Géométrie des vecteurs de tâches pour l'association et la combinaison de modèles

Loïc Fosse

Orange Innovation, Lannion, France
Aix Marseille Univ., CNRS, LIS, Marseille, France

loic.fosse@orange.com

Introduction et motivations

Les modèles de langues sont de plus en plus grands $\Rightarrow$ leur adaptation est de plus en plus coûteuse. Pour palier à cela :

- adaptations efficaces (adapteurs, *prefix-tuning*, adaptations de rang faible, ... *etc*)
- combinaison de modèles *a posteriori* (inspiré des méthodes ensemblistes)
- combinaison de méthodes efficaces

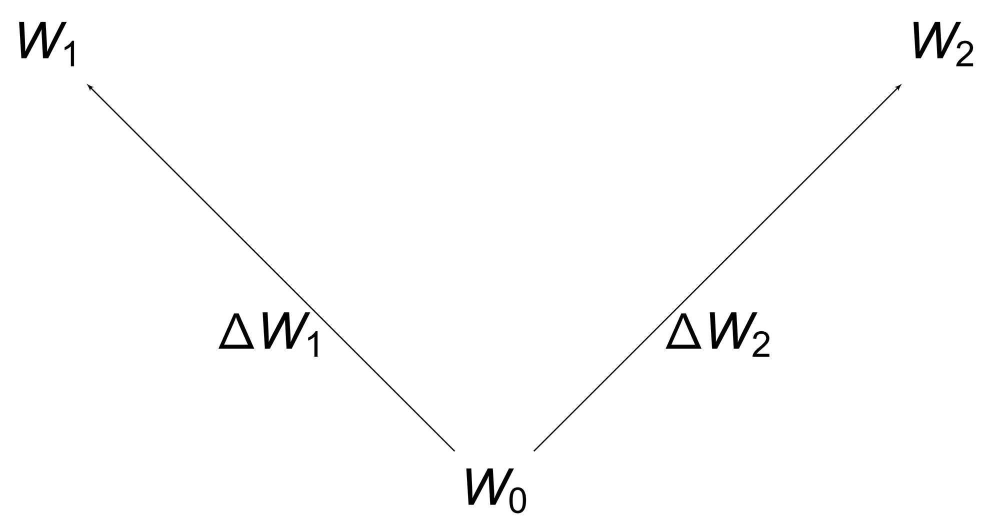
Un mot arrive sur le devant de la scène : l'**arithmétique des tâches** basé sur la notion de **vecteurs de tâches**.

Vecteur de tâches

- W_0 : poids du modèle pré-entraînés
- W_t : poids du modèle adapté à une tâche t

Le vecteur de tâches est défini par :

$$\Delta W_t = W_t - W_0$$



Questions :

- Ce vecteur est-il un « vecteur de tâche » ?
- Si oui comment apprécier les différences entre ces derniers ?
- Avons nous des propriétés arithmétiques entre ces vecteurs ?

Adaptations de rang faible

Soit $W \in \mathbb{R}^{d \times d}$:

- *full-finetuning* : $W_t = W_0 + \Delta W$ avec $\Delta W \in \mathbb{R}^{d \times d}$
- LoRA : $W_t = W_0 + BA$ avec $A \in \mathbb{R}^{r \times d}$ et $B \in \mathbb{R}^{d \times r}$ ($r << d$)

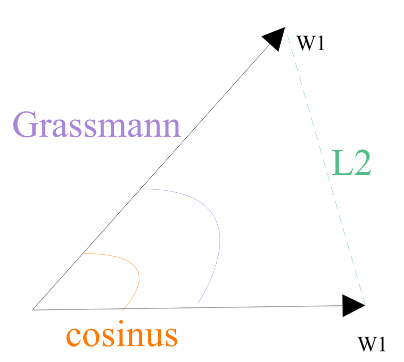
LoRA : estimation de faible dimension (intrinsèque) du vecteur de tâche

Dans la partique :

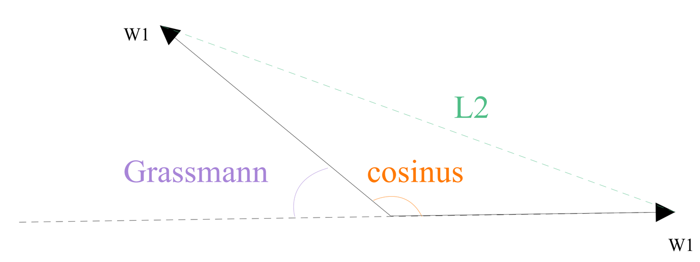
- applicable sur chaque couche linéaire
- pratique courante : Requêtes et Valeurs

Distances entre vecteurs de tâches

- l_2 : $||W_1 - W_2||_2$ position absolue entre les paramètres
- cos : $|1 - cos(W_1, W_2)|$ corrélation entre les paramètres
- Grassmann : $d_G(W_1, W_2) = d(Im(W_1), Im(W_2))$ distance entre les représentations (espaces images)



(a) Grassmann $=$ cos



(b) Grassmann $\neq$ cos

Nous avons une hiérarchie entre les différentes distances

Les tâches (Benchmark GLUE)

Cardinalité des jeux de données :

	COLA	MRPC	RTE	QNLI	QQP	MNLI	SST2	SNLI	YELP	IMDB
train	8.55k	3.67k	2.49k	105k	364k	393k	67.3k	550k	560k	25k
dev	1.04k	408	277	5.46k	40.4k	19.65k	872	10k	-	-
test	-	-	-	-	-	-	-	10k	38k	25k

Performances des modèles (métriques standards de GLUE) :

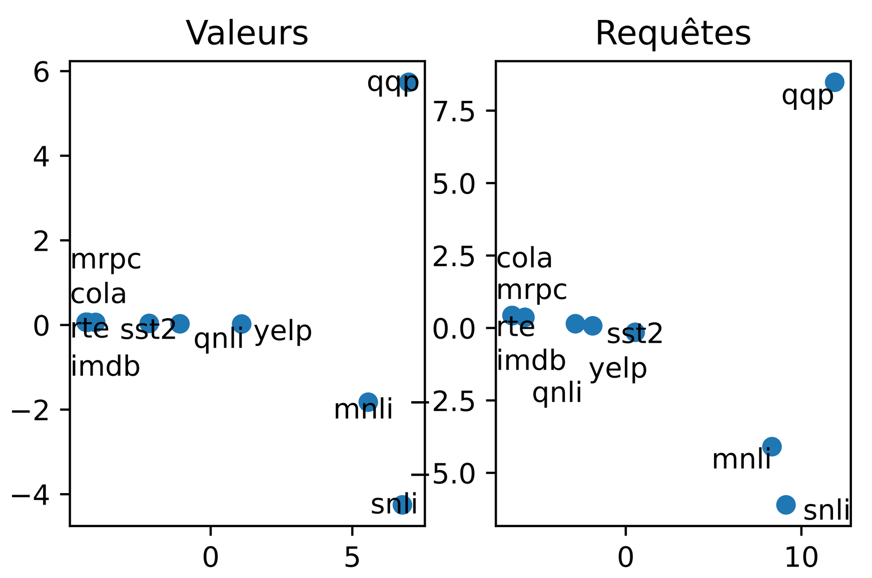
	COLA	MRPC	RTE	QNLI	QQP	MNLI	SST2	SNLI	YELP	IMDB
f-FT	56.36	85.78	69.66	91.87	86.36	85.75	93.69	90.61	97.73	93.95
LoRA	54.98	86.52	71.84	92.28	86.02	86.73	93.92	90.61	98.01	95.34

LoRA $\approx$ f-FT : le finetuning LoRA s'est bien passé !

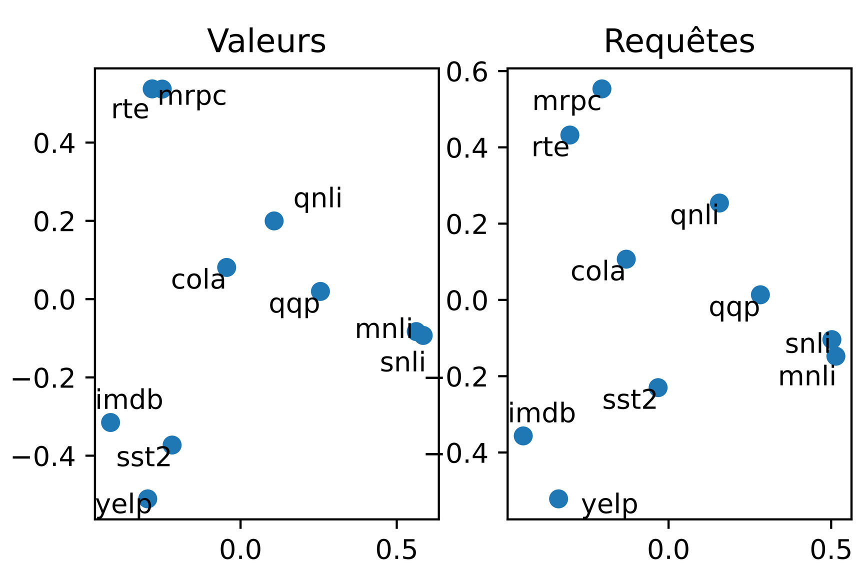
Visualisation du vecteur de tâche

Protocole pour visualiser les distances entre décompositions de rang faible :

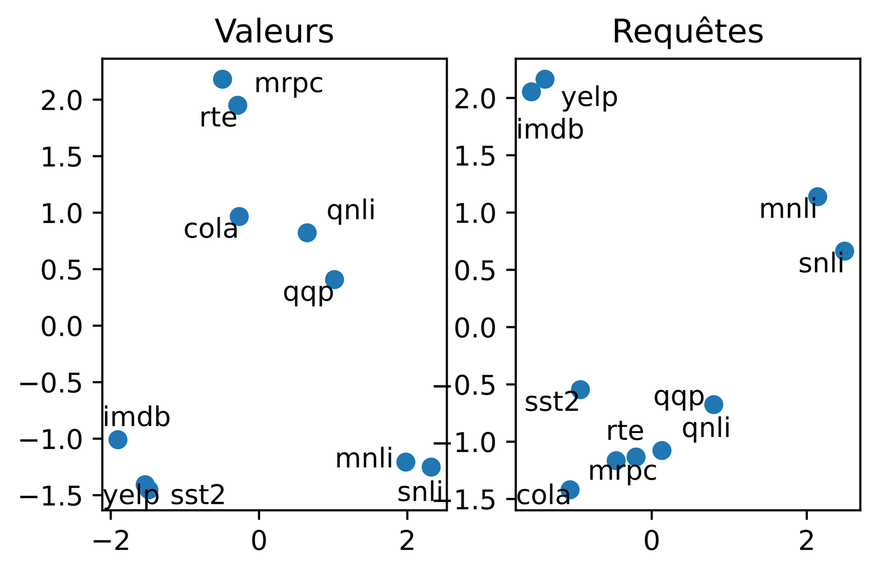
- $T_d(i, j, k)$: distance d entre le vecteur de tâche i et j sur la couche k
- $\bar{T}(i, j) = \frac{1}{K} \sum_k T_d(i, j, k)$
- $\bar{T}(i, j)$ matrice symétrique $\rightarrow$ PCA pour visualiser cette matrice



(a) L_2



(b) cosinus



(c) Grassmann

Combinaisons de modèles

1. $Comb((A_i, B_i)\ ;\ (A_j, B_j)) = (\frac{1}{2}(A_i + A_j), \frac{1}{2}(B_i + B_j))$
2. $\delta(i, j) = perf((A_i, B_i)) - perf(Comb((A_i, B_i)\ ;\ (A_j, B_j)))$
3. $corr_{spe}(\delta(i, j), d(i, j))$

	L_2	cos	d_G	cos^{5+}	d_G^{5+}
Requêtes	0.45 (0.21)	0.65 (0.16)	0.65 (0.25)	0.65 (0.16)	0.61 (0.25)
Valeurs	0.48 (0.23)	0.68 (0.21)	0.64 (0.19)	0.74 (0.20)	**0.77 (0.14)**

Interprétation géométrique

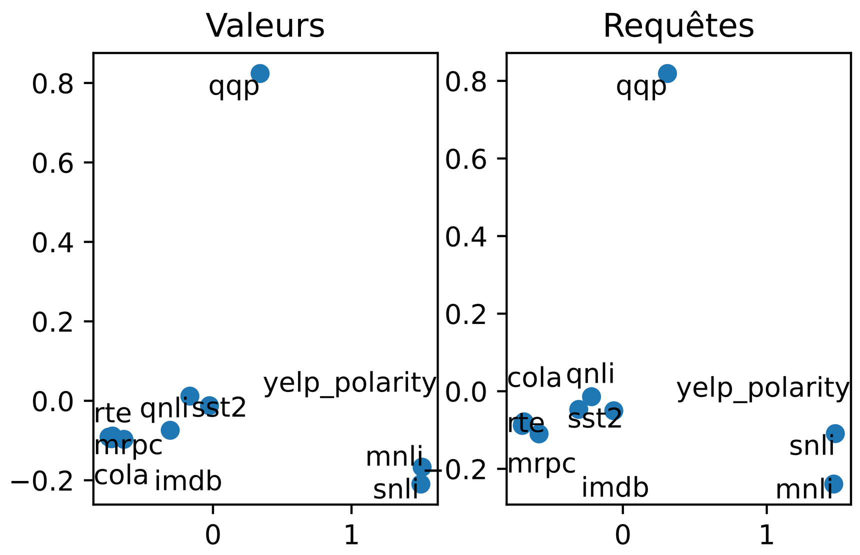
Grassmann et cosinus $\rightarrow$ mêmes interprétations. Cependant nous avons :

$$d_G(B_1 A_1, B_2 A_2) = d_G(B_1, B_2)\ . \tag{1}$$

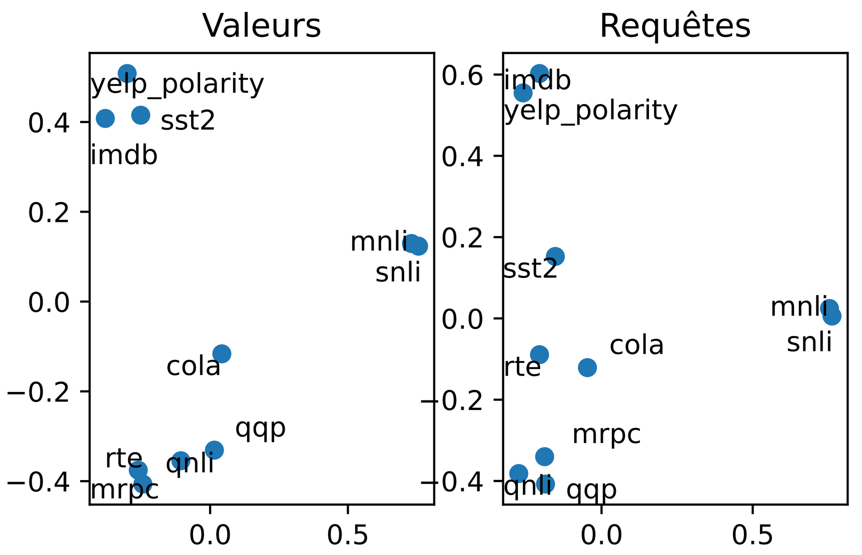
Ainsi les matrice A ne semblent pas participer au vecteur de tâche, explication :

compression $\Rightarrow$ distortion $\Rightarrow$ perte d'information

$d = 768$ — A — $d = r = 8$ — B — $d = 768$



(a) A



(b) B

Les matrices B encodent plus d'informations.

Messages à emporter

- le vecteur de tâche semble être porteur d'informations sur la tâche, à condition d'utiliser les bonnes métriques
- la distance L_2 trop restrictive pour l'évaluation des distances entre vecteurs de tâches $\rightarrow$ utilisation de métriques plus vectorielles (cosinus / Grassmann)
- la distinction entre les tâches est déterminée par la matrice B