REPLY AUTONOMOUS

LIVRE BLANC

AU-DELÀ DES AGENTS LLM : ENTRAÎNEMENT DE MODÈLES D'IA POUR LES SIGNAUX AUTOMOBILES

Les systèmes de séries temporelles sont partout — mais comprendre les signaux désordonnés du monde réel limite toujours les performances.

LIVRE BLANC

SOMMAIRE

TL ; DR

Les résultats montrent des améliorations claires. Au-delà d'une meilleure gestion des dépendances temporelles longues et d'une robustesse plus forte à travers divers scénarios de conduite, le progrès le plus significatif réside dans la capacité du modèle à capturer correctement des événements extrêmes. Contrairement aux approches traditionnelles basées sur LSTM ou GRU, qui tendent à lisser les prédictions et à sous-réagir à des dynamiques rares, **le modèle basé sur Transformer reste sensible aux déviations marquées du signal**. Cela est particulièrement important dans les applications automobiles, où les valeurs aberrantes ne sont pas du bruit mais des événements critiques (comme une perte soudaine de contrôle, des manœuvres agressives, des situations de maniabilité proche de la limite ou des transitoires de perte de stabilité). Prédire avec précision ces comportements rares mais à fort impact est essentiel pour des systèmes axés sur la sécurité. En plus d'une précision améliorée dans ces régimes, le modèle montre une plus grande stabilité et généralisation globale, avec une dérive réduite dans le temps et un comportement plus cohérent à travers des données de test non vues. Les mécanismes d'attention fournissent également un aperçu des signaux et des périodes temporelles qui guident la prédiction, ajoutant une couche précieuse d'interprétabilité à ces décisions cruciales.

Cela dit, de fortes performances hors ligne ne se traduisent pas automatiquement par une préparation à la production. Le déploiement sur les plateformes embarquées introduit de nouvelles contraintes, notamment une mémoire limitée, des budgets de calcul et des exigences strictes de latence en temps réel. La chaîne d'outils, de la formation à l'exportation au format standardisé et à l'intégration finale dans les environnements automobiles, reste un effort d'ingénierie non trivial. Assurer un comportement déterministe et le respect des normes critiques pour la sécurité est une autre étape importante avant l'industrialisation.

En regardant vers l'avenir, cette approche va bien au-delà d'un simple capteur virtuel. Les modèles de séries temporelles basés sur l'attention peuvent soutenir des estimations supplémentaires de la dynamique du véhicule, la prédiction de durabilité et de fatigue, les stratégies de maintenance prédictive et la modélisation du comportement du conducteur. Plus largement, tout système industriel caractérisé par des données séquentielles structurées et de longues dépendances temporelles peut bénéficier de ce changement architectural. Les transformateurs ne sont plus limités au traitement du langage naturel ; Ils deviennent des outils puissants pour raisonner sur les signaux dans des systèmes physiques complexes.

LE PROBLÈME

Estimer une variable de dynamique cachée du véhicule à partir des signaux de production peut sembler simple en théorie, mais les données réelles sont fondamentalement complexes. Contrairement aux environnements simulés, les signaux embarqués dans les véhicules automobiles sont bruyants, imparfaits, asynchrones et parfois extrêmes. Ces caractéristiques rendent la modélisation des séries temporelles nettement plus difficile que les problèmes de régression standards.

Premièrement, les signaux réels sont bruyants. Les capteurs sont affectés par les vibrations, les variations de température, les effets de quantification et les interférences électroniques. Les vitesses des roues peuvent fluctuer en raison d'irrégularités de la route, les accéléromètres capturent à la fois la dynamique réelle et les vibrations structurelles, et les mesures de direction peuvent inclure des jeux mécaniques ou des artefacts de filtrage. Un modèle doit apprendre le comportement physique sous-jacent tout en ignorant les variations fallacieuses. Le surapprentissage au bruit conduit à des prédictions instables, tandis que le sur-lissage supprime des dynamiques critiques.

Deuxièmement, les signaux dérivent avec le temps. Les ajustements des capteurs changent, l'usure des pneus évolue, les variations de charge des véhicules et les conditions environnementales varient. Même si deux séquences de conduite se ressemblent, la distribution statistique des signaux peut ne pas être identique. Un modèle robuste doit généraliser à ces variations lentes sans dégrader les performances. La dérive est particulièrement difficile car elle est graduelle et souvent pas explicitement étiquetée.

Troisièmement, les données réelles contiennent souvent des manquants et des irrégularités. Les messages CAN peuvent être retardés ou supprimés. Certains signaux ne sont pas échantillonnés à chaque pas de temps. Lors de certaines manœuvres, certains signaux peuvent saturer ou devenir peu fiables. Les modèles traditionnels supposent des séquences propres et uniformément échantillonnées ; Les données de production répondent rarement à cette hypothèse. Le modèle doit rester stable même lorsque certaines parties de la séquence d'entrée sont imparfaites.

Enfin, et de manière plus cruciale dans le domaine automobile, les données réelles contiennent des événements rares mais à fort impact. Les manœuvres extrêmes, la perte soudaine de contrôle, les situations proches de la limite ou les interventions agressives du pilote sont statistiquement rares mais physiquement cruciales.

D'un point de vue purement statistique, ces valeurs aberrantes représentent une petite fraction de l'ensemble de données. D'un point de vue sécurité, ce sont les moments les plus importants à prédire avec précision. De nombreux modèles classiques de séquences tendent à lisser les prédictions vers le comportement moyen, ce qui peut entraîner une sous-estimation lors de transitoires brusques ou de dynamiques extrêmes. Dans les applications critiques pour la sécurité, cet effet de lissage pose problème.

Tous ces facteurs réunis rendent la modélisation des séries temporelles automobiles réelles significativement plus difficile que les tâches de régression contrôlée en laboratoire. Le défi n'est pas seulement d'obtenir une faible erreur moyenne, mais aussi de maintenir robustesse, stabilité et sensibilité aux dynamiques critiques dans des conditions de conduite diverses et imparfaites.

C'est précisément pour cela qu'une architecture capable d'attention sélective, de pondération dynamique des caractéristiques et de raisonnement temporel à longue portée devient précieuse dans ce contexte.

L'APPROCHE

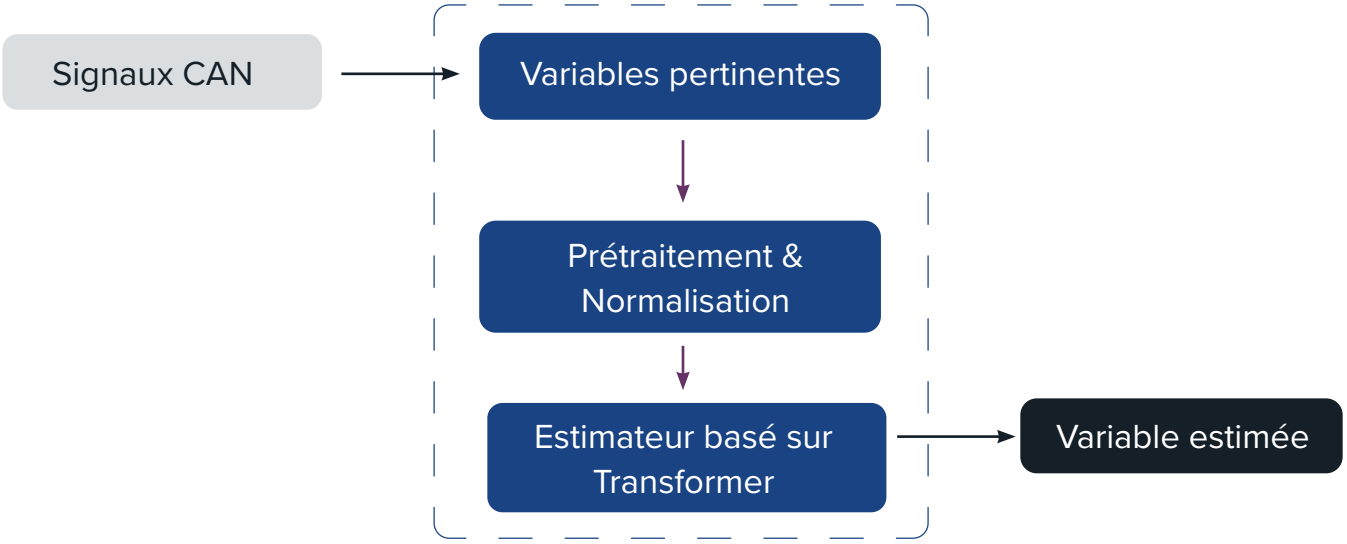
Estimer une variable de dynamique de véhicule cachée à partir des signaux de production nécessite plus que la simple sélection d'une architecture de modèle. Elle exige un pipeline structuré qui relie les données brutes des capteurs et un raisonnement temporel robuste. Notre approche a donc été conçue comme un système de bout en bout, combinant ingénierie du signal, prétraitement discipliné et modélisation de séquences basée sur l'attention au sein d'un cadre cohérent d'IA.

Le processus commence par des signaux **CAN** de niveau production. Ces signaux décrivent le comportement dynamique du véhicule à travers les commandes de direction, la vitesse des roues, les accélérations et les taux de rotation. Bien qu'individuellement informatives, leur véritable valeur émerge lorsqu'elles sont considérées conjointement au fil du temps. Le défi n'est pas simplement de mapper les entrées aux sorties, mais de capturer les relations temporelles et les interactions non linéaires qui définissent la dynamique des véhicules.

Plutôt que de fournir tous les signaux disponibles dans un modèle de manière indiscriminée, **la première étape consiste à identifier les caractéristiques les plus pertinentes.** Cette étape combine une expertise de domaine en dynamique des véhicules avec une analyse exploratoire des données. Des techniques d'évaluation statistique et de réduction de dimensionnalité sont utilisées pour évaluer la redondance, la structure de corrélation et la contribution à l'information entre les signaux. **L'objectif est double : conserver le sens physique tout en réduisant la complexité inutile.** Un ensemble de fonctionnalités compact et bien structuré améliore la stabilité, l'interprétabilité et la généralisation du modèle.

Une fois les entrées concernées définies, **elles sont passées par une étape structurée de prétraitement et de normalisation.** Même lorsque les ensembles de données sont propres et synchronisés, les magnitudes des signaux peuvent différer considérablement. L'angle de direction, la vitesse des roues et les mesures inertielles fonctionnent sur différentes échelles numériques, ce qui peut biaiser l'optimisation basée sur le gradient si elle n'est pas bien gérée. Un pipeline de normalisation cohérent garantit que chaque fonctionnalité contribue proportionnellement pendant l'entraînement. Cette étape est conçue pour être reproductible et modulaire, permettant des expérimentations cohérentes à travers différentes familles de modèles.

Au niveau du système, l'architecture peut se résumer comme suit :



Pipeline IA

Ce pipeline reflète une séparation délibérée des préoccupations. La sélection du signal assure la pertinence physique, le prétraitement assure la stabilité numérique, et l'estimateur se concentre uniquement sur la modélisation temporelle.

Au cœur du système se trouve un modèle de séquence basé sur l'attention. Les architectures récurrentes traditionnelles telles que les **LSTM** et les **GRU** reposent sur des états cachés compressés pour encoder les informations passées. Bien qu'efficace, ce mécanisme peut être difficile lorsque des dépendances à longue distance ou des changements de régime brusques sont impliqués. En revanche, l'estimateur basé sur Transformer exploite les mécanismes d'auto-attention pour évaluer dynamiquement les relations sur des pas de temps.

Cela signifie que le modèle ne se contente pas de « se souvenir » du passé ; **Il pèse activement quels moments sont les plus pertinents** pour la prédiction actuelle. En dynamique des véhicules, cela est particulièrement important. La pertinence des commandes de direction passées, des variations de vitesse de lacet ou des schémas d'accélération peut varier selon la manœuvre exécutée. Un modèle basé sur l'attention s'adapte à ce contexte en temps réel.

Un autre aspect important de l'architecture est sa capacité à apprendre dynamiquement l'importance variable. Au lieu de traiter tous les signaux d'entrée de manière égale en permanence, le modèle peut changer de focus selon le régime de conduite. Lors du mouvement en régime stationnaire, certains signaux peuvent dominer ; lors de manœuvres agressives, d'autres peuvent devenir critiques. Cette capacité de pondération dynamique améliore la robustesse et l'interprétabilité.

D'un point de vue modélisateur, le système fonctionne sur des segments de conduite séquentiels. Les données sont structurées en fenêtres temporelles qui préservent la continuité temporelle, permettant au modèle d'apprendre à la fois les transitoires à court terme et les tendances comportementales à long terme. Cette configuration garantit que l'estimateur capture l'évolution dynamique complète de l'état du véhicule plutôt que des instantanés isolés.

La philosophie de conception globale met l'accent sur la modularité et l'extensibilité. L'étape de sélection des caractéristiques peut être adaptée à différents objectifs d'estimation. Le pipeline de prétraitement peut accueillir des ensembles de données alternatifs. L'estimateur basé sur l'attention peut être remplacé ou complété par d'autres architectures si les contraintes de déploiement l'exigent. Cette flexibilité garantit que la solution n'est pas limitée à un seul cas d'usage mais peut se généraliser à d'autres problèmes de détection virtuelle et de dynamique des véhicules.

En structurant le système de cette manière, nous allons au-delà de l'expérimentation isolée de modèles pour nous orienter vers un flux de travail industriel reproductible et industriel, adapté à l'estimation de séries temporelles pertinentes pour la sécurité.

ÉTUDE DE CAS

Étude de cas : détection virtuelle dans un système automobile à grande échelle (client anonymisé).

La méthodologie proposée a été appliquée et validée dans le cadre d'un programme d'estimation de la dynamique des véhicules mené pour un important acteur automobile. L'étude s'est concentrée sur l'estimation de l'angle de glissement latéral du véhicule en utilisant uniquement des signaux embarqués de niveau production.

L'angle de glissement latéral est une variable critique dans la dynamique des véhicules. Il décrit la différence entre l'orientation longitudinale du véhicule et sa direction réelle du mouvement. Bien qu'essentiel pour le contrôle de stabilité, les systèmes de sécurité et les fonctions avancées d'assistance à la conduite, il n'est pas directement mesurable dans les véhicules de série sans capteurs dédiés et coûteux. En conséquence, la plupart des solutions industrielles reposent sur des observateurs basés sur des modèles pour l'estimer.

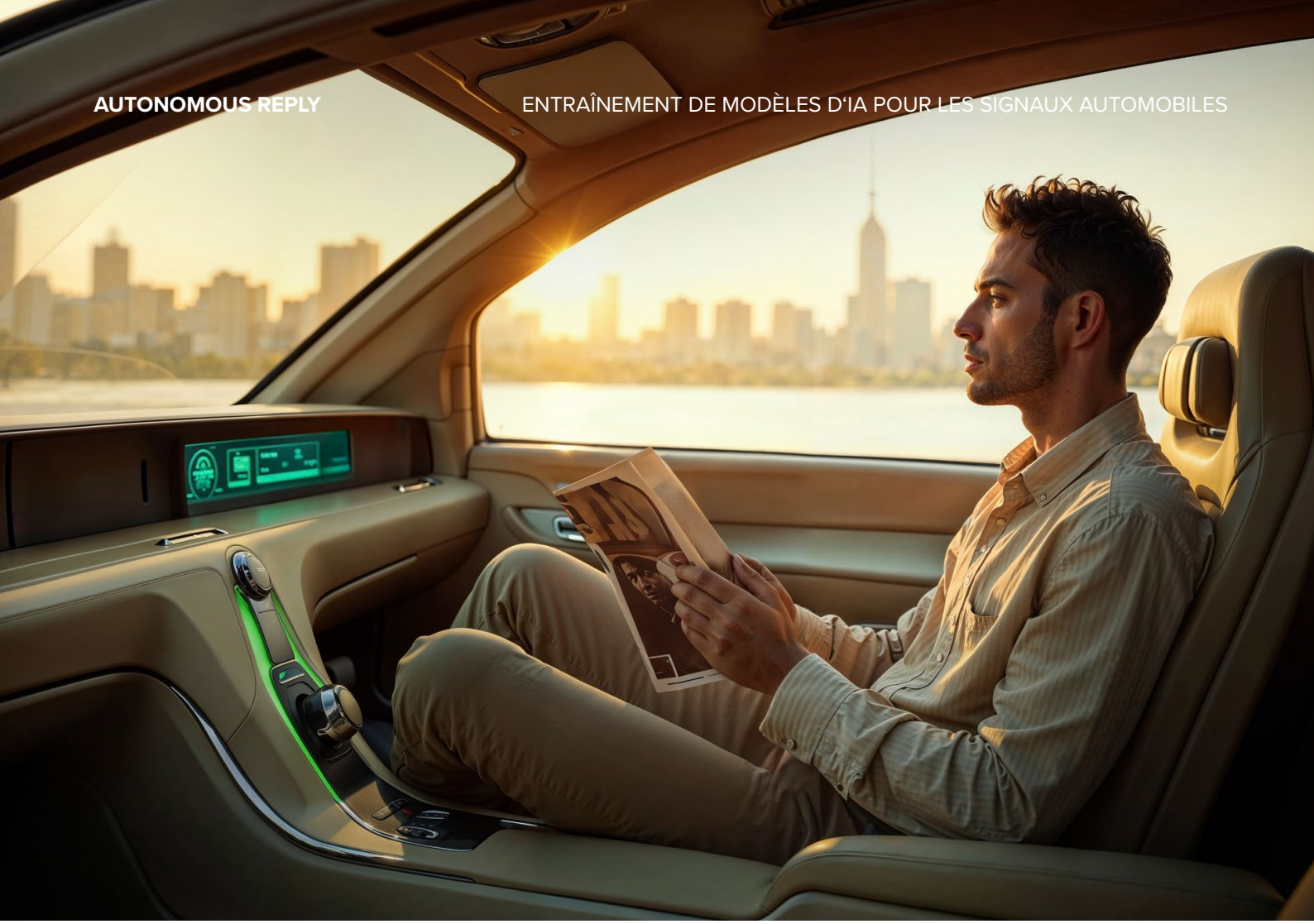
Référence industrielle : estimation basée sur l'observateur

La solution existante du client reposait sur un observateur piloté par la physique dérivé des équations de dynamique des véhicules. Cet observateur intègre l'entrée de direction, la vitesse des roues, les mesures inertielles et les paramètres du véhicule pour reconstruire l'angle de glissement latéral en temps réel.

Les approches basées sur l'observateur offrent plusieurs avantages. Ils sont efficaces en calcul, physiquement interprétables et compatibles avec les environnements automobiles embarqués. Cependant, ils reposent sur des hypothèses concernant le comportement des pneus, les paramètres du véhicule et les conditions de fonctionnement. Leurs performances peuvent se dégrader sous de fortes non-linéarités, des manœuvres agressives ou lorsque les paramètres du système s'écartent de la calibration nominale.

L'objectif de l'étude n'était pas de remplacer cette méthode à l'aveugle, mais d'évaluer si une approche basée sur les données pouvait :

- Égaler ou dépasser sa précision d'estimation
- Rester robuste sous des conditions dynamiques et non linéaires
- Améliorer la sensibilité aux événements rares et extrêmes
- Fournir des analyses complémentaires



Analyse des données et ingénierie du signal

Le projet a débuté par une analyse approfondie des signaux embarqués disponibles. Le jeu de données comprenait des sessions de conduite enregistrées couvrant un large éventail de conditions de fonctionnement, incluant la conduite en régime stationnaire, les manœuvres modérées et les scénarios très dynamiques.

Plutôt que de transmettre tous les signaux disponibles directement dans les modèles, nous avons mené une analyse structurée des données exploratoires. Cela comprenait des études de corrélation, l'analyse de variance et des techniques de réduction de dimensionnalité telles que l'analyse en composantes principales.

L'objectif était de comprendre :

- Redondance entre signaux
- Contribution relative à l'information
- Sensibilité des signaux à la dynamique latérale
- Risque de multicolinéarité

Cette évaluation du signal basée sur les données a été combinée à une expertise du domaine en dynamique des véhicules pour identifier un ensemble de caractéristiques compact et physiquement significatif. L'objectif n'était pas simplement la réduction de la dimensionnalité, mais l'ingénierie structurée des caractéristiques fondée à la fois sur la physique et les statistiques.

Les signaux sélectionnés ont ensuite été

organisés dans un format expérimental cohérent, adapté à la modélisation séquentielle. Un pipeline de prétraitement a été mis en place pour garantir la reproductibilité entre les expériences. La normalisation par caractéristiques a été appliquée pour tenir compte des différences d'échelle entre les signaux, assurant une optimisation stable et évitant le biais en faveur des entrées de grande intensité.

STRATÉGIE DE BENCHMARKING MULTI-MODÈLES

Pour évaluer correctement la valeur ajoutée des architectures basées sur l'attention, l'étude a suivi une stratégie progressive de benchmarking à travers plusieurs familles de modèles.

1. Modèles classiques d'apprentissage automatique

Comme première base, des modèles classiques de régression ont été mis en œuvre. Cela incluait la régression linéaire et les méthodes d'ensemble basées sur des arbres. L'objectif était d'évaluer jusqu'où les modèles non séquentiels pouvaient aller lorsqu'ils disposaient de fenêtres temporelles conçues.

Parmi ceux-ci, les arbres de décision boostés par gradient, y compris les implémentations XGBoost, ont démontré de fortes capacités d'approximation non linéaire. Cependant, ces modèles ne raisonnent pas nativement dans le temps.

Les informations temporelles doivent être encodées manuellement via des fenêtres glissantes ou des éléments agrégés, ce qui peut limiter l'expressivité lors de la modélisation de dépendances à longue portée.

Ces modèles fournissaient une référence utile pour la limite inférieure en termes de performances réalisables sans modélisation profonde des séquences.

2. Réseaux neuronaux récurrents (LSTM et GRU)

La deuxième catégorie de modèles a introduit le traitement de séquences natif via des réseaux neuronaux récurrents. Les architectures LSTM et GRU ont toutes deux été mises en œuvre et évaluées.

Ces modèles traitent les séries temporelles séquentiellement, maintenant un état interne caché qui évolue au fil du temps. Ils sont largement utilisés dans les applications automobiles de séries temporelles et représentent une base solide en apprentissage profond.

Différentes longueurs de séquence et configurations architecturales ont été explorées pour évaluer :

- Sensibilité aux manœuvres transitoires
- Capacité à capturer des dépendances temporelles plus longues
- Stabilité sur des suites étendues
- Robustesse face aux changements de régime dynamiques .

Bien que les modèles récurrents aient montré des améliorations nettes par rapport aux références classiques de l'apprentissage automatique, certains comportements ont été observés lors d'événements très dynamiques. En particulier, leur tendance à lisser les prédictions sous des variations extrêmes du signal est devenue un point important d'analyse

3. Modélisation basée sur l'attention : transformateur à fusion temporelle

La famille de modèles finale évaluée était une architecture basée sur l'attention : **le Temporal Fusion Transformer.**

Contrairement aux modèles récurrents qui compressent l'information en un vecteur d'état caché, l'architecture Transformer **calcule des poids d'attention explicites sur des pas de temps**. Cela permet au modèle de déterminer dynamiquement quelles informations historiques sont les plus pertinentes à chaque étape de prédiction.

L'architecture intègre également des mécanismes de sélection de variables, lui permettant d'apprendre quels signaux comptent le plus dans différents contextes de conduite. Cette capacité est particulièrement pertinente en dynamique des véhicules, où l'importance de la direction, du taux de lacet ou de l'accélération peut varier considérablement selon la manœuvre.

Le modèle a été configuré pour fonctionner sur des segments de conduite séquentiels, cohérents avec les bases récurrentes, assurant une comparaison équitable entre les architectures.

Protocole d'entraînement et de validation

Un axe clé de l'étude était de garantir une évaluation généralisée significative. Plutôt que de diviser aléatoirement les échantillons de temps, le jeu de données était partitionné au niveau de la session de conduite. Des séquences de conduite entières étaient réservées à la validation afin de simuler des conditions de déploiement réelles.

Cette approche a permis de tester les modèles sur des contextes de conduite invisibles plutôt que sur de simples horodatages invisibles de la même trajectoire.

Une attention particulière a été portée aux scénarios rares mais pertinents pour la sécurité dans l'ensemble de données. Cela inclut des manœuvres agressives et des dynamiques latérales élevées, statistiquement rares mais stratégiques sur le plan opérationnel. Le cadre d'évaluation a donc pris en compte non seulement l'erreur moyenne de prédiction, mais aussi le comportement lors de régimes dynamiques extrêmes.

Toutes les familles de modèles (basées sur l'observateur, apprentissage automatique classique, réseaux de neurones récurrents et architectures basées sur Transformer) ont été évaluées selon ce cadre expérimental cohérent.

RÉSULTATS

Tous les modèles ont été évalués selon la même configuration de validation à l'aide de séances de conduite invisibles. La performance a été évaluée à l'aide de RMSE, MAE et R^2. En plus des métriques globales, nous avons inspecté visuellement le comportement de prédiction sur toute la plage dynamique.

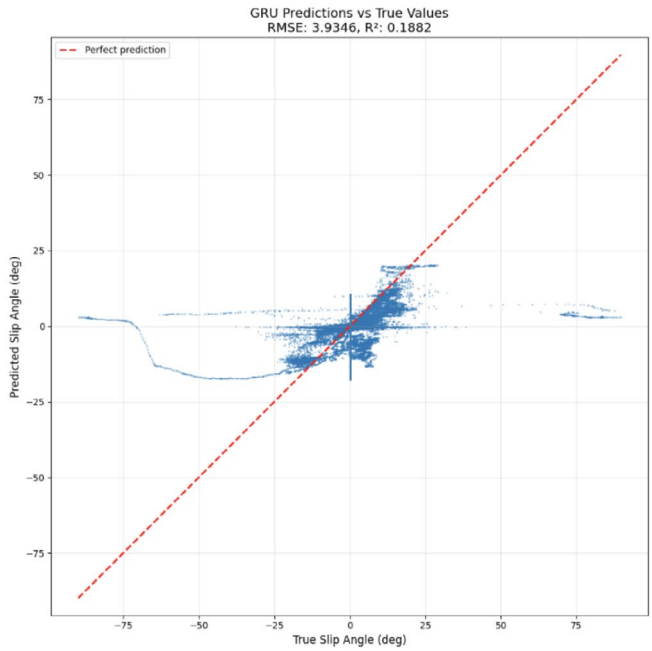
Le tableau ci-dessous résume la performance relative des différentes approches.

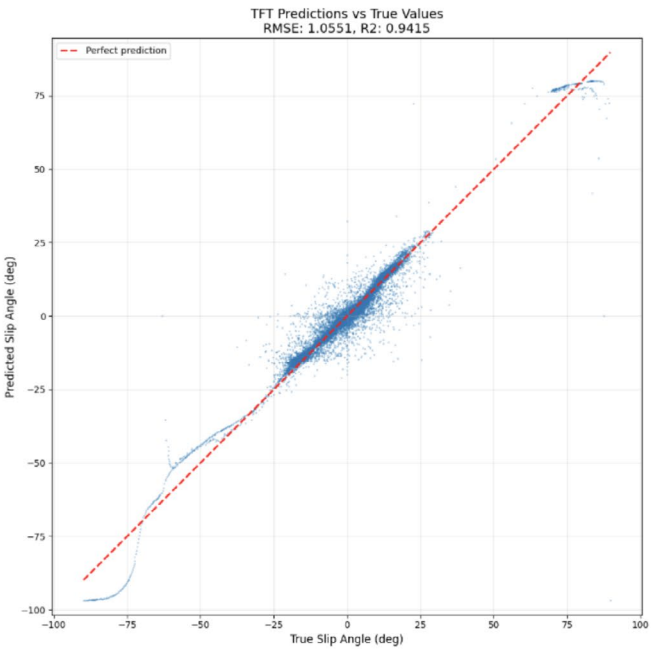
Famille de modèles	Exemples	Notes	Performance relative
Estimateurs classiques	Filtres de Kalman, observateurs	Fort biais inductif ; peut avoir des difficultés avec des non-linéarités complexes / événements rares	Rapide, interprétable et fiable lorsque le modèle physique et les paramètres sont précis ; peut se dégrader sous de forts effets non linéaires des pneus, de la dérive des paramètres et des manœuvres rares ou à fort glissement élevé.
Modèles basés sur des arbres	XGBoost	Une base tabulaire forte, rapide pour l'entraînement et l'exécution, nécessite une ingénierie des caractéristiques pour les effets temporels	Compétitif avec les RNN (RMSE ~3,90°, R^2 ~0,20), mais nécessite de l'ingénierie des caractéristiques pour représenter les effets temporels.
Mannequins récurrents	LSTM/GRU	Bon pour les schémas temporels locaux ; peut plafonner sur les dépendances à long terme	Apprenez les schémas temporels à court terme directement à partir des signaux ; dans nos tests, le GRU a légèrement battu le LSTM (RMSE 3,93°, R^2 0,19 contre RMSE 4,12°, R^2 0,11) mais les deux ont généralisé modestement et ont eu tendance à atteindre leur pic précoce.
Série temporelle basée sur l'attention Transformer	Transformateur de fusion temporelle	Apprend l'alignement à travers le temps (et les capteurs) et évolue avec les données/params	Capture les dépendances entre signal croisé et temps ; dans nos tests, il a clairement surpassé les RNN (RMSE 1,06°, R^2 0,94). L'interprétabilité (importance variable) est optionnelle et peut être lente.
Benchmark interne client	Référence propriétaire	Point de référence utilisé par le client	Observateur d'état déterministe basé sur un modèle utilisant le gain de rétroaction de sortie ; faible en calcul et adapté à l'ingénierie, mais la précision dépend de la fidélité du modèle et de l'ajustement selon les régimes.

La méthode basée sur l'observateur fournissait une référence de référence forte et fiable. Les modèles classiques d'apprentissage automatique tels que **XGBoost** présentaient des performances compétitives mais reposaient sur l'ingénierie manuelle des caractéristiques temporelles. Les modèles récurrents (LSTM et GRU) ont amélioré l'apprentissage temporel mais ont montré des limites dans la capture de dynamiques extrêmes.

Le modèle de série temporelle basé sur Transformer a obtenu la meilleure performance globale parmi les approches apprises.

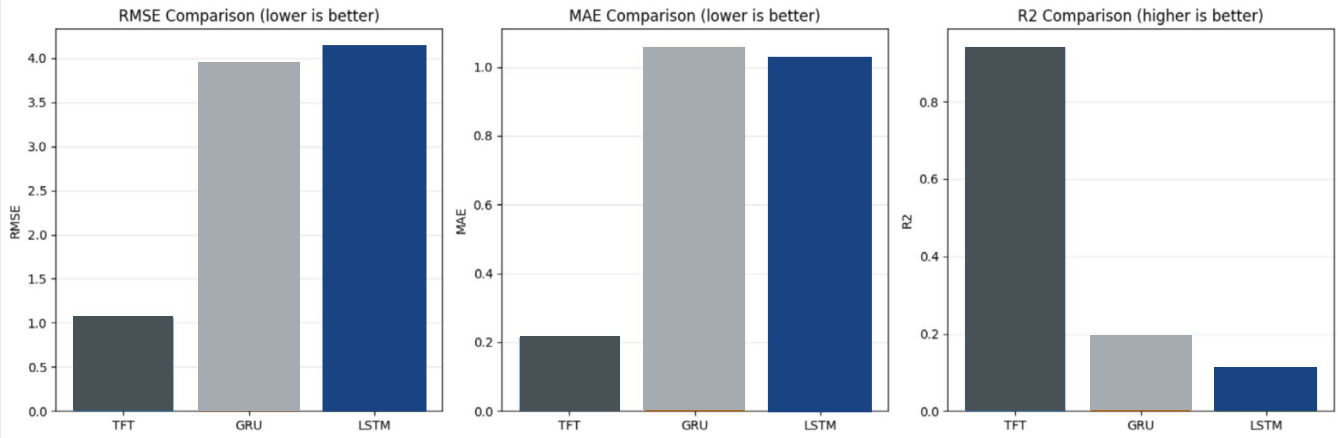
Les graphiques suivants montrent les valeurs prédites versus vraies pour les données de validation représentatives.





Les modèles récurrents tendent à lisser les valeurs extrêmes et à sous-prédire les régions de grande intensité. En revanche, le modèle basé sur le Transformer suit plus étroitement les valeurs nominales et extrêmes, avec une dispersion réduite autour de la ligne de prédiction idéale.

Les métriques agrégées à travers l'ensemble de validation sont présentées ci-dessous.



Les métriques agrégées à travers l'ensemble de validation sont présentées ci-dessous. Le modèle basé sur le transformateur obtient une erreur plus faible (RMSE, MAE) et une puissance explicative (R^2) plus élevée comparé au LSTM et au GRU. L'amélioration est particulièrement visible lors des segments très dynamiques, où la modélisation basée sur l'attention maintient la sensibilité aux transitions nettes.

APPROCHE DE LA RÉALITÉ DE PRODUCTION

Former un modèle performant dans un environnement contrôlé est une étape importante. Le déployer dans un système automobile embarqué est un autre défi totalement différent. Combler cet écart nécessite de s'attaquer à des contraintes rarement visibles lors de la recherche et de l'expérimentation.

La première considération majeure est l'**outillage** et la **portabilité** des modèles. Les architectures basées sur des transformateurs sont généralement développées dans des environnements Python utilisant des cadres d'apprentissage profond de haut niveau. Passer de cet écosystème à une chaîne d'outils automobile intégrée introduit des frictions. Les pipelines d'exportation de modèles doivent être robustes, compatibles avec les temps d'exécution d'inférence et validés sur plusieurs représentations intermédiaires. La conversion de modèles en formats portables tels qu'**ONNX** peut révéler des opérateurs non pris en charge, des incompatibilités de versions ou des composants architecturaux qui ne se traduisent pas correctement dans des environnements embarqués. Ces problèmes n'invalident pas l'approche, mais ils nécessitent une ingénierie et une validation soigneuses pour garantir que le modèle entraîné se comporte de manière identique une fois déployé.

La **latence** est la deuxième contrainte clé. Dans l'estimation de la dynamique des véhicules, les prévisions doivent être produites dans des budgets stricts en temps réel. Contrairement à l'expérimentation hors ligne, où le temps d'inférence est secondaire, le déploiement embarqué exige un temps d'exécution déterministe et borné.

Les architectures transformateurs, bien que puissantes, introduisent une surcharge computationnelle supplémentaire par rapport aux observateurs légers ou aux modèles peu profonds. Les mécanismes d'attention s'adaptent à la longueur de la séquence et à la taille du modèle, rendant les choix de conception architecturale essentiels. La taille de la fenêtre de séquence, le nombre de têtes d'attention et les dimensions cachées doivent être sélectionnés non seulement pour la précision mais aussi pour l'efficacité de l'exécution.

L'**empreinte mémoire** est une autre limitation pratique. Les plateformes automobiles embarquées fonctionnent sous des budgets limités de RAM et de stockage flash. Un modèle performant dans un environnement de station de travail en développement peut dépasser les limites de mémoire d'une unité de contrôle électronique.

Un **profilage minutieux** est donc nécessaire pour évaluer le nombre de paramètres, le stockage d'activation intermédiaire et les schémas d'allocation à l'exécution.

Pour répondre à ces contraintes, les techniques de compression des modèles deviennent essentielles. Plusieurs stratégies peuvent être envisagées. **La taille** peut réduire la redondance des paramètres en supprimant les poids de faible importance après l'entraînement. **La quantification** peut convertir les représentations en virgule flottante en formats de moindre précision, réduisant significativement la consommation de mémoire et accélérant l'inférence sur du matériel compatible. La distillation de connaissances offre une autre voie, où un modèle étudiant plus petit est entraîné à reproduire le comportement d'un Transformer plus grand tout en conservant la plupart de ses avantages en termes de performance. Ces techniques introduisent des compromis entre compacité et précision d'estimation, et doivent être évaluées avec soin dans un contexte de sécurité pertinent.

Au-delà des considérations computationnelles, le déterminisme et les **exigences de validation** jouent un rôle crucial. Les systèmes automobiles exigent un comportement prévisible dans toutes les conditions de fonctionnement. Cela signifie vérifier que les sorties du modèle restent stables face aux changements de précision numérique, aux entrées extrêmes et aux cas frontières. Il est important de noter qu'aucune de ces contraintes ne rend les architectures basées sur l'attention impraticables. Ils déplacent simplement le problème de la modélisation pure vers l'ingénierie au niveau système. Avec un réglage architectural approprié, des longueurs de séquence contrôlées et des stratégies de compression ciblées, les estimateurs basés sur Transformer peuvent être adaptés aux environnements embarqués. Le chemin du prototype de recherche au système de production n'est pas trivial, mais il est techniquement faisable lorsqu'il est abordé méthodiquement.

En fin de compte, déployer un modèle d'IA dans un contexte automobile ne se limite pas à la performance prédictive. Il s'agit d'**équilibrer la précision, la latence, l'empreinte mémoire, le déterminisme et la maintenabilité** au sein des réalités des systèmes embarqués.

LA VISION D'ENSEMBLE

Les architectures transformers ont profondément transformé l'intelligence artificielle. Les modèles Foundation entraînés sur d'immenses corpus de texte, d'images et d'audio ont démontré que les architectures basées sur l'attention peuvent apprendre des représentations riches et généraliser à travers les tâches à une échelle sans précédent. Les grands modèles de langage, les transformateurs de vision et les systèmes multimodaux servent désormais de base à de nombreux produits pilotés par l'IA.

Pourtant, cette transformation s'est principalement concentrée sur les données non structurées : **langage, pixels et flux audio.**

Les systèmes industriels fonctionnent dans un régime fondamentalement différent. Ils produisent des signaux structurés, pilotés par la physique, qui évoluent au fil du temps et sont étroitement liés à la dynamique mécanique, électrique ou aérodynamique. Ces signaux sont contraints par des exigences de sécurité, des limitations en temps réel et des réalités intégrées au déploiement. Leurs propriétés statistiques ne sont pas façonnées par les schémas du langage humain, mais par les lois physiques et les interactions des systèmes.

Appliquer des modèles de fondation génériques à de tels domaines n'est ni suffisant ni approprié. Ce qui est requis, ce sont des modèles de fondation spécifiques à un domaine pour les signaux : des architectures conçues pour raisonner sur des données de séries temporelles structurées, capables de capturer des dynamiques non linéaires, des événements critiques rares et des dépendances temporelles à longue portée.

L'approche présentée dans cet ouvrage reflète ce changement plus large. Plutôt que de construire des applications uniquement autour d'agents sur de grands modèles de langage, l'accent se porte sur l'entraînement de modèles basés sur l'attention directement sur les signaux de domaine. Cela représente une évolution complémentaire de l'intelligence artificielle : du raisonnement sur les mots et les images au raisonnement sur les systèmes dynamiques.

À mesure que les industries deviennent de plus en plus définies par logiciel, la capacité à entraîner et déployer des modèles natifs du signal robustes, interprétables et efficaces deviendra un facteur de différenciation stratégique. La prochaine génération d'IA ne comprendra pas seulement le langage, elle comprendra aussi les machines.

OÙ CELA S'APPLIQUE ENSUITE

Bien que ce travail se soit concentré sur un défi d'estimation spécifique en dynamique des véhicules, la méthodologie sous-jacente est largement applicable aux systèmes intelligents modernes.

Dans les véhicules définis par **logiciel (SDV)**, la détection virtuelle et l'estimation avancée de l'état deviennent des capacités centrales. À mesure que les véhicules évoluent vers des architectures de plus en plus axées sur les logiciels, les modèles de séries temporelles basés sur l'attention peuvent permettre une surveillance améliorée, une redondance et une maintenance prédictive sans nécessiter de capteurs physiques supplémentaires.

Au sein des stacks d'autonomie, l'estimation robuste de l'état interne reste fondamentale. Les systèmes de perception, de planification et de contrôle reposent sur des représentations précises et stables de la dynamique des systèmes. Les estimateurs basés sur transformateurs offrent un cadre flexible pour modéliser les comportements non linéaires et gérer des dépendances temporelles complexes, en particulier dans des conditions de cas particuliers.

Les mêmes principes s'appliquent naturellement à **la robotique** et aux **systèmes aériens** sans pilote, où l'intelligence intégrée doit interpréter des signaux en évolution rapide sous des budgets de calcul contraignés.

Dans les **environnements aérospatiaux** et **industriels**, estimer les états cachés à partir de mesures indirectes est un défi récurrent. Les architectures basées sur l'attention offrent un moyen évolutif de résoudre ces problèmes d'estimation tout en maintenant une adaptabilité aux conditions de fonctionnement évolutives.

Une orientation particulièrement prometteuse réside dans les approches hybrides qui combinent des modèles basés sur la physique avec un apprentissage basé sur les données. Plutôt que de remplacer les méthodes établies, les estimateurs basés sur l'attention peuvent les compléter, renforçant la robustesse dans des régimes où les modèles purement analytiques peinent tout en préservant structure et interprétabilité.

Plus largement, ce travail illustre comment les techniques modernes de modélisation de séquences peuvent dépasser les domaines traditionnels et devenir des éléments fondamentaux dans des systèmes dynamiques en temps réel critiques pour la sécurité.

APPEL À L'ACTION

Si vous vous attaquez à des problèmes de séries temporelles à enjeux élevés et souhaitez explorer des modèles de signaux basés sur l'attention, **contactez-nous pour discuter d'un benchmark, d'un atelier ou d'un pilote.**


REPLY
AUTONOMOUS