UMR IRISA

Institut de Recherche en Informatique et Systèmes Aléatoires

Une études statistiques des plongements dans les modèles *transformers* pour le français

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

I- introduction : Quelques éléments de contexte

Observations de convergences

Dans une étude sur des *LSTM* hiérarchique [NGUYEN et al., 2020] les poids d'attention tendent à avoir un distribution uniforme.

# LSTM	Carte d'attention sur biLSTM
1	A blond little boy in an orange sweatshirt with red sleeves is using scissors to cut something
2	A blond little boy in an orange sweatshirt with red sleeves is using scissors to cut something
3	A blond little boy in an orange sweatshirt with red sleeves is using scissors to cut something

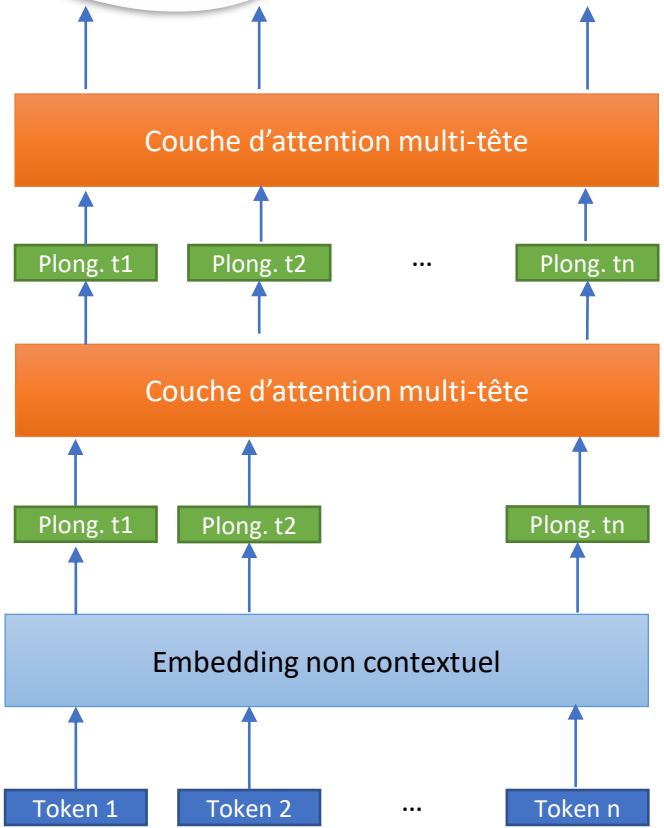
Etude sur la géométrie des plongements

Kawin Ethayarajh. *How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo. And GPT-2 Embeddings.* EMNLP-IJCNLP 2019

Hypothèse de travail

Nous soupçonnons une convergence géométrique des plongements cachés au sein d'une phrase (INTRAPHRASE) lors de la montée en abstraction dans des réseaux de type BERT pré-entrainés sur des corpus français

I- Introduction : Bref rappel sur la notion des plongements cachés dans *BERT*



Plongements cachés dit **contextuels** (issues d'une couche d'attention multi-tête)

Plongements cachés dit **non-contextuels** (issues d'une couche classique d'embedding)

Montée en abstraction dans le modèle

- Un réseau *transformers* est un empilement de couches d'attention multi-têtes.
- La sortie de chaque couche nous donne des plongements cachés de mêmes dimensions

II- Méthodologie : Les architectures utilisées

Flaubert VS CamemBERT	FlauBERT	CamemBERT
Dimension des plongements	768	768
Nombre de couches (numérotation à partir de 0)	13 (12 couches d'attention)	13 (12 couches d'attention)
Nombre de tête d'attention	12	12
Type de couches	XML	RoBERTa
Nombre de paramètres	140 M	110 M
Tâche de pré-entrainement	Language masqué	Language masqué
Nombre de données d'entrainement	71 GB (OPUS + Wikipedia)	138 GB (OSCAR + CCNET + wikipedia)

II- Méthodologie : les métriques pour évaluer la convergence

Considérons , $e_i^k(s) \coloneqq \{e_{ij}^k(s), j \in [1,d]\}$ le plongement du *i-ème token* au niveau k du Transformers

Pour quantifier la convergence des plongements, nous employons les 2 métriques suivantes:

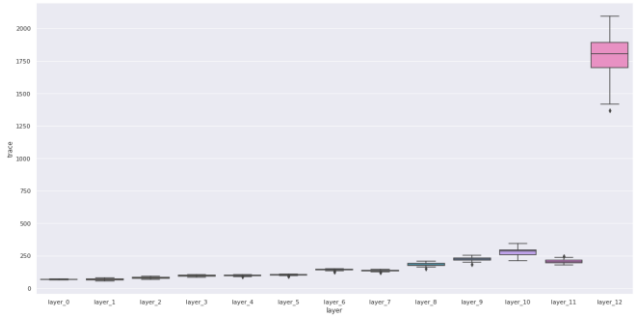
Convergence Absolue	Mesure de similarité
Considérons la variance sur chaque dimension : $$v_j^k(s) = var(e_{ij}^k(s))$$ On considérera pour chaque couche k : $$\sum_j v_j^k(s)$$	Utilisation de la similarité intraphrase (Ethayarajh) : $$c_{ij}^k(s) = \frac{e_i^k(s)\, e_j^k(s)}{\|e_i^k(s)\| \, \|e_j^k(s)\|}$$ Similarité du cosinus au sein d'une phrase s

Ces métriques se concentrent sur la mesure de dispersion des plongements INTRAPHRASE.

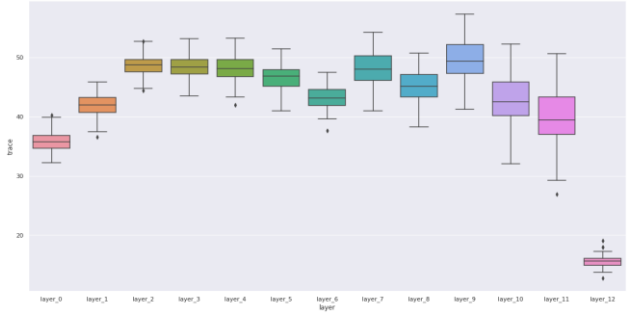
III – Les résultats : distribution des métriques sur les modèles pré-entrainés

Des comportements différents sur les réseaux pré-entrainés

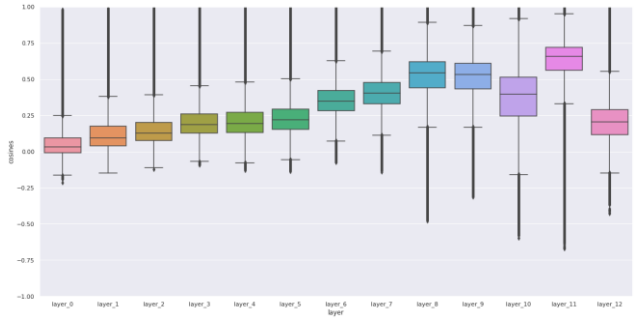
Faire attention aux échelles !



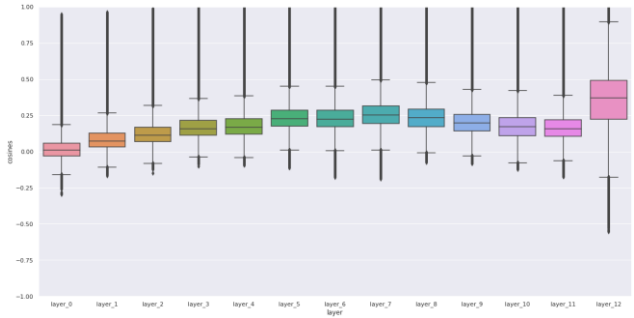
Variance **FlauBERT**



Variance **CamemBERT**



Similarités **FlauBERT**

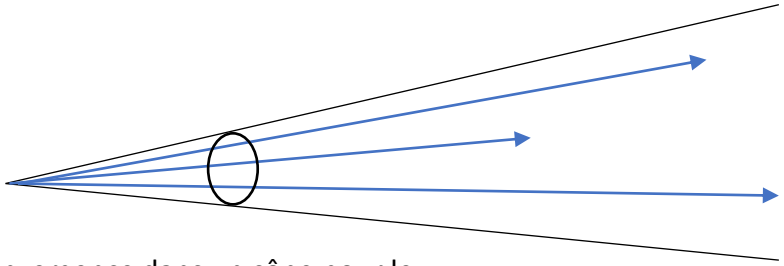


Similarités **CamemBERT**

III – Les résultats : résumé des distributions pour les versions p-e

Une différence entre les deux modèles pré-entrainés.

FlauBERT	CamemBERT
- *Explosion* de la variance sur la dernière couche - *Baisse* de la similarité sur la dernière couche - **Effet de dispersion sur la dernière couche. (pas de convergence observée)**	- Diminution de la variance sur la dernière couche. - Augmentation de la similarité - **Concentration des plongements dans un cône étroit. (première observation de convergence)**



Convergence dans un cône pour le réseau CamemBERT

III- Les Résultats : spécialisation des modèles

Corpus FLUE/CLS :

Subset: CLS

Split: train

text (string)	label (class label)	idx (int)
Prison Break, c'est un peu l'histoire d'un pétard mouillé ; ou comment , à partir d'une idée originale et complètement…	0 (negative)	1
Rare son les Opéras qui respecte la mise en scène et les décors Originaux Une vrai merveille Un vrai bonheur Pour conclure…	1 (positive)	2

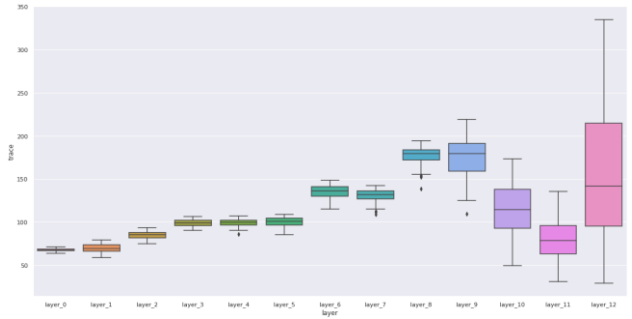
Nombre d'époques	1
Nombre de données d'entrainement	5000
Nombre de données de tests	1000

Précision des deux modèles :
96%

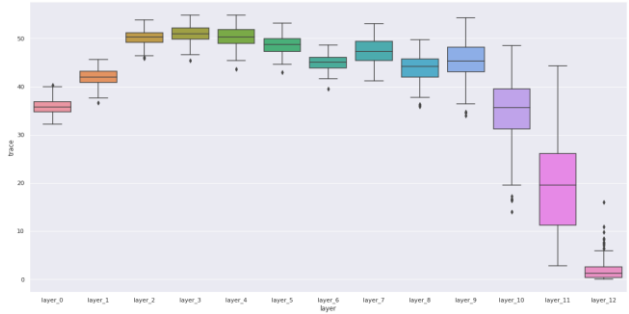
III – Les résultats : distribution sur les modèles spécialisés

Des comportements « similaires » sur les modèles spécialisés : concentration dans un cône étroit des plongements au sein d'une phrase

Convergence de la variance (couche 12)

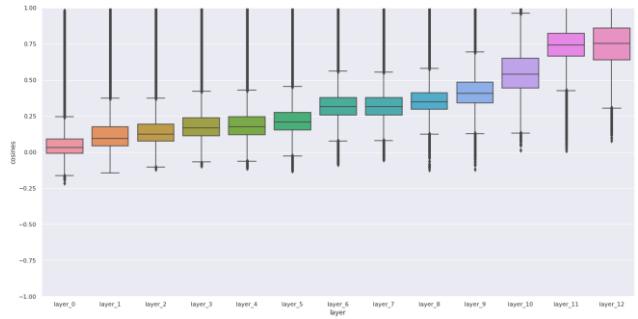


Variance **FlauBERT**

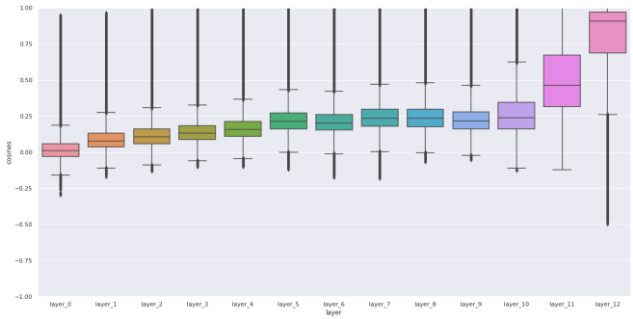


Variance **CamemBERT**

Augmentation de la similarité (couche 12)



Similarités **FlauBERT**



Similarités **CamemBERT**

III – Les résultats : résumé des distribution sur les modèles spécialisés

FlauBERT et camemBERT

- Cette fois-ci, comportement **similaire** pour les deux transformers.

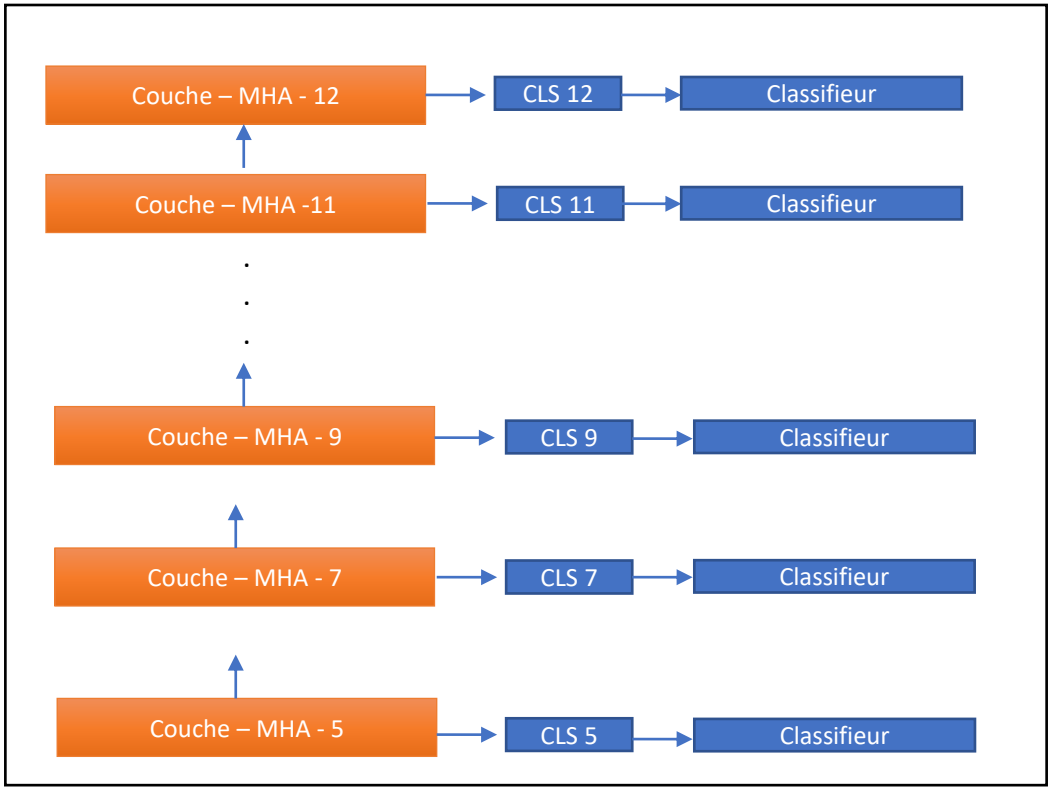
- La spécialisation pour la tâche donnée semble provoquer la concentration dans un cône étroit pour les deux architectures. Cette concentration est tout de même plus prononcée pour CamemBERT.

- Nous observons ainsi bien la convergence évoquée plus tôt (pour les deux réseaux).

- **La spécialisation semble forcer la convergence des plongements lors de la montée en abstraction**.

III- Les résultats : Lien entre la convergence et la performance

Ajout d'une expérience pour l'étude d'un possible lien entre la convergence et la performance



- Récupération du plongement du *token* CLS pour les différents modèles à différents étages de nos architecture
- Construction *d'une tête de classification* sur ces différents plongements

- Expérience inspirée de [Rogers et al. (2020)]

III- Les résultats : Lien entre la convergence et la performance

Précision de la tête de classification (en pourcentage)

	CLS 5	CLS 7	CLS 9	CLS 11	CLS 12
FlauBERT pré-entrainé	69,8	71,4	72	72,8	73
Flaubert Spécialisé	77,7	85,6	91,9	92,9	93,1
CamemBERT pré-entrainé	73,1	81,3	89,7	84,4	89,6
CamemBERT spécialisé	77,9	86,5	93,4	94,6	94,5

- Les modèles spécialisés présentent des comportements similaires, avec une précision qui augmentent lors de la montée en abstraction, tout comme la convergence.
- Pour les versions pré-entrainés, les plongements de CamemBERT permettent d'obtenir une précision nettement meilleure

III- Les résultats : Lien entre la convergence et la performance

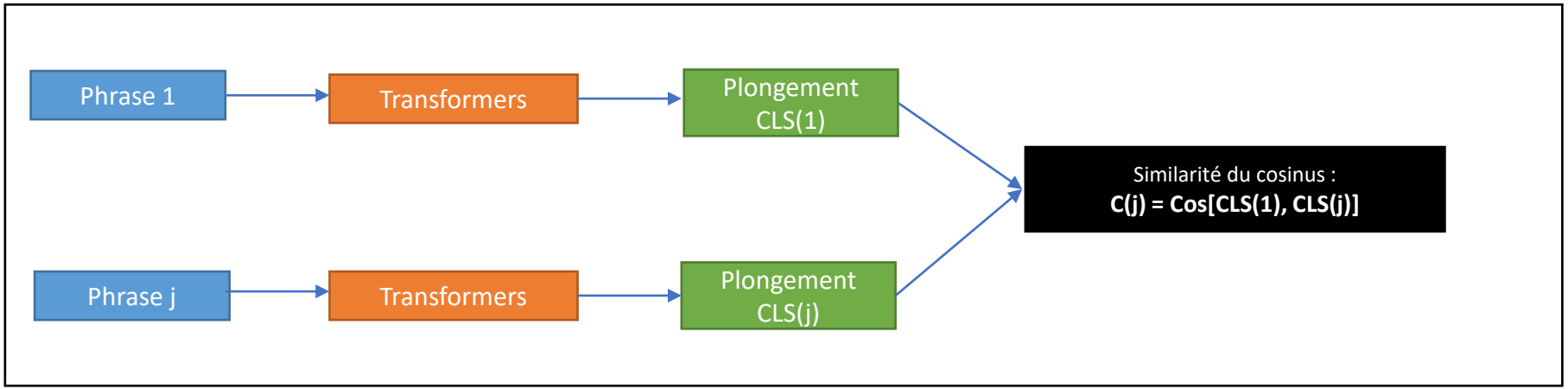
- Pour les deux réseaux spécialisés les résultats sont comparables (à l'image de la convergence observée précédemment)

- Pour les version pré-entrainées on remarque cependant que CamemBERT (qui présente aussi une forte convergence dans un cône étroit) donne des résultats très largement supérieurs. Une fois de plus les résultats pour les versions pré-entrainées présentent des différences.

- **Il semblerait qu'il y ait un lien étroit entre la performance et la convergence.**

III – Présentation des résultats : 1 cône = 1 classe ?

Cette fois-ci regardons la similarité entre les différentes phrases (dépassons le cadre intra-phrases)

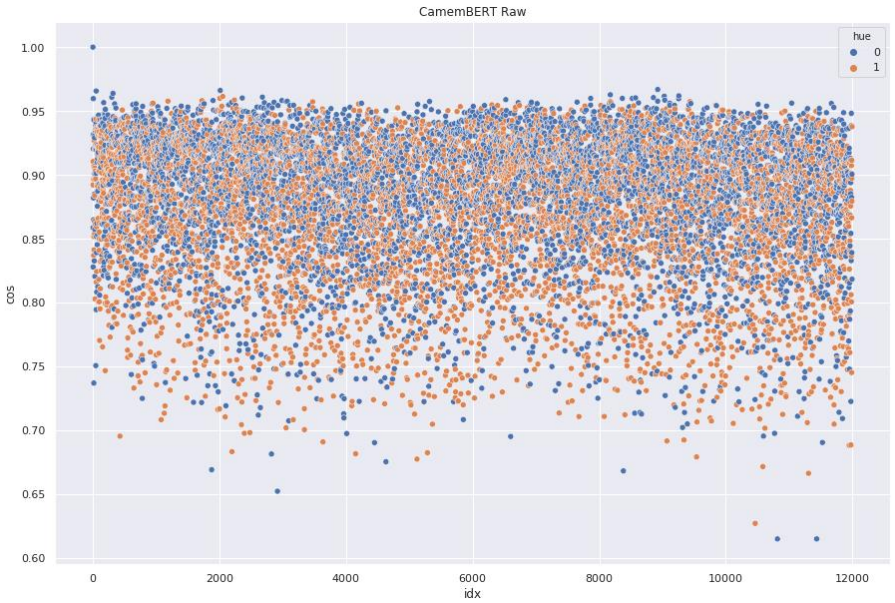


- Pour chaque phrase j de notre jeu de données nous allons calculer la similarité du cosinus entre son plongement CLS final et le plongement CLS final de notre phrase 1.

- Nous allons observer la répartition des coefficients C(j)

III – Présentation des résultats : 1 cône = 1 classe ? Les modèles pré-entrainés

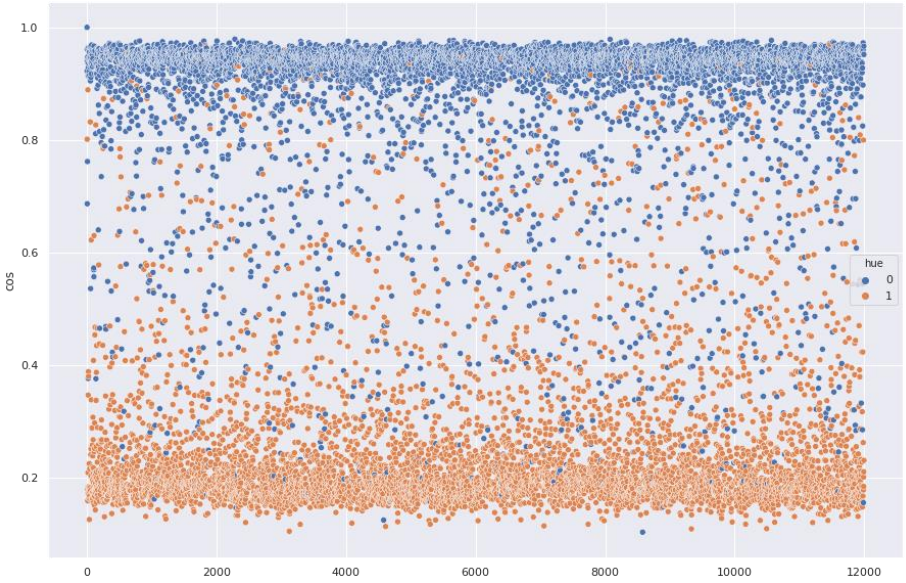
FlauBERT



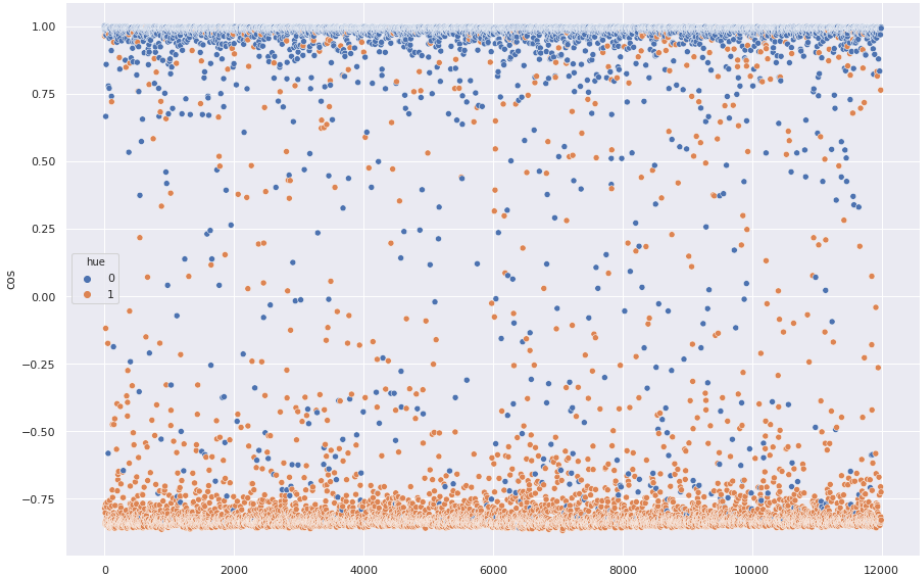
CamemBERT

Pour les modèles pré-entrainés, nous n'observons pas de schéma particulier, les distributions sont uniformes nous n'avons pas de séparation des classes.

III – Présentation des résultats : 1 cône = 1 classe ? Les modèles spécialisés



FlauBERT



CamemBERT

Pour les deux modèles spécialisés, nous avons une séparation des classes :

- Les cônes semblent être des propriétés de classe et non des propriétés de phrases
- Pour Flaubert les cônes de chaque classe sont orthogonaux
- Pour camemBERT la séparation et plus violente et les cônes sont opposés

IV – Conclusion

Les questions ouvertes que cette étude a soulevées

- Les différences de comportements entre les modèles pré-entrainés restent pour l'instant inexpliquées – première piste sur les données de pré-entrainement et les tâches de pré-entrainement.

- La convergence dans un cône est visible sur une tâche de classification de document, que se passe-t-il si on effectue une tâche d'étiquetage où cette fois-ci nous aurons une plusieurs classes différentes au sein de chaque phrase.

- La convergence pour la tâche de classification de document permet de mettre en avant l'apparition de cônes de classes. Est-ce possible d'utiliser ce cône de classe pour un critère d'entrainement ?

Duc-Hau Nguyen, Guillaume Gravier, Pascale Sébillot. A study of the Plausibility of Attention between RNN Encoders in Natural Language Inference. ICMLA 2020

Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo. And GPT-2 Embeddings. EMNLP-IJCNLP 2019

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2020). A primer in bertology : What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842–866.

Merci à tous !

Contacts :
- loic.fosse@insa-rennes.fr
- duc-hau.nguyen@irisa.fr
- guig@irisa.fr
- pascale.sebillot@irisa.fr

www.irisa.fr @irisa_lab

Institut de Recherche en Informatique et Systèmes Aléatoires