Hierarchical BiLSTM Max Pooling (Talman et al., 2018)

Qu'est ce que GLUE ?

GLUE (General Language Understanding Evaluation) est une plate-forme de référence et d'analyse multi-tâches pour la compréhension du langage naturel.

Il faut obtenir ainsi le meilleur score possible dans les 11 tâches d'évaluation (voir lien ci-dessous) imposé par la plate-forme.

- https://gluebenchmark.com/leaderboard
- https://gluebenchmark.com/tasks

Classement GLUE de Hierarchical BiLSTM Max Pooling (Talman et al., 2018) : 78

Score obtenus : 63.6

Dans l'ensemble, notre modèle améliore l'état de l'art précédent pour le jeu de donnée SciTail et obtient de bons résultats pour le corpus SNLI et Multi-Genre Natural Language Inference. Notre architecture proposée suit une approche basée sur l'intégration de phrases pour NLI.

Architecture du modèle Hierarchical BiLSTM Max Pooling (Talman et al., 2018)

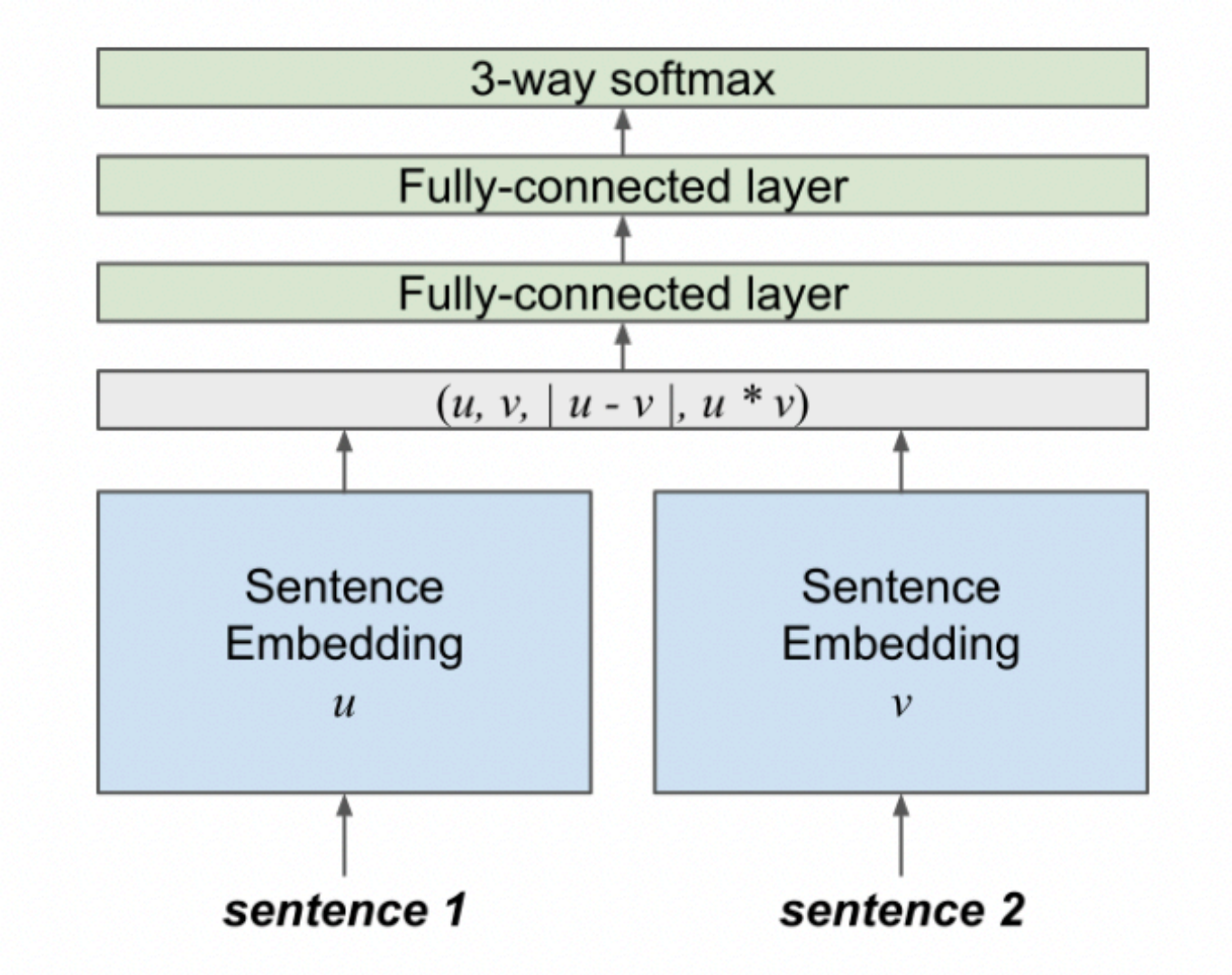
Le modèle propose une hiérarchie de couches BiLSTM (LSTM bidirectionnels) et max pooling qui implémente une stratégie de raffinement itérative et produit des résultats de pointe sur l'ensemble de données SciTail ainsi que des résultats solides pour SNLI et MultiNLI.

Les BiLSTM utilisent deux LSTM pour s’entraîner sur l’entrée séquentielle.

Les cellules LSTM (Long Short Term Memory) possèdent une mémoire interne appelée cellule. La cellule permet de maintenir un état aussi longtemps que nécessaire. Cette cellule consiste en une valeur numérique que le réseau peut piloter en fonction des situations.

L'idée de base derrière ces approches est d'encoder séparément les phrases de prémisse et d'hypothèse, puis de les combiner à l'aide d'un classificateur de réseau neuronal.

L'architecture à été implémenté à l'aide de PyTorch.



Le modèle illustré sur l'image ci-dessus contient des incorporations de phrases pour les deux phrases d'entrée, où la sortie des incorporations de phrases est combinée à l'aide d'une heuristique introduite par Mou, en rassemblant la concaténation (u, v), la différence absolue par élément |u - v|, et le produit par élément u * v.

Le vecteur combiné est ensuite transmis à un perceptron multicouche à 3 couches (MLP) avec un classificateur softmax à 3 voies.

Les deux premières couches du MLP utilisent à la fois l'abandon et une fonction d'activation ReLU.

Pour les représentations de phrases, nous incorporons d'abord les mots individuels avec des incorporations de mots pré-entraînées. La séquence des mots incorporés est ensuite transmise à l'encodeur de phrase qui utilise le BiLSTM avec le max pooling.

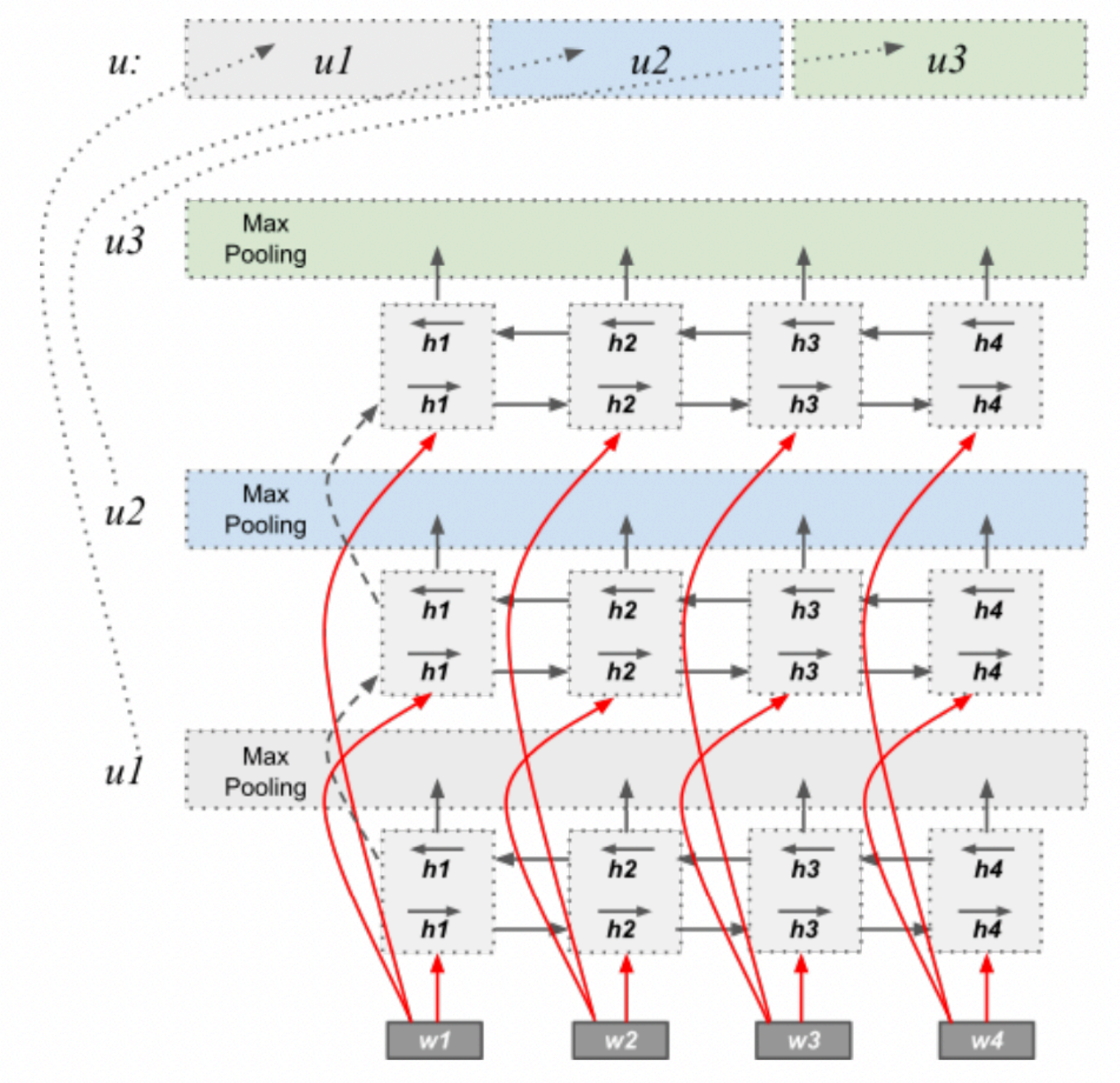
Pour résumer les différences entre ce modèle et le BiLSTM classique, nous pouvons énumérer les trois principaux aspects suivants :

- Chaque couche de notre modèle est un BiLSTM distinct initialisé avec les états cachés et cellulaires de la couche précédente.
- Chaque couche de notre modèle reçoit les mêmes incorporations de mots que son entrée.
- La représentation finale de la phrase est la concaténation du max mis en commun sortie de chaque couche du réseau d'encodage.

Pour le modèle proposé, nous avons utilisé un algorithme d'optimisation de descente de gradient basé sur la règle de mise à jour d'Adam, qui est pré-implémentée dans PyTorch.

Le taux d'apprentissage était diminué d'un facteur de 0,2 après chaque époque si le modèle ne s'améliorait pas.

Architecture de l'encodeur de phrases HBMP



Les résultats montrent que le HBMP est plus performant que chacun des autres modèles testé dans le papier, ce qui justifie l'utilisation de cette configuration (voir l'image ci-dessus) en faveur d'autres architectures.

De plus, nous pouvons voir que les différents composants contribuent tous au score final. L'assemblage d'informations provenant de trois couches BiLSTM distinctes (avec des paramètres indépendants) améliore les performances comme nous pouvons le voir dans la comparaison entre BiLSTM-Ens et BiLSTM-Ens-Tied.

L'initialisation entraînable ne semble pas ajouter à la capacité du modèle et indique que l'initialisation hiérarchique que nous proposons est effectivement bénéfique. Enfin, le fait d'alimenter tous les BiLSTM du modèle avec les mêmes encastrements d'entrée entraîne une amélioration par rapport au modèle empilé qui ne relit pas l'information d'entrée.

Les résultats montrent que le modèle fonctionne mieux que chacun des autres modèles, qui prend en charge l'utilisation de notre configuration en faveur d'architectures alternatives. De plus, nous pouvons voir que les différentes composantes contribuent toutes au score final.

Résultat

Dans cet article, nous avons introduit une architecture de raffinement itératif (le modèle proposé) basé sur des couches BiLSTM avec une mise en commun maximale qui atteint un nouvel état de l'art pour SciTail et de bons résultats dans la catégorie d'encodage de phrases SNLI et MultiNLI.

SNLI : Le corpus Stanford Natural Language Inference (SNLI) est un ensemble de données de 570 000 paires de phrases écrites par des humains, étiquetées manuellement avec les étiquettes dorées implication, contradiction et neutre. L'ensemble de données est divisé en ensembles d'entraînement (550 152 paires), de développement (10 000 paires) et de test (10 000 paires).

SciTail : SciTail est un ensemble de données NLI créé à partir d'examens scientifiques à choix multiples composés de 27 000 paires de phrases. Chaque question et le choix de la bonne réponse ont été convertis en une déclaration assertive pour former l'hypothèse. L'ensemble de données est divisé en ensembles d'entraînement (23 596 paires), de développement (1 304 paires) et de test (2 126 paires). Contrairement aux ensembles de données SNLI et MultiNLI, SciTail n'utilise que deux étiquettes : implication et neutre.

MultiNLI : Le corpus Multi-Genre Natural Language Inference (MultiNLI) est un corpus à large couverture pour l'inférence en langage naturel, composé de 433 000 paires de phrases écrites par l'homme étiquetées avec implication, contradiction et neutre. Contrairement au corpus SNLI, qui tire la phrase de prémisse des légendes d'images, MultiNLI se compose de paires de phrases de dix genres distincts d'anglais écrit et parlé. L'ensemble de données est divisé en ensembles d'entraînement (392 702 paires), de développement (20 000 paires) et de test (20 000 paires).

		Predicted - HBMP				Predicted - InferSent			
		entail	contradict	neutral	*recall*	entail	contradict	neutral	*recall*
Gold	entail	3047	58	263	**90.5%**	2967	95	306	**88.1%**
	contradict	117	2840	280	**87.7%**	154	2756	327	**85.1%**
	neutral	357	240	2622	**81.5%**	346	302	2571	**79.9%**
	precision	**86.5%**	**90.5%**	**82.8%**		**85.6%**	**87.4%**	**80.2%**	

Table 6. SNLI confusion matrices for HBMP and InferSent.

		HBMP			InferSent		
		entail	neutral	*recall*	entail	neutral	*recall*
Gold	entail	632	210	**75.0%**	673	169	**79.9%**
	neutral	88	1196	**93.1%**	140	1144	**89.1%**
	precision	**88.0%**	**85.0%**		**82.8%**	**87.1%**	

Table 7. SciTail confusion matrices for HBMP and InferSent based on the development set.

		Predicted - HBMP				Predicted - InferSent			
		entail	contradict	neutral	*recall*	entail	contradict	neutral	*recall*
Gold	entail	2781	196	486	**80.3%**	2614	278	587	**75.1%**
	contradict	372	2354	514	**72.7%**	449	2241	523	**69.7%**
	neutral	528	443	2158	**69.0%**	477	507	2139	**68.5%**
	precision	**75.6%**	**78.7%**	**68.3%**		**73.8%**	**74.1%**	**65.8%**	

Table 8. MultiNLI-matched confusion matrices for HBMP and InferSent based on the development set.

Nous démontrons que notre modèle surpasse InferSent dans presque tous les cas, avec une réduction substantielle de la confusion entre les classes de relations inférentielles.

L'analyse linguistique sur MultiNLI révèle également que notre approche est robuste dans les différentes catégories et surpasse InferSent sur, par exemple, les antonymes et les négations qui nécessitent une bonne niveau d'abstraction sémantique.

BiLSTM avec max pooling obtient les meilleurs résultats non seulement en NLI mais aussi dans de nombreuses autres tâches NLP nécessitant des représentations de sens au niveau de la phrase. Ils montrent également que le modèle formé sur les données NLI atteint de bonnes performances sur diverses tâches d'apprentissage par transfert.