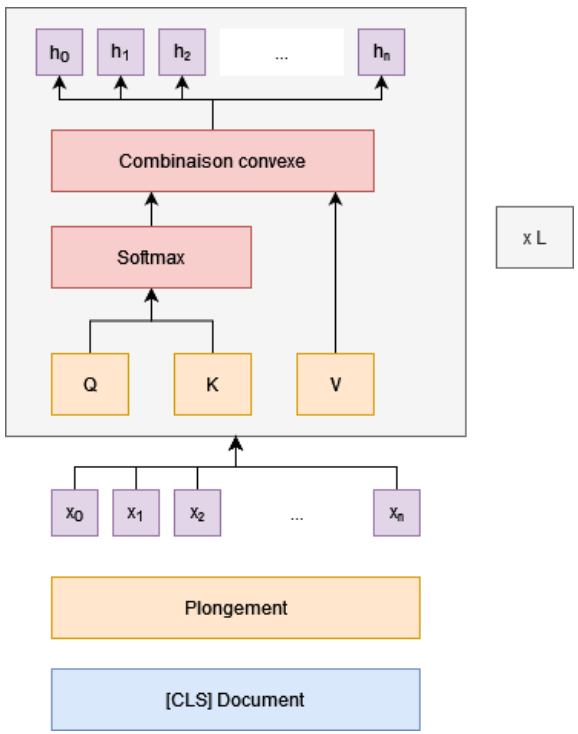
Géométrie de l'auto-attention en classification : quand la géométrie remplace l'attention

Loïc FOSSE - Duc-Hau NGUYEN - Guillaume GRAVIER - Pascale SÉBILLOT

Univ. Rennes, CNRS, Inria / IRISA, Campus de Beaulieu, 35042 Rennes

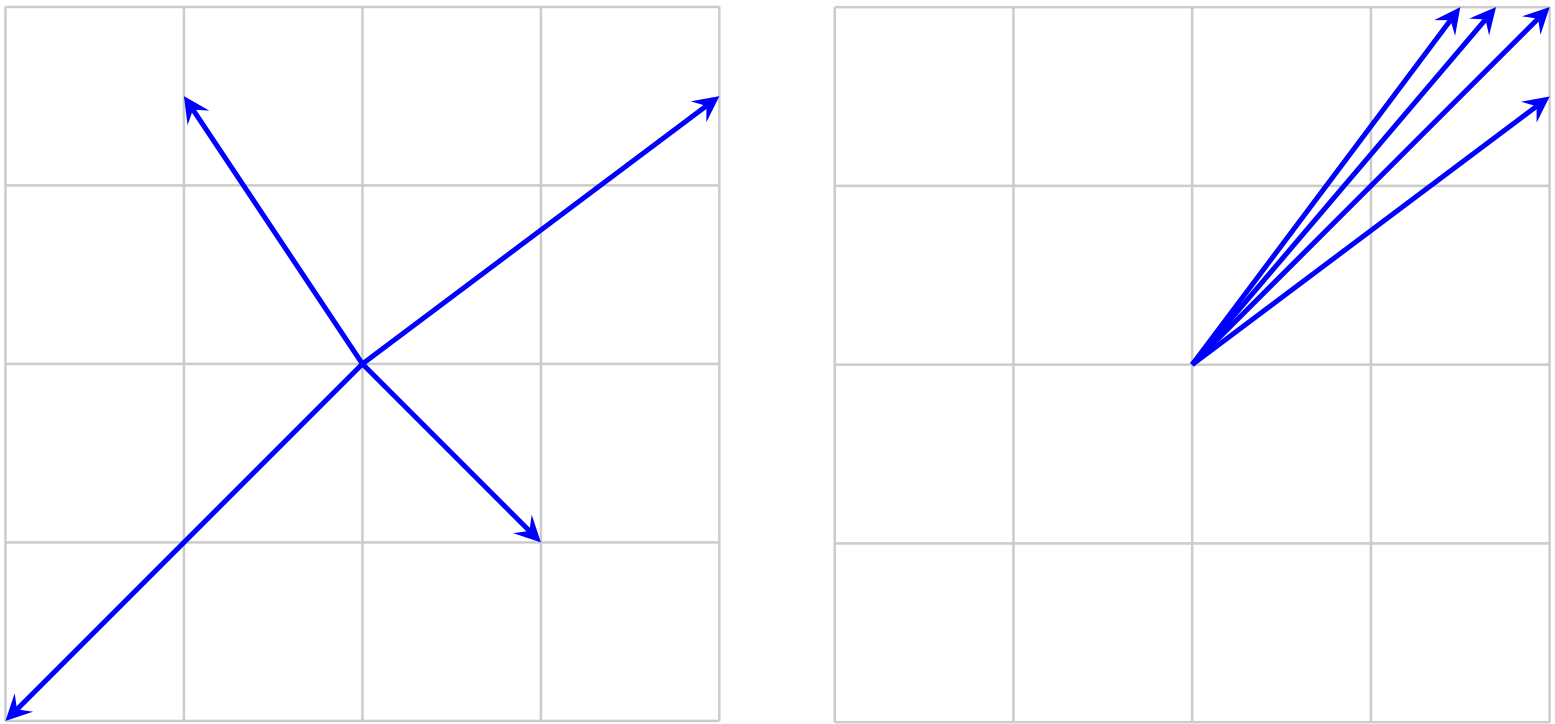
May 28, 2023

Les modèles de type *transformers*



- [Ethayarajh, 2019] étudie les propriétés géométriques des états cachés
- une propriété est mise en avant : **l'anisotropie**

L'anisotropie: concentration des plongements dans une direction de l'espace considéré.



(a) Plongements isotropes

(b) Plongements anisotropes

Figure 1: Exemple visuel de l'anisotropie

L'anisotropie : un phénomène bien connu

- [Ethayarajh, 2019] étudie dans un premier temps ce phénomène sur BERT, GPT2 et ELMo.
- [Fosse et al., 2022] reprend ceci en introduisant une comparaison entre modèles pré-entrainés et les modèles affinés sur des tâches de classification.

Conclusions principales

- **Affiner sur une tâche de classification renforce l'anisotropie.**

Les questions :

- Quel est le lien entre anisotropie et classification ?
- Comment cette anisotropie se construit-elle ?

Les tâches de classification :

- E-SNLI [Camburu et al., 2018] : inférence
- Hatexplain [Mathew et al., 2021] : détection de discours haineux
- YelpHat [Sen et al., 2020] : classification en polarité

Classification sur balise `[CLS]`

1 cône = 1 classe pour BERT ?

L'anysotropie soulève une question en classification : **est ce que la direction du cône dépend de la classe ?**

Dans une tâche de classification :

1. pour chaque classe i on définit r_i : un représentant de la classe i. Un représentant est le plongement `[CLS]` d'une phrase bien classée.

2. Ensuite pour chaque phrase restante du corpus on calcule : $cos([\texttt{CLS}], r_i)$

Résultats sur E-SNLI

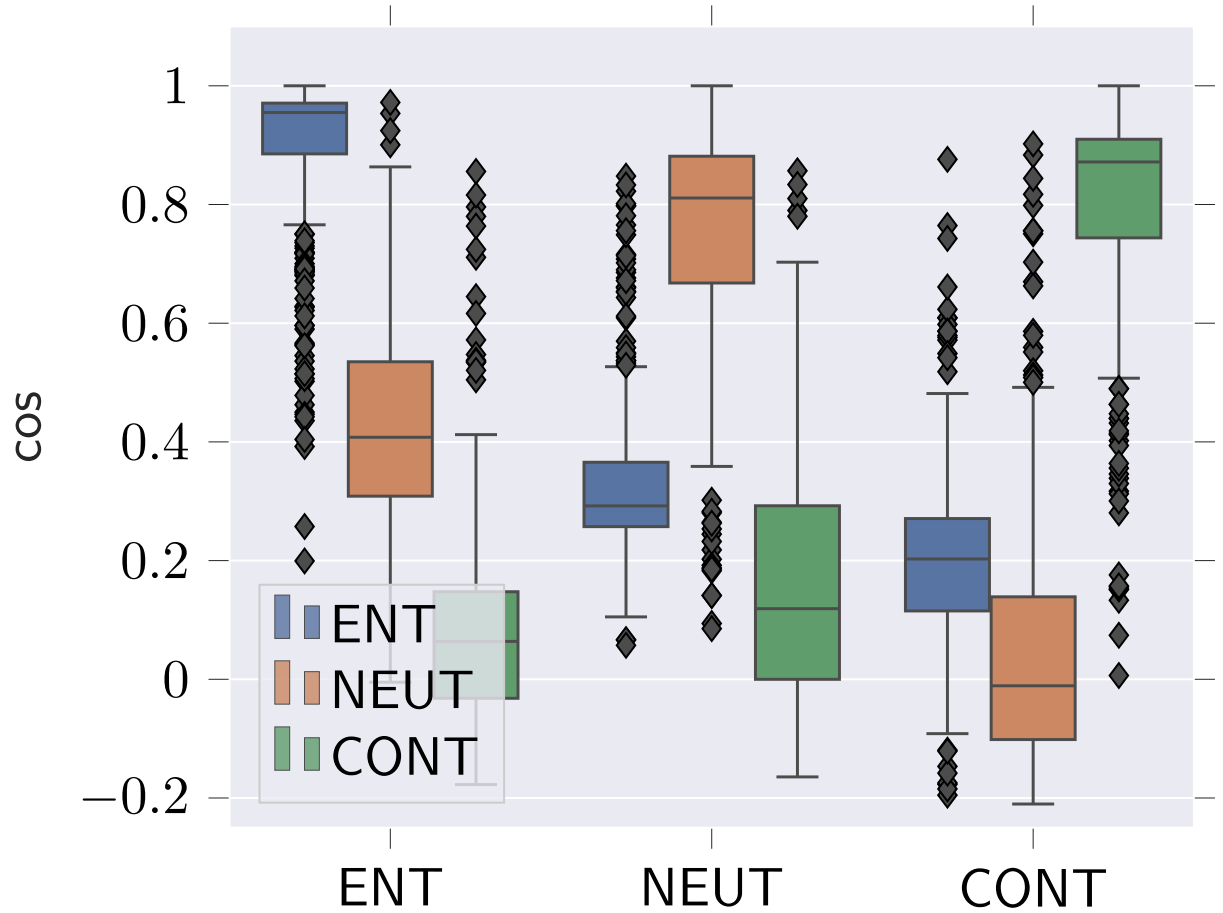


Figure 2: Abscisse: r_i la classe du représentant choisi

Examples sur Yelp-Hat (critiques de restaurants)

phrase	**+/-**	**cos**
Last summer I had an appointment to get new tires and had to wait a super long time. I also went in this week for them to fix a minor problem with a tire they put on. They "fixed" it for free, and the very next morning I had the same issue. I called to complain, and the "manager" didn't even apologize!!! So frustrated. Never going back. They seem overpriced, too.	-	
Been coming to cafe rio for awesome fast Mexican food for years. Lived in Utah for a while...just like you Camilla k. I loved the ones in Utah and thought I would give this one a chance. The manager guy, don't know his name, is a total jerk. The food wasn't even warm. It was cold. My wife and I are sitting here right now and I'm so upset that I have to leave this review right now. Just awful service and not even good food anymore. Gradually getting worse. I'm going to costa vida from now on.	-	0,96
SO GOOD!!!!!!! The only roll I got that wasn't good was a lobster roll. **It just had no flavor**. Everything else I had was AMAZING!!! Now that I'm done raving about the food, **I do have two complaints**. 1) **The hostess wasn't super friendly** or anything. **She was really hard to understand and made no effort** to speak more clearly so we'd know what she was talking about. 2) There's no where near enough tables. The place is an okay size but there's probably only like 10 tables? Maybe I'm remembering wrong. There's also no sushi bar, which I don't care about, but some people do.	+	0,93

Si on regroupe ...

- A-t-on une convergence des plongements dans une direction de l'espace ? **Oui**
- Cette convergence dépend-elle de la classe considérée ? **Oui**, on peut séparer les classes en plusieurs cônes qui sont orthogonnaux.

Conclusion

La direction prise par un plongement dans l'espace semble déterminer la classe.

Le mécanisme d'auto-attention

Projections

$$q_i = (\mathbb{X}\mathbf{Q})_i \quad k_i = (\mathbb{X}\mathbf{K})_i \quad v_i = (\mathbb{X}\mathbf{V})_i \tag{1}$$

Les poids d'attention comme similarités entre requètes et clés

$$a_{ij} = \frac{exp\left(q_i \cdot k_j/\sqrt{d}\right)}{\sum_{l=1}^{n} exp\left(q_i \cdot k_l/\sqrt{d}\right)} \tag{2}$$

Les plongements finaux comme combinaison **convexe** des vecteurs de valeur

$$y_i = \sum_{j=1}^{n} a_{ij} v_j \tag{3}$$

Interprétation géométrique

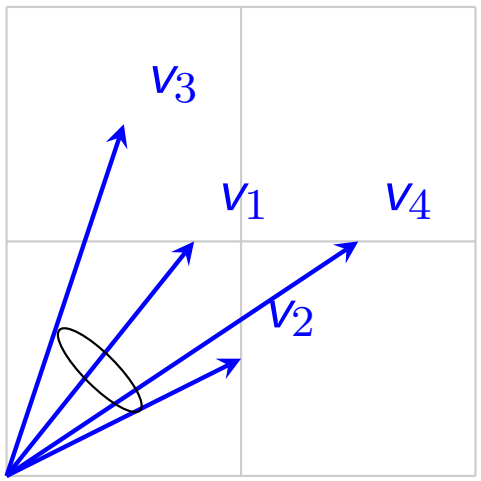


Figure 3: Illustration d'un cône défini par l'ensemble des valeurs v_j

- **la direction prise par un plongement dépend de l'intensité des poids d'attention** a_{ij}.

L'anysotropie : une propriété des modèles transformers

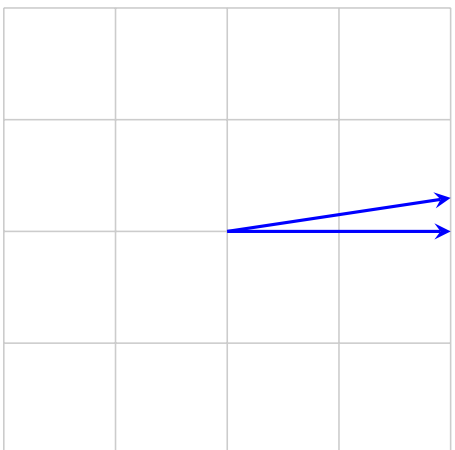
- Retrouve-t-on l'anisotropie à l'intérieur des modèles ? (clés, valeurs)

Pour répondre à cette question:

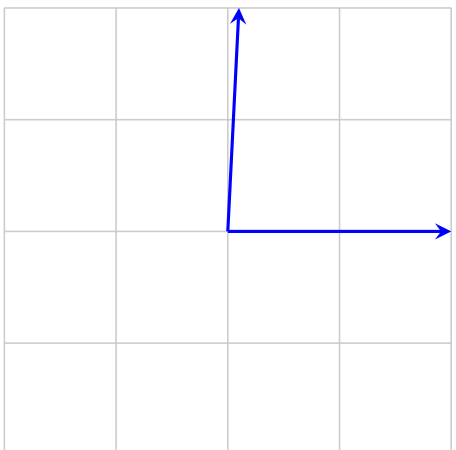
- construction d'un modèle basé sur le mécanisme d'auto-attention.
- mesure de l'anisotropie des **clés** et des **valeurs**

Mesure de conicité [Ethayarajh, 2019]

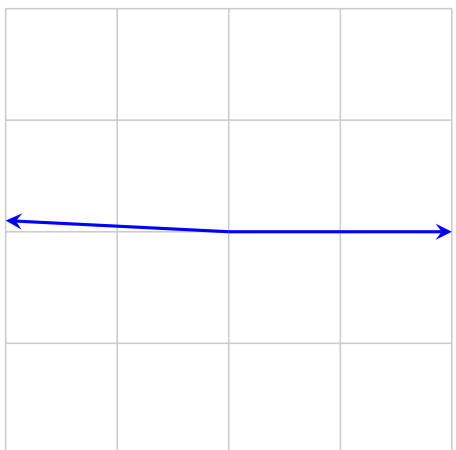
$$\text{SIM}(A) = \frac{2}{n(n-1)} \sum_{i<j} \left[(AA^t) . \frac{1}{\sqrt{\text{diag}(AA^t)\text{diag}(AA^t)^t}} \right] (i,j)$$



(a) $SIM(A) \approx 1$



(b) $SIM(A) \approx 0$



(c) $SIM(A) \approx -1$

Métrique *SIM* sur le modèle d'auto-attention

		HatexPlain					YelpHat					E-SNLI	
		l=1	**l=2**	**l=3**	**l=4**	**l=5**	**l=1**	**l=2**	**l=3**	**l=4**	**l=5**	**l=1**	**l=2**
A=K	L=1	0.713	-	-	-	-	0.578	-	-	-	-	0.657	-
	L=2	0.691	0.683	-	-	-	0.649	0.556	-	-	-	0.719	0.489
	L=3	0.698	0.688	0.841	-	-	0.597	0.431	0.537	-	-	-	-
	L=4	0.614	0.620	0.748	0.840	-	0.714	0.717	0.702	0.860	-	-	-
	L=5	0.624	0.647	0.777	0.913	0.973	0.584	0.542	0.761	0.816	0.959	-	-
A=V	L=1	0.542	-	-	-	-	0.372	-	-	-	-	0.524	-
	L=2	0.510	0.740	-	-	-	0.409	0.511	-	-	-	0.182	0.746
	L=3	0.605	0.688	0.88	-	-	0.461	0.371	0.494	-	-	-	-
	L=4	0.592	0.561	0.785	0.931	-	0.429	0.613	0.803	0.904	-	-	-
	L=5	0.606	0.673	0.858	0.958	0.992	0.417	0.624	0.780	0.9023	0.972	-	-

- l'anisotropie se construit hiérarchiquement au niveau des clés des valeurs

Sur une couche : anistropie faible

		HatexPlain					YelpHat					E-SNLI	
		l=1	**l=2**	**l=3**	**l=4**	**l=5**	**l=1**	**l=2**	**l=3**	**l=4**	**l=5**	**l=1**	**l=2**
	L=1	**0.713**	-	-	-	-	**0.578**	-	-	-	-	**0.657**	-
	L=2	0.691	0.683	-	-	-	0.649	0.556	-	-	-	0.719	0.489
A=K	L=3	0.698	0.688	0.841	-	-	0.597	0.431	0.537	-	-	-	-
	L=4	0.614	0.620	0.748	0.840	-	0.714	0.717	0.702	0.860	-	-	-
	L=5	0.624	0.647	0.777	0.913	0.973	0.584	0.542	0.761	0.816	0.959	-	-
	L=1	**0.542**	-	-	-	-	**0.372**	-	-	-	-	**0.524**	-
	L=2	0.510	0.740	-	-	-	0.409	0.511	-	-	-	0.182	0.746
A=V	L=3	0.605	0.688	0.88	-	-	0.461	0.371	0.494	-	-	-	-
	L=4	0.592	0.561	0.785	0.931	-	0.429	0.613	0.803	0.904	-	-	-
	L=5	0.606	0.673	0.858	0.958	0.992	0.417	0.624	0.780	0.9023	0.972	-	-

- l'anisotropie se construit hiérarchiquement au niveau des clés des valeurs
- les modèles avec peu de couches ne sont pas anisotropes.

Bilan intermédiaire

- Les plongements sont anisotropes.
- La direction du cône détermine la classe.
- Ce cône se construit de manière hiérarchique à travers les couches.
- Les équations de l'attention : **la direction du cône est guidée par les poids d'attention**

$$y_i = \sum_{j=1}^{n} a_{ij} v_j$$

Utilisation des poids d'attention a posteriori

- L'attention comme comme explication
- Problèmes : les poids d'attention sont trop uniformes, [Nguyen et al., 2022]
- Rechercher la parcimonie de l'attention.

Régularisation de [Nguyen et al., 2022]

On ajoute un terme de **pénalisation** (basé sur l'entropie de Shanon) pour rendre les poids d'attention plus parcimonieux.

$$\tilde{\mathcal{L}}(s) = \mathcal{L}(s) + \lambda H(a_{0,j})$$

λ étant un hyperparamètre contrôlant l'intensité de régularisation.

		$\lambda = 0$	$\lambda = 0.0001$	$\lambda = 0.001$	$\lambda = 0.01$	$\lambda = 0.1$
Hatexplain	*H*	0.988	0.31	0.111	0.021	0.01
	Key	0.745	0.669	0.687	0.732	0.727
	Value	**0.578**	**0.637**	**0.782**	**0.828**	**0.772**
E-SNLI	*H*	0.999	0.576	0.64	0.000	0.000
	Key	0.674	0.737	0.795	0.833	0.707
	Value	**0.507**	**0.45**	**0.526**	**0.868**	**0.925**
YelpHat	*H*	0.47	0.48	0.57	0.49	0.07
	Key	0.631	0.657	0.634	0.607	0.672
	Value	**0.303**	**0.391**	**0.375**	**0.366**	**0.483**

Conclusions

- Le rôle de l'anisotropie ? **Les plongements s'organisent dans des structures géométriques semblables à des cônes dont la direction détermine la classe.**
- Comment l'anisotropie se construit ? **Elle se construit directement dans le modèles au niveau des projections intermédiaires.** Oui mais qu'est-ce qui fait que le modèle envoie des plongements dans une direction ?
- Si on vient perturber la distribution des poids d'attention on modifie la direction $\Rightarrow$ le modèles renforce son anisotropie pour compenser, anéantissant les chances d'explication à posteriori.

Discussions

- Sur une tâche d'etiquetage ?
- Sur une tâche de classification à beaucoup (vraiment beaucoup) de classe ? Le modèle construit une direction par classe. Ces directions semblent à premières vues orthogonales. On est en dimension 768 mais est ce que le modèle utilise 768 dimensions, et si non est ce que nous avons une borne supérieur du nombre de classe que on peut séparer ?

Merci à tous

- loic.fosse@insa-rennes.fr / loic.fosse@orange.com
- guig@irisa.fr
- duc-hau.nguyen@irisa.fr
- Pascale.Sebillot@irisa.fr


Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. (2018).
e-SNLI: Natural language inference with natural language explanations.
In *Advances in Neural Information Processing Systems*, volume 31, pages 9539–9549.


Ethayarajh, K. (2019).
How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings.
In *2019 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 55–65.


Fosse, L., Nguyen, D., Sébillot, P., and Gravier, G. (2022).

Une étude statistique des plongements dans les modèles transformers pour le français.
In *29th Conference Traitement Automatique des Langues Naturelles*, pages 247–256.


Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021).
HateXplain: A benchmark dataset for explainable hate speech detection.
In *35th AAAI Conference on Artificial Intelligence*, pages 14867–14875.


Nguyen, D., Gravier, G., and Sébillot, P. (2022).
Filtrage et régularisation pour améliorer la plausibilité des poids d'attention dans la tâche d'inférence en langue naturelle.
In *29th Conference Traitement Automatique des Langues Naturelles*, pages 95–103.


Sen, C., Hartvigsen, T., Yin, B., Kong, X., and Rundensteiner, E. (2020).
Human attention maps for text classification: Do humans and neural networks focus on the same words?
In *60th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608.