle comme étant significatifs, c'est-à-dire que le poids de ces arcs est supérieur à celui des arcs non inclus. Pour chaque seuil établi, nous considérons uniquement les arcs de G' dont le poids dépasse ce seuil. Nous éliminons ensuite les nœuds isolés de G' tout en conservant les mêmes nœuds dans $\tilde{G}$.

4 Résultats et Discussions

Les courbes évaluant le score de G' par rapport aux graphes de référence sont présentées pour chaque cas d'usage dans la Figure 2. Dans le cas de l'accouchement (Figure 2a), nous observons qu'au-dessus d'un seuil de 0,13 ou 0,15, selon la référence établie, le score atteint 1. Cela indique que tous les arcs de G' sont inclus dans $\tilde{G}$, confirmant l'existence d'un seuil au-delà duquel les arcs correspondent précisément à la vérité terrain et démontrant que le modèle attribue correctement des poids d'auto-attention élevés aux arcs pertinents. Avant ce seuil, le score reste positif mais ne dépasse pas 0,25, ce qui suggère que, bien que certains arcs ne soient pas inclus dans la vérité terrain, leur influence est relativement mineure.

Dans le cas de l'hypertensivité (cf. Figure 2b), les courbes montrent un déclin constant, ce qui signifie que plus le seuil de sélection des arcs est élevé, moins il y a d'arcs en commun avec la référence établie, ou bien leur poids est inférieur à celui des arcs exclus. Ces résultats

(a) Cas d'usage accouchement

(b) Cas d'usage hypertensivité

FIG. 2 – *Courbes illustrant l'évolution du score d'évaluation en fonction du seuil d'auto-attention utilisé pour sélectionner les arcs dans le graphe d'auto-attention G'.*

indiquent que, dans ce cas d'usage, le modèle ne porte pas suffisamment attention aux paires de diagnostics validées par l'expertise médicale.

Bien que le modèle sur lequel nous appliquons notre approche n'atteigne pas un taux de rappel élevé — diagnostiquant correctement 190 sur 2000 individus dans le cas de l'accouchement et 514 sur 2000 dans le cas de l'hypertension — l'objectif est de déterminer si ces résultats sont le fruit d'un surapprentissage. Les analyses révèlent que le modèle établit des liens diagnostics plus pertinents pour l'accouchement que pour l'hypertension, ce qui est surprenant compte tenu du meilleur taux de rappel observé pour l'hypertension. Cette anomalie est interprétée comme étant due à la complexité du cas de l'hypertension, qui présente une grande variété de parcours diagnostiques pouvant mener à des prédictions correctes. Ces résultats démontrent la valeur ajoutée de notre approche, dont l'analyse crée un lien direct entre les données d'entraînement utilisées et la performance obtenue, permettant ainsi un ajustement potentiel de l'échantillon utilisé. Enfin, le protocole développé pour créer un graphe de référence permet d'évaluer le graphe d'auto-attention issu de l'apprentissage. Les courbes d'évolution de score (cf. Figure 2) montrent que l'évaluation suit la même tendance quelle que soit l'expertise médicale comparée, validant ainsi le graphe de référence comme représentatif des connaissances médicales générales.

5 Conclusion

Notre méthode vise à valider l'apprentissage des modèles basés sur des mécanismes d'attention en représentant les liens d'auto-attention appris sous forme de graphe et en évaluant leur pertinence par rapport à un référence établie. Nous appliquons notre approche sur le modèle Behrt, entraîné pour prédire le diagnostic de la prochaine visite à l'hôpital à partir d'une série de visites précédentes.

Nous proposons une méthode pour créer un graphe de référence à partir d'un protocole simple à mettre en œuvre. Dans nos expérimentations, nous considérons individuellement les graphes de références définis par différents experts médicaux. Par la suite, nous souhaitons unifier ces différentes références en une seule et développer une méthode pour pondérer les arêtes établies par l'expertise médicale, permettant ainsi une évaluation plus précise de la distribution de l'auto-attention apprise par le modèle.

Pour générer le graphe d'auto-attention, nous utilisons les poids d'attention de la dernière couche du modèle. Par la suite, nous souhaitons concevoir une méthode d'intégration de l'attention de toutes les couches, afin d'évaluer l'apprentissage de manière plus complète. De plus, nous aimerions analyser chaque tête d'attention individuellement pour examiner si les poids associés peuvent avoir des interprétations distinctes dans le contexte médical. Enfin, nous aimerions étudier l'impact du choix de l'opération d'agrégation des poids d'auto-attention.

Finalement, les résultats obtenus démontrent que notre approche permet une évaluation plus granulaire de l'apprentissage du modèle que les mesures de performances classiques. Nos expériences sur deux cas d'usage ont révélé un phénomène inattendu : le modèle ayant de meilleures performances initiales établit des liens diagnostiques moins pertinents que celui semblant a priori moins performant. Notre méthode a ainsi mis en lumière des biais liés aux données d'entraînement, difficilement perceptibles avec des mesures standard comme le rappel. En offrant une analyse automatique et approfondie de l'apprentissage des relations entre les éléments des séquences, notre approche améliore la compréhension et la confiance dans les prédictions du modèle.

Références

Bengio, Y. et al. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks 5*(2), 157–166.

Devlin, J. et al. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the ACL*, pp. 4171–4186.

Li, Y. et al. (2020). BEHRT : transformer for electronic health records. *Scientific Reports 10*(1), 7155.

Nerella, S. et al. (2023). Transformers in healthcare : A survey. *arXiv preprint arXiv :2307.00067*.

Shortliffe, E. H. et al. (2018). Clinical decision support in the era of artificial intelligence. *Jama 320*(21), 2199–2200.

Siebra, C. A. et al. (2024). Transformers in health : a systematic review on architectures for longitudinal data analysis. *Artificial Intelligence Review 57*(2), 1–39.

Simonyan, K. et al. (2014). Deep inside convolutional networks : visualising image classification models and saliency maps. In *ICLR*.

Vaswani, A. et al. (2017). Attention is all you need. *31st NIPS*, 6000–6010.

Wen, Q. et al. (2023). Transformers in time series : a survey. In *32nd IJCAI*, pp. 6778–6786.

Summary

Transformers have revolutionized sequential data processing but lack explainability, particularly problematic in regulated fields like healthcare. Our work introduces a graph-based visualization of attention learning and a metric for validating learned connections against ground truth. Testing on Behrt (a diagnostic prediction model) demonstrates how our method reveals inter-diagnosis relationships and dataset biases.