Knowledge Neurons in Multilingual Pretrained Language Models

Yinghao Cai
Southeast University
Email:yinghaocai@foxmail.com

I. Abstract

Large-scale pretrained language models perform impressively in storing knowledge [1, 2] . The paper, [3] proposes a concept, knowledge neurons and the method to extract them. Besides, the paper also studies the application of knowledge neurons. However, the study is limited in the model pretrained on a single language while in practice, knowledge is often expressed multilingually. In this paper, I aim to study the generation of the method of extracting knowledge neurons in large-scale pretrained multilingual language model.Caparing the results of experiments on single-language dataset and multilingual dataset, I find that there is a barrier between different languages. Also, the application of the knowledge neurons is also verified. The code is available at https://github.com/Fuyao233/Knowledge-Neurons-in-Multilingual-Pretrained-Language-Models.

II. Introduction

LARGE-SCALE pretrained language models[4] are believed to be able to storage factual knowledge showed in the training corpus. [1, 2] Hence,pretrained language models can play a role as the knowledge base through predicting text.[2] Recently, the research in [5] shows a huger language models have the stronger ability to storage factual knowledge. Nevertheless, the work mainly focuses on the prediction accuracy and the paper, Knowledge Neurons in Pretrained Transformers, focuses on the details how the factual knowledge is stored in Transformers,spicifically in the Forward Feedback Neural network(FFN).

In [3],the author views the FFN parts(i.e.,two-layer full connected network) in Transformer as a special self-attention layer, which can memory factual knowledge through key-value mechanism.[6] Inspired by this idea,the author introduces the knowledge neurons, a group of neurons expressing the relational fact and raises the attribution method to identify these neurons. Figure 1 shows the identified knowledge neurons.

In the subsequent research, the author researches on the knowledge neurons' effects on knowledge expression. It is found that suppressing and reinforcing the activation of knowledge neurons will significantly affect the probability predicted by the Transformer. Also, the activation responds more to the corresponding factual knowledge.

Knowledge neurons are also used to leverage the stored knowledge. In [3], the author presents two interesting applications: edit facts and erase relations without fine-tuning.

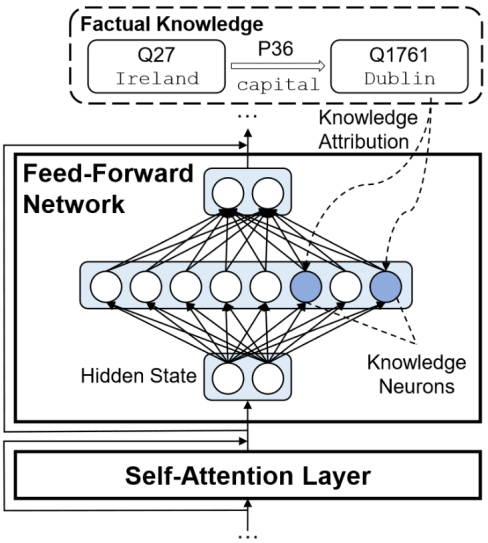

Fig. 1. Knowledge neurons in FFN

However, study of the knowledge neurons is limited in the language models pretrained on the specific language while knowledge in practice is both written and queried in many languages. In my study, the distribution of knowledge neurons in multilingual pretrained language models, such as XLM [7], X-FACTR [8]. To properly handle language variations, I adopt the methods in [8] to expand probing methods from single- to multi-word entities, and the decoding algorithms to generate multi-token predictions.

My study focuses on three questions:

- Are knowledge neurons activated by the same knowledge expressed by different languages ? And what about the various prompts in single language?
- Do knowledge neurons in multilingual language models gather in the top blocks either?
- Will the knowledge stored in the multilingual language model be modified operating on the knowledge neurons?

To answer these questions, experiments are conducted on the pretrained language model, BERT-based-multilingual and two datasets, PARAREL_en and T-REx_multilingual. In experimental part, I will introduce these two datasets in detail.

To solve the first question, I extract the knowledge neurons in multilingual BERT with two datasets and find the answer. The knowledge neurons will respond to not only the facts

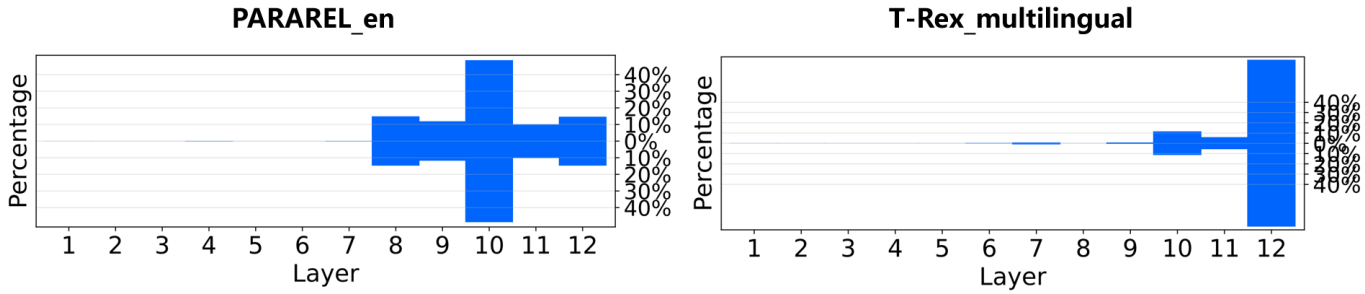

Fig. 2. Percentage of knowledge neurons identified by our method in each mulitilingual BERT layer.The left shows the distribution on the PARAREL_en and the right on the T-Rex_multilingual.

expression by different languages but also the various prompts. To solve the second question, I draw the distribution of the knowledge neurons and conclude that most knowledge neurons are distributed in the top layers of the model. The finding agrees with [3]. To solve the third question, I follow the idea from [3]. The paper shows that the activation of knowledge neurons is positively related to the knowledge expression. So I modify the activation of the extracted knowledge neurons and compare the prediction.

The conclusion of this paper, namely the answers of the above questions can be listed as follow:

- Knowledge neurons in multilingual pretrained language model are activated by not only the prompts in one specific language but also multilingual prompts. However, the knowledge neurons shared by multilingual prompts are less than them by single-language.
- Most knowledge neurons in multilingual language models are distributed in the top layers in multilingual language model. And the knowledge neurons in multilingual language model are distributed in higher layer.
- Knowledge stored in the multilingual language model can also modified without any fine-tune.

III. Background

A. Connection between FFN and self-attention in Transformer[9]

Transformer[9] is a one of the most popular encoder-decoder architectures in NLP. It consists of encoder and decoder. Its encoder has 6 identical blocks in which there are two sub-layer, self-attention layer and forward feedback network. For the input $X \in \mathbf{R}^{N \times d_{model}}$, the output of self-attention layer and FFN are as follows:

$$Q_h = XW_h^Q, K_h = XW_h^K, V_h = XW_h^K, \tag{1}$$

$$Self - Att_h(X) = softmax(\frac{Q_h K_h^T}{\sqrt{d_{model}}})V_h, \tag{2}$$

$$FFN(X) = gelu(XW_1)W_2 \tag{3}$$

where $W_h^Q, W_h^K, W_h^V, W_1, W_2$ are parameters and $gelu(\cdot)$ is the activation function raised in [10].

Except the difference in activation function, $Self - Att_h(\cdot)$ and $FFN(\cdot)$ are similar according to Equation 2 and Equation 3. So the input of FFN can be seen as queries and the two layers of FFN can be been as keys and values.[6]

B. Multilingual Language Model

Although the language model, BERT can be pretrained on more than one hundred languages, the information expressed in different languages shows significant difference. Hence, knowledge is not shared among models trained on different languages. Studies [7][8] aim to propose proper training methods of multilingual language models.

Multilingual models benefit the downstream tasks, such as translation or text classification in low-resource languages by leveraging information learned from other languages.

IV. Identification of Knowledge neurons

A. Knowledge attribution

Inspired by [11], the attribution method proposed is based on integrated gradients [12].

First, the probability of the correct knowledge in prediction for the prompt, x is denoted as follow:

$$\mathrm{P}_x\left(\hat{w}_i^{(l)}\right) = p\left(y^* \mid x, w_i^{(l)} = \hat{w}_i^{(l)}\right) \tag{4}$$

where $w_i^{(l)}$ denotes the $i-th$ intermediate neuron in the $l-th$ FFN; $\hat{w_i}^{(l)}$ is a given constant assigned to $w_i^{(l)}$. And the attribution score is calculated as follow:

$$Attr\left(w_i^{(l)}\right) = \bar{w}_i^{(l)} \int_{\alpha=0}^{1} \frac{\partial \mathrm{P}_x\left(\alpha \bar{w}_i^{(l)}\right)}{\partial w_i^{(l)}} d\alpha \tag{5}$$

where $\bar{w_i}^{(l)}$ is the original value of the neuron. In the integrated equation, it continuously change $w_i^{(l)}$ from 0 to $\bar{w_i}^{(l)}$. The equation accumulates the influence caused by the neuron. The larger score means a neuron has, the more significant it is in the knowledge expression.

In practice, Riemann approximation is used to replace the continuous integrals. The final formula is as follow:

$$\tilde{Attr}\left(w_i^{(l)}\right) = \frac{\bar{w}_i^{(l)}}{m} \sum_{k=1}^{m} \frac{\partial \mathrm{P}_x\left(\frac{k}{m} \bar{w}_i^{(l)}\right)}{\partial w_i^{(l)}} \tag{6}$$

m, the number of approximation steps is usual 20.

Relations	Template #1	Template #2	Template #3
P176 (manufacturer)	[X] is produced by [Y]	[X] is a product of [Y]	[Y] and its product [X]
P463 (member_of)	[X] is a member of [Y]	[X] belongs to the organization of [Y]	[X] is affiliated with [Y]
P407 (language_of_work)	[X] was written in [Y]	The language of [X] is [Y]	[X] was a [Y]-language work

TABLE I
EXAMPLE PROMPT TEMPLATES OF THREE RELATIONS IN PARAREL. [X] AND [Y] ARE THE PLACEHOLDERS FOR THE HEAD AND TAIL ENTITIES, RESPECTIVELY.

B. Knowledge Neuron Refining

The attribution scores will be compared with the threshold t and the neurons with larger values than t are collected in the coarse set of knowledge neurons. These neurons include some "false-positive" neurons which may response for other reasons. For knowledge neurons refining, the author assumes that different prompts of the same fact share the same knowledge neurons. Based on the assumption, diverse formats of prompts are adopted and the knowledge neurons shared by more than $p\%$ prompts are retained.

V. EXPERIMENTS

A. Experimental settings

The experiments are conducted on BERT-base-multilingual-uncased [4], a widely-used language model pretrained on on the top 102 languages with the largest Wikipedia using a masked language modeling (MLM) objective. Similarly as other BERT-base models, it contains 12 Transformer blocks, where the hidden size is 768 and the FFN inner hidden size is 3,072. For each prompt, the attribution threshold t is set to 0.2 times the maximum attribution score. For each relation, the initial refining threshold p% (Section IV.B) is 0.7. Then, the threshold is increased or decreased by 0.05 at a time until the average number of knowledge neurons lies in [2, 5]. The experiments are conducted on a NVIDIA Tesla V100 GPU with 16GB. It takes 26.3 minutes on average to refine knowledge neurons from a specific relation.

B. Datasets

Experiments are conducted on two datasets, PARAREL_en and T-Rex_multilingual .

PARAREL_en is a single-language datasets cited from [3]. It is built on the PARAREL dataset [13]. PARAREL is curated by experts, containing various prompt templates in English for 38 relations from the T-REx dataset [14]. Some prompt templates are shown in Table I.For each relational fact, the head entity in prompt templates is filled and left the tail entity as a blank to predict. In order to guarantee the template diversity, relations with fewer than 4 prompt templates are filtered out and finally kept 34 relations, where each relation has 8.63 different prompt templates on average. The dataset contains253,448 knowledge-expressing prompts in total for 27,738 relational facts.

T-Rex_multilingual is a multilingual datasets. It is built on the mTREXf [8]. In that paper, the author samples subject-object pairs with probability proportional to their frequency for each of the 46 relations in T-REx. For each subject and object, there are various expressions in different languages. Besides, the data from [8] provides multilingual prompt templates Table III. Then I choose 6 languages, English, Dutch, Japanese, Vietnamese, Chinese and Korean language to express a triplet. When producing data, I choose to mask the subject X or object Y with equal probability. And for better refining, I filter out the facts expressed by less than 6 languages. In total, there are 25,896 facts and 155,376 prompts.

Type of Neurons	PARAREL_en	T-Rex_multilingual
Knowledge neurons	5.42	1.67
∩ of intra-rel. fact pairs	2.85	0.62
∩ of inter-rel. fact pairs	0.63	0.38

TABLE II
STATISTICS OF KNOWLEDGE NEURONS. ∩ DENOTES THE INTERSECTION OF KNOWLEDGE NEURONS OF FACT PAIRS. "REL." IS THE SHORTHAND OF RELATION.

C. Baseline and Statics of knowledge neurons

The baseline method calculating the attribution scores adopted in the paper is intuitive, i.e. $Attr_{base}(w_i^{(l)}) = \bar{w}_i^{(l)}$, measuring the sensitivity to the input. The method based on neuron activation is a reasonable baseline. It is motivated by FFNs's analogy with the self-attention mechanism , because self-attention scores are usually used as a strong attribution baseline.[15]

I extract knowledge neurons from the both two datasets and the distribution of the knowledge neurons is in Figure 2 . The figure shows that the knowledge neurons in multilingual language model respond to not only the prompts in a single language but also multilingual prompts. And most of the knowledge neurons are distributed in the top layers of the model. And for T-Rex_multilingual, most of its knowledge are distributed in the 12th layer and for PARAREL_en, the layer containing most knowledge neurons is the 11th layer.

In statistics, on PARAREL_en, there are 5.42 knowledge neurons for each fact and 1.67 on T-Rex_multilingual. The interaction is also computed and listed in the Table II. The intra-rel is the number of knowledge neurons shared by facts in the same relation and the inter-rel is the number of knowledge neurons shared by different relations. For PARAREL_en, the intra-rel is 2.85 and the inter-rel is 0.63 and for T-Rex_multilingual, the intral-rel is 0.62 and the inter-rel is 0.38. The result shows that most facts of the same relation share knowledge neurons and facts of different relation share few knowledge neurons. Besides, the number of knowledge neurons and intrea-rel extracted from T-Rex_multilingual is less than it from PARAREL_en. That means facts expressed

Relations	en	nl	fr
P19(place of birth)	[X] was born in [Y]	[X] werd geboren in [Y]	[X] est [né;X-Gender=MASC—née;X-Gender=FEM] [PREPLOC;Y] [Y]
P69 (educated at)	[X] was educated at the University of [Y]	[X] is opgeleid aan de Universiteit van [Y]	[X] a fait ses études à [ARTDEF;Y] [Y]
P39(position held)	[X] has the position of [Y]	[X] heeft de positie van [Y]	[X] est [Y]

TABLE III
MULTILINGUAL PROMPT TEMPLATES OF THREE RELATIONS IN MTREXF. [X] AND [Y] ARE THE PLACEHOLDERS FOR THE HEAD AND TAIL ENTITIES, RESPECTIVELY.EN MEANS ENGLISH, NL MEANS DUTCH, RU MEANS RUSSIAN.

Prompt Types	Knowledge Neurons	Random Neurons
Change rate↑	38.90%	2.30%
Success rate↑	34.70%	0
ΔIntra-rel. PPL↓	17.2	45.2
ΔInter-rel. PPL↓	14.3	37.9

TABLE IV
CASE STUDIES OF UPDATING FACTS. ↑ MEANS THE HIGHER THE BETTER, AND ↓ MEANS THE LOWER THE BETTER. "REL." IS THE SHORTHAND OF RELATION.

by single language share more knowledge neurons than the facts expressed multilingually.

VI. CASE STUDIES

I study the applications of knowledge neurons, update knowledge, to verified the effectiveness of knowledge neurons in multilingual language model. [3].

A. Update knowledge

Updating knowledge means that the original fact $< h, r, t >$ is changed into $< h, r, t' >$.

1) Methods: After identifying the knowledge neurons of the fact $< h, r, t >$, the author directly makes modification on the value slot as follow:

$$FFN_i^{(val)} = FFN_i^{(val)} - \lambda_1 \mathbf{t} + \lambda_2 \mathbf{t}'$$

where $FFN_i^{(val)}$ is the value slot of the $i-th$ knowledge neuron; $\mathbf{t}$ and $\mathbf{t}'$ are embedding vector of the entities t and t'; λ_1 and λ_2 are hyperparameters. λ_1 and λ_2 are set to 1 and 8 in my experiments.

2) Setup and Evaluation metrics: Two metrics are proposed to asses the fact updating. One is change rate, the ratio that the original prediction of t is changed and the other is the success rate, the ratio that the modified prediction t' become the prediction result. There are two other metrics to evaluate the influence on other knowledge: Δintra-relation PPL and Δinter-relation PPL.

3) Results and analysis: I conduct this part on T-Rex_multilingual. Results are listed in Table IV. The modification of knowledge neurons affects the prediction effectively while the modification of random neurons is insufficient.

VII. CONCLUSION

This paper inspired by [3] studies the generalization of the methods to extract knowledge neurons in multilingual language model and the effectiveness of the knowledge neurons. At first, I propose three questions about the knowledge neurons in the multilingual language model. To solve these questions, I cite the single-language dataset PARAREL_en from [3] and construct the multilingual dataset T-REx_multilingua from[8]. Experiments on multilingual language model show the distribution of the knowledge neurons leading to conclusion that the knowledge neurons shared by multilingual prompts are less than by single-language and the multilingual knowledge neurals tend to gather in the higher layer of the model. Besides this paper also verifies the application of the multilingual knowledge neurons.

REFERENCES

[1] Fabio Petroni et al. "Language Models as Knowledge Bases?" In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 2463–2473.

[2] Zhengbao Jiang et al. "How can we know what language models know?" In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 423–438.

[3] Damai Dai et al. "Knowledge Neurons in Pretrained Transformers". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 8493–8502.

[4] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*. 2019, pp. 4171–4186.

[5] Adam Roberts, Colin Raffel, and Noam Shazeer. "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 5418–5426.

[6] Mor Geva et al. "Transformer Feed-Forward Layers Are Key-Value Memories". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 5484–5495.

[7] Alexis Conneau and Guillaume Lample. "Cross-lingual language model pretraining". In: *Advances in neural information processing systems* 32 (2019).

[8] Zhengbao Jiang et al. "X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models". In: ().

[9] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[10] Dan Hendrycks and Kevin Gimpel. "Gaussian error linear units (gelus)". In: *arXiv preprint arXiv:1606.08415* (2016).

[11] Yaru Hao et al. "Self-attention attribution: Interpreting information interactions inside transformer". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 14. 2021, pp. 12963–12971.

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[13] Yanai Elazar et al. "Measuring and improving consistency in pretrained language models". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 1012–1031.

[14] Hady Elsahar et al. "T-rex: A large scale alignment of natural language with knowledge base triples". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[15] Olga Kovaleva et al. "Revealing the Dark Secrets of BERT". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.