Multi-task BiLSTM + Attn (Wang 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 Multi-task BiLSTM + Attn (Wang et al., 2018) : 78

Score obtenus : 65.6

Ce modèle resemble très forte au modèle Multi-task BiLSTM + Attn en ce qui concerne l'architecture, les domaines où il obtient de bons résultats ainsi que le score obtenu sur GLUE. Le papier qui le présente est une introduction au référentiel GLUE où le modèle a été un des premiers modèles utilisé pour tester ce référentiel. Depuis la création de ce référentiel, une cinquantaine de modèle obtiennent de meilleure résultats que lui sur tout les plans donc il est devenu obsolète.

GLUE

La capacité humaine à comprendre le langage est générale, flexible et robuste. En revanche, la plupart des modèles NLU au-dessus du niveau du mot sont conçus pour une tâche spécifique et se débattent avec des données hors du domaine. Si nous aspirons à développer des modèles dont la compréhension va au-delà de la détection de correspondances superficielles entre les entrées et les sorties, il est essentiel de

développer un modèle plus unifié qui puisse apprendre à exécuter une gamme de tâches linguistiques différentes dans différents domaines.

Pour faciliter la recherche dans cette direction, nous présentons le référentiel General Language Understanding Evaluation (GLUE) : une collection de tâches NLU comprenant la réponse à des questions, l'analyse de sentiments et l'implication textuelle, et une plateforme en ligne associée pour l'évaluation, la comparaison et l'analyse de modèles. GLUE n'impose pas de contraintes sur l'architecture des modèles au-delà de la capacité à traiter des entrées d'une seule phrase ou d'une paire de phrases et à faire les prédictions correspondantes. Pour certaines tâches GLUE, les données d'entraînement sont abondantes, mais pour d'autres, elles sont limitées ou ne correspondent pas au genre de l'ensemble de test. GLUE favorise donc les modèles qui peuvent apprendre à représenter les connaissances linguistiques d'une manière qui facilite l'apprentissage efficace des échantillons et le transfert efficace des connaissances entre les tâches. Aucun des jeux de données de GLUE n'a été créé de toutes pièces pour le benchmark ; nous nous appuyons sur des jeux de données préexistants parce qu'ils ont été implicitement reconnus par la communauté NLP comme étant difficiles et intéressants. Quatre des jeux de données contiennent des données de test privées, qui seront utilisées pour s'assurer que le benchmark est utilisé de manière équitable.

Architecture

L' architecture de base la plus simple est basée sur des codeurs de phrases en vecteurs, et met de côté la capacité de GLUE à évaluer des modèles avec des structures plus complexes. Le modèle utilise un BiLSTM (LSTM bidirectionnels) à deux couches, 1500D avec max pooling et 300D GloVe word embeddings. Pour les tâches à une seule phrase, nous encodons la phrase et passons le vecteur résultant à un classificateur. Pour les tâches de paires de phrases, nous encodons les phrases indépendamment pour produire des vecteurs u, v, et passons [u , v ; |u - v| ; u ∗ v] à un classificateur (détail de cette phase dans un autre état de l'art). Le classifieur est un MLP avec une couche cachée de 512D.

En gros dans ce papier, pour introduire et présenter GLUE, nous avons fait une série de test sur différents modèles où le modèle Multi-task BiLSTM + Attn était présent et nous avons obtenus les résultats ci-dessous :

Chaque colonne du tableau corrpespond à une tâche spécifique (que je peux décrire et ajouter au besoin) :

Model	Avg	Single Sentence CoLA	SST-2	Similarity and Paraphrase MRPC	QQP	STS-B	Natural Language Inference MNLI	QNLI	RTE	WNLI
Single-Task Training										
BiLSTM	63.9	15.7	85.9	69.3/79.4	81.7/61.4	66.0/62.8	70.3/70.8	75.7	52.8	**65.1**
+ELMo	66.4	**35.0**	90.2	69.0/80.8	85.7/65.6	64.0/60.2	72.9/73.4	71.7	50.1	**65.1**
+CoVe	64.0	14.5	88.5	73.4/81.4	83.3/59.4	67.2/64.1	64.5/64.8	75.4	53.5	**65.1**
+Attn	63.9	15.7	85.9	68.5/80.3	83.5/62.9	59.3/55.8	74.2/73.8	77.2	51.9	**65.1**
+Attn, ELMo	66.5	**35.0**	90.2	68.8/80.2	**86.5/66.1**	55.5/52.5	**76.9/76.7**	76.7	50.4	**65.1**
+Attn, CoVe	63.2	14.5	88.5	68.6/79.7	84.1/60.1	57.2/53.6	71.6/71.5	74.5	52.7	**65.1**
Multi-Task Training										
BiLSTM	64.2	11.6	82.8	74.3/81.8	84.2/62.5	70.3/67.8	65.4/66.1	74.6	57.4	**65.1**
+ELMo	67.7	32.1	89.3	**78.0/84.7**	82.6/61.1	67.2/67.9	70.3/67.8	75.5	57.4	**65.1**
+CoVe	62.9	18.5	81.9	71.5/78.7	84.9/60.6	64.4/62.7	65.4/65.7	70.8	52.7	**65.1**
+Attn	65.6	18.6	83.0	76.2/83.9	82.4/60.1	72.8/70.5	67.6/68.3	74.3	58.4	**65.1**
+Attn, ELMo	**70.0**	33.6	**90.4**	**78.0**/84.4	84.3/63.1	74.2/72.3	74.1/74.5	**79.8**	58.9	**65.1**
+Attn, CoVe	63.1	8.3	80.7	71.8/80.0	83.4/60.5	69.8/68.4	68.1/68.6	72.9	56.0	**65.1**
Pre-Trained Sentence Representation Models										
CBoW	58.9	0.0	80.0	73.4/81.5	79.1/51.4	61.2/58.7	56.0/56.4	72.1	54.1	**65.1**
Skip-Thought	61.3	0.0	81.8	71.7/80.8	82.2/56.4	71.8/69.7	62.9/62.8	72.9	53.1	**65.1**
InferSent	63.9	4.5	85.1	74.1/81.2	81.7/59.1	75.9/75.3	66.1/65.7	72.7	58.0	**65.1**
DisSent	62.0	4.9	83.7	74.1/81.7	82.6/59.5	66.1/64.8	58.7/59.1	73.9	56.4	**65.1**
GenSen	66.2	7.7	83.1	76.6/83.0	82.9/59.8	**79.3/79.2**	71.4/71.3	78.6	**59.2**	**65.1**

Nous constatons que l'entraînement multi-tâches donne de meilleurs résultats globaux que l'entraînement monotâche pour les modèles utilisant l'attention ou ELMo.

L'attention a généralement un effet global négligeable ou négatif dans l'entraînement à une seule tâche, mais elle est utile dans l'entraînement multi-tâches.

Nous constatons une amélioration constante dans l'utilisation des incorporations ELMo à la place des incorporations GloVe ou CoVe, en particulier pour les tâches à une seule phrase.

		Coarse-Grained				Fine-Grained					
Model	**All**	**LS**	**PAS**	**L**	**K**	**UQuant**	**MNeg**	**2Neg**	**Coref**	**Restr**	**Down**
Single-Task Training											
BiLSTM	21	25	24	16	16	70	53	4	21	-15	**12**
+ELMo	20	20	21	14	17	70	20	**42**	33	-26	-3
+CoVe	21	19	23	20	18	71	47	-1	33	-15	8
+Attn	25	24	30	20	14	50	47	21	38	-8	-3
+Attn, ELMo	**28**	**30**	**35**	**23**	14	**85**	20	**42**	33	-26	-3
+Attn, CoVe	24	29	29	18	12	77	50	1	18	-1	**12**
Multi-Task Training											
BiLSTM	20	13	24	14	22	71	17	-8	31	-15	8
+ELMo	21	20	21	19	21	71	**60**	2	22	0	**12**
+CoVe	18	15	11	18	**27**	71	40	7	**40**	0	8
+Attn	18	13	24	11	16	71	1	-12	31	-15	8
+Attn, ELMo	22	18	26	13	19	70	27	5	31	-26	-3
+Attn, CoVe	18	16	25	16	13	71	26	-8	33	**9**	8
Pre-Trained Sentence Representation Models											
CBoW	9	6	13	5	10	3	0	13	28	-15	-11
Skip-Thought	12	2	23	11	9	61	6	-2	30	-15	0
InferSent	18	20	20	15	14	77	50	-20	15	-15	-9
DisSent	16	16	19	13	15	70	43	-11	20	-36	-09
GenSen	20	28	26	14	12	78	57	2	21	-15	**12**

En conclusion, GLUE est une plateforme et une collection de ressources pour évaluer et analyser les systèmes de compréhension du langage naturel. On peut remarquer que, dans l'ensemble, les modèles formés conjointement à nos tâches obtiennent de meilleures performances que les performances combinées des modèles formés pour chaque tâche séparément.

On constate aussi l'utilité des mécanismes d'attention et des méthodes d'apprentissage par transfert telles que ELMo dans les systèmes NLU, qui se combinent pour surpasser les meilleurs modèles de représentation de phrases sur le benchmark GLUE, mais laissent encore une marge d'amélioration.

En revanche, lorsqu'on évalue ces modèles sur notre jeu de données de diagnostic, nous constatons qu'ils échouent (souvent de manière spectaculaire) sur de nombreux phénomènes linguistiques, ce qui suggère des pistes de travail pour l'avenir. En résumé, la question de savoir comment concevoir des modèles NLU à usage général reste sans réponse, et nous pensons que GLUE peut fournir un terrain fertile pour relever ce défi.