Borsa di studio attivata ai sensi di quanto disposto dal D.M. n. 1061 del 10/08/2021

Titolo del progetto: Deep learning ecosostenibile per il Natural Language Processing su larga scala

La borsa sarà attivata sul seguente corso di dottorato accreditato per il XXXVII ciclo:
INGEGNERIA INFORMATICA

Responsabile scientifico: Roberto Navigli
Area per la quale si presenta la richiesta: GREEN
Numero di mensilità da svolgere in azienda: 12
Azienda: Babelscape
Il Dipartimento è disponibile a cofinanziare per un importo pari a euro: 7000
Dipartimento finanziatore: DIPARTIMENTO DI INGEGNERIA INFORMATICA, AUTOMATICA E GESTIONALE - ANTONIO RUBERTI- con delibera del 20/9/2021

Progetto di ricerca:
I modelli linguistici sono diventati la pietra angolare dell'attuale Natural Language Processing in quanto codificano molte informazioni linguistiche in forma latente e permettono la comprensione e la generazione di testi su larga scala, anche in più lingue contemporaneamente (ad esempio, consentendo la traduzione automatica o il dialogo allo stato dell'arte). Tuttavia, questi modelli sono cresciuti considerevolmente nel tempo, con centinaia di milioni di parametri (es. BERT) e, più recentemente, diversi miliardi di parametri da addestrare (es. GPT-3 e WuDao). Purtroppo, le emissioni di carbonio per l'addestramento di tali modelli sono molto elevate (Patterson et al., 2021): è stato stimato che i sistemi di intelligenza artificiale, e in particolare i modelli di NLP, hanno oggi un'impronta di carbonio molto significativa (https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/?sh=5bcadc756b43): l'addestramento di un singolo modello di deep learning può generare fino a 626.155 libbre di emissioni di CO2, più o meno l'impronta di carbonio totale di cinque automobili durante la loro intera vita. Come punto di confronto, un americano produce in media circa 36.000 libbre di emissioni di CO2 in un anno (Strubell et al., 2020). Sfortunatamente, questa tendenza può solo peggiorare e si aggiunge alle attuali preoccupazioni sul cambiamento climatico (https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/?sh=5bcadc756b43).

Questa tendenza è motivata dalla necessità di andare oltre gli embedding statici di parole e ottenere embedding di parole contestualizzate, portando all'uso di architetture neurali ricorrenti (ad esempio BiLSTMs) per ottenere embedding contestualizzati generalizzati e alla comparsa di ELMO (Peters et al., 2018). Questo percorso ha trovato il suo apice con la prima architettura Transformer (Vaswani et al., 2017) che ha sfruttato il meccanismo di Attention che ha superato gli approcci ricorsivi e ha permesso la realizzazione del primo modello Transformer pre-trained, BERT (Devlin et al., 2019).

Similmente a quanto accaduto all'inizio dello scorso decennio nella computer vision con l'emergere delle ResNet, i modelli Transformer hanno preso la scena del NLP, grazie alla loro adattabilità in nuovi compiti facendo leva sul cosiddetto transfer learning. Essendo pre-addestrati su massicce quantità di testo eterogeneo, possono generalizzare facilmente quando affrontano nuovi compiti che sembravano impossibili con gli approcci precedenti.

Il noto benchmark GLUE (Wang et al.,2018) traccia le prestazioni su una serie di compiti di comprensione del

linguaggio naturale (NLU). Le BiLSTMs hanno raggiunto prestazioni medie di 64,2 punti. ELMO ha portato le prestazioni a 67,7 punti, 70 se combinato con Attention. Negli ultimi tre anni dalla comparsa di BERT, le attuali architetture basate su Transformer hanno superato le prestazioni umane. Al momento in cui scriviamo, ERNIE 3.0 (Sun et al.,2021) è in testa con 91,1 punti, più di 20 punti sopra la baseline di ELMO.

Tuttavia, Ernie 3.0 ha 10 miliardi di parametri, mentre ELMO ne ha 93,6 milioni. Nonostante l'aumento delle prestazioni con le dimensioni del modello, questa tendenza ha un ovvio rovescio della medaglia. L'addestramento e l'esecuzione dell'inferenza con questi modelli diventano più dispendiosi in termini di energia all'aumentare delle loro dimensioni. Mentre i componenti e le risorse dei computer sono diventati più potenti ed efficienti, la tendenza alla crescita dei grandi modelli linguistici supera di gran lunga il loro ritmo di evoluzione. Inoltre, i modelli spesso richiedono hardware specializzato come GPU o TPU. Combinato con il prezzo e la scarsità di tali componenti, la creazione di modelli di linguaggio per l'NLP su larga scala che siano ecosostenibili è diventata una, se non la, sfida "green" per i ricercatori e le imprese che si occupano di AI (Patterson et al.,2021).

C'è quindi un urgente bisogno di modelli di linguaggio più leggeri ed efficienti, e di metodi veloci ed economici per adattarli a nuovi domini e compiti. Questa necessità ha già motivato un crescente interesse per gli approcci di Deep Learning che renderanno i modelli più leggeri ed efficienti. Per esempio, il pruning, in cui l'ipotesi del "biglietto della lotteria" viene messa in pratica e alcune parti del modello vengono eliminate senza perdita di prestazioni, mira a creare un modello linguistico più piccolo (Wang et al.,2020). Anche la distillazione ha guadagnato trazione, dove un modello di lingua più grande viene utilizzato come modello "insegnante", e la sua conoscenza viene distillata in uno più piccolo, cercando di sacrificare la minor quantità di prestazioni (Jiaoet al.,2020). Inoltre, sul lato dell'architettura, uno dei principali colli di bottiglia dei modelli Transformer è il meccanismo di attenzione e la complessità quadratica (O(N2)) della sua computazione. Diversi tentativi sono stati fatti per facilitare la complessità di calcolo dell'attenzione sostituendola con trasformate di Fourier (Lee-Thorp et al.,2021), o rendendola lineare nella lunghezza della sequenza mascherando parti del meccanismo di attenzione (Beltagy et al. 2020).

L'obiettivo di questo dottorato sarà quello di fornire metodi efficienti per impiegare e addestrare modelli linguistici per compiti di NLU, con particolare attenzione a quelli sviluppati presso il laboratorio NLP della Sapienza guidato dal prof. Roberto Navigli, quali multilinguismo, Word Sense Disambiguation, Semantic Role Labeling, Semantic Parsing e Machine Translation.

La chiave della ricerca sarà il fatto che le reti neurali profonde (DNN) di grandi dimensioni ma attivate in maniera sparsa, cioè che utilizzano un numero notevolmente inferiore di parametri, possono consumare una piccola frazione dell'energia delle DNN grandi e dense senza doverne sacrificare la precisione. Gli approcci attuali per rendere i modelli più efficienti si sono concentrati sulla fase di preaddestramento, in modo da preservare le prestazioni sulle baseline di NLU come GLUE. Tuttavia, c'è la necessità di applicare modelli efficienti a una gamma ampia di compiti più complessi a valle dell'NLU e di sviluppare metodi per rendere i modelli più efficienti pur ottenendo prestazioni competitive. Per fare ciò, il candidato dovrà comprendere adeguatamente come sviluppare nuove architetture che, partendo dalle tecniche esistenti di Deep Learning basate sul Pruning o sulla Distillazione, propongano nuovi metodi che si concentrino su particolari compiti NLU, permettendo l'uso di grandi modelli linguistici nel modo più efficiente e meno dispendioso di energia.

Bibliografia

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of

the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, FangWang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural languageunderstanding. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4163–4174, Online.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms.

David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. CoRR, abs/2104.10350.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized wordrepresentations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. Proceedings of the AAAI Conference on Artificial Intelligence, 34(09):13693–13696.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu,Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, XuanOuyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. CoRR, abs/2107.02137.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is allyou need. In Advances in Neural Information Processing Systems, volume 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium.

Ziheng Wang, Jeremy Wohlwend, and Tao Lei. 2020. Structured pruning oflarge language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6151–6162, Online.

Titolo del progetto (inglese): Eco-sustainable deep learning for large-scale Natural Language Processing

Progetto di ricerca (inglese):
Language modeling has become the cornerstone of current Natural Language Processing in that they encode much linguistic information in latent form and enable text understanding and generation on a large scale, also across languages (e.g. enabling state-of-the-art machine translation or dialogue). However, these models have grown considerably over time, with hundred millions of parameters (e.g. BERT) and, more recently, several billion parameters to be trained (e.g. GPT-3 and WuDao). Unfortunately, the carbon emissions for training such models is very important (Patterson et al., 2021). It has been estimated that Artificial Intelligence systems, and in particolar NLP models, have a meaningful carbon footprint today (https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/?sh=5bcadc756b43): training a single deep learning model can generate up to 626,155 pounds of CO2 emissions, more or less the total carbon footprint of five cars over their entire lifetime. As a point of

comparison, an American produces on average around 36,000 pounds of CO2 emissions in a year (Strubell et al., 2020). Unfortunately, this trend can only get worse and adds to the current worries about climate change (https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/?sh=5bcadc756b43).

This trend is motivated by the need to move further beyond static word embeddings into contextualized ones led to the use of recurrent neural architectures (e.g. BiLSTMs) to obtain generalized contextualized embeddings and the appearance of ELMO (Peters et al., 2018). This journey found its peak with the first Transformer architecture (Vaswani et al., 2017) that exploited the Attention mechanism to break from recursive approaches and enabled the crafting of the first pre-trained Transformer model, BERT (Devlin et al., 2019).

Similar to what happened at the beginning of the last decade in computer vision with the emergence of ResNets, Transformer models have taken the scene of NLP, thanks to their adaptability into new tasks by leveraging them with transfer learning. By being pre-trained on massive amounts of heterogeneous text, they can generalize when facing new tasks that seemed impossible with previous approaches.

The well-known GLUE leaderboard (Wang et al.,2018) tracks performance on an array of Natural Language Understanding (NLU) tasks. BiLSTMs reached a performance of 64.2 on average. ELMO brought it up to 67.7 points, 70 if combined with Attention. Within the last three years since the appearance of BERT, current Transformer based architectures have surpassed human performance. At the time of writing, ERNIE 3.0 (Sun et al.,2021) leads it with 91.1 points, more than 20 points over the ELMO baseline.

However, Ernie 3.0 has 10 billion parameters, while ELMO has 93.6 million. Despite performance increases with model size, this trend has an obvious downside. Training and performing inference with these models become more energy-consuming with their size. While computer components and resources have grown more powerful and efficient, the growth tendency of large Language Models surpasses their pace. Moreover, the models often require specialized hardware such as GPUs or TPUs. Combined with the price and shortage of such components, it can become a challenge for researchers and enterprises to use them (Patterson et al.,2021).

Therefore there is an urgent need for lighter and more efficient Language Models, and fast and cheap methods to adapt them to new domains and tasks.This need has already motivated a growing interest in Deep Learning approaches that will make the models lighter or more efficient. For instance, pruning, where the ”lottery ticket” hypothesis is put into practice and certain parts of the model are dropped without loss of performance, aims to make a language model smaller (Wang et al.,2020). Distillation has also gained traction, where a larger language model is used as a ”teacher” model, and its knowledge is distilled to a smaller one, trying to sacrifice the least amount of performance (Jiaoet al.,2020). Moreover, on the architecture side, one of the main bottlenecks of Transformer models is the Attention mechanism and the quadratic complexity (O(N2)) of its computation. Several attempts have been made to ease the computation complexity of attention by substituting it with Fourier transformations (Lee-Thorp et al.,2021), or making it linear on the sequence length by masking parts of the attention mechanism (Beltagy et al. 2020).

The goal of this Ph.D. will be to provide efficient ways to employ and train Language Models for NLU tasks, with a particular focus on those developed at the Sapienza NLP lab led by prof. Roberto Navigli, such as Multilinguality, Word Sense Disambiguation, Semantic Role Labeling, Semantic Parsing and Machine Translation. Key to the research will be the fact that large but sparsely activated deep neural networks (DNNs), i.e. using a considerably lower number of parameters, can consume a small fraction of the energy of large, dense DNNs without having to sacrifice accuracy. The current approaches to making models more efficient have focused on the pretraining phase, such that they preserve performance on the NLU baselines such as GLUE. However, there is a need to apply model efficiency to a higher range and more complex downstream tasks and develop methods to make models more efficient while

performing competitively.To do so, the applicant will need to properly understand how to develop new architectures that, starting from existing Deep Learning techniques based on Pruning or Distillation, will put forward new methods that focus on particular NLU tasks, enabling the use of large Language Models in a more efficient and less energy-consuming way.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, FangWang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural languageunderstanding. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4163–4174, Online.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms.

David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. CoRR, abs/2104.10350.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized wordrepresentations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. Proceedings of the AAAI Conference on Artificial Intelligence, 34(09):13693–13696.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu,Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, XuanOuyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. CoRR, abs/2107.02137.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is allyou need. In Advances in Neural Information Processing Systems, volume 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium.

Ziheng Wang, Jeremy Wohlwend, and Tao Lei. 2020. Structured pruning oflarge language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6151–6162, Online.