

Stage M2 : Transformers frugaux par sélection adaptative des têtes d'attention et des tokens

Contexte : Développement de mécanismes adaptatifs permettant d'ajuster le coût de calcul des Transformers à la complexité de chaque entrée, dans la continuité de nos travaux sur *BudgetFormer*.

Niveau : M2 / école d'ingénieur

Durée : 5 ou 6 mois

Début : Février ou Mars

Lieu : Laboratoire DAVID, UVSQ

Description du projet

Les Transformers reposent sur des mécanismes d'attention dont le coût dépend notamment du nombre de têtes activées et de la longueur des séquences traitées. Pourtant, toutes les entrées ne présentent pas la même complexité : mobiliser systématiquement l'ensemble des têtes et des tokens peut entraîner des calculs superflus, une consommation mémoire élevée et une empreinte énergétique inutile.

Dans notre travail récent sur **BudgetFormer** [1], nous avons étudié l'allocation dynamique des têtes d'attention en fonction de l'entrée, avec des résultats prometteurs en classification de texte.

Ce stage vise à prolonger cette approche en combinant deux niveaux de sélection : les **têtes d'attention**, qui déterminent les mécanismes de représentation mobilisés, et les **tokens**, qui déterminent la quantité d'information traitée par ces mécanismes. L'objectif est de concevoir un Transformer dont le budget de calcul s'adapte conjointement à la complexité et aux besoins informationnels de chaque entrée.

Au-delà de la classification, le stage explorera le **question-answering** et des tâches de génération, pour lesquelles la sélection doit préserver les informations fines nécessaires à la production de réponses pertinentes.

Objectifs détaillés

1. Sélection conjointe des têtes et des tokens

Concevoir un mécanisme adaptatif qui choisit, pour chaque entrée, les têtes d'attention à activer et les tokens à conserver. Étudier différentes stratégies de sélection et leur apprentissage conjoint, en tenant compte du compromis entre qualité des prédictions et coût computationnel.

2. Représentation et apprentissage de séquences variables

Étudier les difficultés liées à la représentation tensorielle de séquences dont les longueurs diffèrent après sélection. Explorer des solutions compatibles avec l'entraînement par lots, telles que les masques dynamiques, le regroupement de séquences de tailles comparables et les représentations à longueur variable. L'objectif est de limiter les calculs induits par le remplissage artificiel (*padding*) et de rendre la sélection effectivement avantageuse sur le plan matériel.

3. Extension à des tâches fines et génératives

Évaluer les stratégies proposées au-delà de la classification, notamment sur des tâches de question-answering et de génération. Étudier comment préserver les informations locales et les dépendances à longue portée, tout en réduisant le nombre de tokens et de têtes mobilisés.

4. Évaluation de la frugalité

Comparer les modèles proposés aux architectures de référence selon plusieurs critères : qualité des résultats, nombre d'opérations (FLOPs), mémoire utilisée, latence et débit d'inférence. Lorsque les outils le permettent, estimer également la consommation énergétique afin de caractériser les gains réels au-delà des seuls indicateurs théoriques.

Contexte scientifique et état de l'art

Plusieurs familles de travaux cherchent à réduire le coût des Transformers. BigBird [2] utilise une attention clairsemée pour traiter des séquences longues avec une complexité réduite. D'autres approches s'intéressent directement à la réduction du nombre de tokens. DynamicViT [3], par exemple, élimine progressivement les tokens jugés peu informatifs dans les Transformers de vision. Learned Token Pruning [4] et MoH [5] étudient également une suppression adaptative des tokens pour réduire le coût de calcul.

La sélection conjointe des tokens et des têtes a déjà été explorée pour le traitement du langage, notamment par Lee et al. [6]. Par ailleurs, Diwan et al. [7] montrent que l'efficacité relative des variantes de Transformers dépend de la modalité, de la longueur des entrées et de la métrique de performance considérée.

Ces travaux établissent l'intérêt de l'adaptation du calcul, mais plusieurs questions restent ouvertes pour le cadre étudié : comment coordonner les décisions de sélection des têtes et des tokens en fonction d'une même entrée, comment apprendre cette politique malgré les contraintes de taille des tenseurs [8], et comment préserver les informations nécessaires aux tâches fines et génératives ?

Le projet s'appuie sur **BudgetFormer** [1] pour étudier ces questions dans un cadre unifié. L'ambition est de dépasser une réduction isolée du nombre de têtes ou de tokens pour apprendre une allocation conjointe du calcul, adaptée à chaque entrée et à la tâche considérée.

Verrous scientifiques

- **Couplage des décisions de sélection** : la réduction du nombre de têtes et celle du nombre de tokens ne sont pas indépendantes. Une tête pertinente pour une séquence complète ne l'est pas nécessairement pour une séquence raccourcie. Il faut donc apprendre des décisions coordonnées plutôt que juxtaposer deux mécanismes de pruning.
- **Apprentissage avec des représentations de tailles variables** : la sélection dynamique produit des séquences de longueurs différentes, difficiles à regrouper efficacement dans des tenseurs de taille fixe. Le recours au padding peut annuler une partie des gains attendus. Il faut étudier des stratégies d'apprentissage et d'exécution qui préservent l'efficacité du calcul réellement effectué.
- **Préservation de l'information fine** : en classification, une représentation globale peut suffire à produire une étiquette. En question-answering et en génération, la réponse peut dépendre d'un token rare, d'un détail local ou d'une dépendance éloignée. Une sélection trop agressive risque donc de supprimer une information essentielle. Il faut définir des critères de sélection adaptés à la tâche et mesurer précisément le compromis entre frugalité et qualité.
- **Gains théoriques et gains réels** : une diminution du nombre de FLOPs ne garantit pas toujours une accélération sur GPU ou CPU. L'irrégularité des séquences, les opérations de sélection et les accès mémoire peuvent introduire des surcoûts. L'évaluation devra distinguer complexité théorique, latence mesurée et consommation effective des ressources.

Méthodologie et tâches envisagées

- Revue bibliographique sur la sélection des têtes, le pruning de tokens, les Transformers efficaces et l'optimisation des modèles génératifs.
- Reproduction et analyse du mécanisme de sélection des têtes de *BudgetFormer*, puis mise en place de références avec sélection des tokens seule et des têtes seule.
- Conception d'une stratégie de sélection conjointe, avec étude de plusieurs mécanismes de score, de budget de calcul et de régularisation.
- Expérimentation de solutions pour les longueurs variables : masques dynamiques, regroupement par longueur et réduction du surcoût dû au padding.

- Évaluation progressive sur la classification, puis sur des tâches de question-answering et, selon l'avancement, de génération.
- Comparaison des performances, de la latence, des FLOPs, de la mémoire et, si possible, de l'énergie consommée. Analyse des cas où la sélection dégrade la qualité et des compromis obtenus.

Profil recherché

- Étudiant en M2 ou en dernière année d'école d'ingénieur en intelligence artificielle, apprentissage automatique ou informatique.
- Bonne maîtrise de Python et d'un framework d'apprentissage profond, idéalement PyTorch.
- Connaissances en Transformers, mécanismes d'attention et traitement du langage naturel.
- Intérêt pour la recherche expérimentale, l'optimisation des modèles et l'analyse rigoureuse des performances.

Encadrement et modalités pratiques

- **Bilal FAYE** : maître de conférences, laboratoire DAVID, UVSQ.
- **Tom DEVYNCK** : doctorant, laboratoire DAVID, UVSQ.
- **Mustapha LEBBAH** : professeur, laboratoire DAVID, UVSQ.
- **Hanane AZZAG** : professeure, laboratoire LIPN, Université Sorbonne Paris Nord.
- **Durée** : 5 ou 6 mois.
- **Date de début** : Février ou Mars.
- **Lieu principal** : laboratoire DAVID, UVSQ.
- **Contact** : uvsqai@gmail.com.

Dossier de candidature

- CV à jour ;
- relevés de notes récents ;
- description de projets ou portfolio (facultatif).

Références bibliographiques

- [1] Bilal FAYE, Abdoulaye MBAYE, Hanane AZZAG et Mustapha LEBBAH. “Adaptive Head Budgeting for Efficient Multi-Head Attention”. In : (2026). arXiv : 2604.22583 [cs.LG]. URL : <https://arxiv.org/abs/2604.22583>.
- [2] Manzil ZAHEER et al. “Big Bird : Transformers for Longer Sequences”. In : *Advances in Neural Information Processing Systems*. T. 33. 2020. URL : <https://arxiv.org/abs/2007.14062>.
- [3] Yongming RAO, Wenliang ZHAO, Benlin LIU, Jiwen LU, Jie ZHOU et Cho-Jui HSIEH. “Dynamic-ViT : Efficient Vision Transformers with Dynamic Token Sparsification”. In : *Advances in Neural Information Processing Systems*. T. 34. 2021. URL : <https://arxiv.org/abs/2106.02034>.
- [4] Woosuk KWON, Joseph HASSOUN et Kurt KEUTZER. “Learned Token Pruning for Transformers”. In : *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022, p. 784-794. DOI : 10.1145/3534678.3539260.

- [5] Peng JIN, Bo ZHU, Li YUAN et Shuicheng YAN. “MoH : Multi-head Attention as Mixture-of-Head Attention”. In : *arXiv preprint arXiv :2410.11842* (2024).
- [6] Chonghan LEE, Md Fahim Faysal KHAN, Rita BRUGAROLAS BRUFAU, Ke DING et Vijaykrishnan NARAYANAN. “Token and Head Adaptive Transformers for Efficient Natural Language Processing”. In : *Proceedings of the 29th International Conference on Computational Linguistics*. 2022. URL : <https://aclanthology.org/2022.coling-1.404/>.
- [7] Anuj DIWAN, Eunsol CHOI et David HARWATH. “When to Use Efficient Self Attention ? Profiling Text, Speech and Image Transformer Variants”. In : (2023). arXiv : 2306.08667 [cs.LG]. URL : <https://arxiv.org/abs/2306.08667>.
- [8] Daniel BOLYA, Cheng-Yang FU, Xiaoliang DAI, Peizhao ZHANG, Christoph FEICHTENHOFER et Judy HOFFMAN. “Token merging : Your vit but faster”. In : *arXiv preprint arXiv :2210.09461* (2022).

Perspectives : ce stage pourra contribuer au développement de Transformers plus adaptatifs et moins coûteux, capables d’allouer leurs ressources de calcul en fonction de l’entrée et de la tâche. Il pourra également ouvrir la voie à des travaux de recherche sur l’efficacité des modèles de langage et des systèmes génératifs.