ABCI 3.0開発加速利用 (2024年度) 成果概要 (公開用)

課題名: Towards Interpretable Foundation Models: Sparse Auto- Encoder-Based Transformer Architectures 自然言語理解チーム 代表者氏名: 乾健太郎	月20日から2025年3月31日まで 研究所革新知能統合研究センター ^{IB}
---	--

成果概要: Transformer-based neural network architectures are the key technology underlying the rapid progress of generative AI. However, the internal computational processes of these models is difficult to understand and interpret. Previous work has predominantly relied on post-hoc interpretation methods, such as training sparse auto-encoders (SAEs). SAEs disentangle internal representations into more easily interpretable features but suffer from a low degree of internal validity. Here, we developed a SAE-like component that can inserted at points of interest in the Transformer architecture. This SAE-based Transformer is then trained from scratch, resulting in a model that is inherently interpretable at the insertion points.

成果のポイント: Our main result achieved during the accelerated development program is a proof-of-concept which shows that it is possible to insert sparse autoencoder-like components directly during pretraining of transformer-based neural networks. Since naïve insertion of the SAE-like components degraded performance, we developed strategies for improving and speeding up the convergence the training of SAE-based Transformer architectures. The resulting highly interpretable models incurred only a small performance

Layer 6 - Neuron 2386 - Example 1 - Rank 2051 - Token Entropy 6.6562

Entropy: 6.6562

<s> Start ling __new __finding : __ 6 0 0 __million __years __ago ; __ a __bi ological __m ish ap __chi __of __the __University __of __Oregon 's __Institute __of __M ole cular __Bi ology ; __discuss es __his __animals . <0x0A> (You Tube / Univers ity __of __O reg an) __If __life __is __effectively __an __end 1 __passed __on __from __one __being __to __the __next , __then __evolution __is __the __high - st ake __wrong , __and __you 'II __live __but den ed __by _a __mal ada pt ive __mut ation __or __gen ettic _ ary __history __of __life . __A __New __Th erm od ynam ics __Theory __of __the __Origin __of __life . __a __prim ord ial _soup , __a __b olt __of __light ning __and __a __col oss al __stroke __of __luck , __ __have __little __to __do __with __it . __Ung ulate __E volution __is __true , _why __do __we __still <0x0A> Where _are __all __the __mon key - men _I __was __promised ? __" __Well , __if __you __o __change __over __time __without __proof __on __a __mon key - man __level , __here _are __a __bui __on .__Well , __seven __anyway .__Ele ph ants _are __E vol ving __to __to se __Their __T us ks __revealed __in __detail __species , __has __been __revealed __in __un pre ced ented __detail __in __g <s> __We __need __strong __leadership __from __leaders __and __significant __role __models <0x0A> W omen 's __Econom ic __Security <0x0A> The __persistent __gender __pay __gap , __work p __in __the __work force __due __to __car ing __respons ib ilities , __comp ounds __throughout __their __ _sav ings . <0x0A> - _The _gender _pay _gap _in _Australia , _the _difference _between _v 5 % _(\$ 1 5 , 1 7 6 _p . a .). <0x0A> - _Women _around _the _age _of _ 4 5 _tend _to _have __males __of __similar __age __and __income __level __due __to __breaks __from __paid __work __for _super ann u ation _balance _for _ret iring _women _was _around _ 6 5 % _of _the _mediar _approaching _ 5 5 - 6 4 _years) _and _\$ 1 8 3 , 0 0 0 _for _men . <0x0A> - _The _Australian _and _occupation , _a _pattern _that _has _pers isted _over _the _past _two _dec ades . _ <0x0A> This __economic __dis adv antage __leaves __women __at __greater __risk __of __pover ty __a _type _of _work _that _women _tend _to _eng age _also _means _they _are _more _like __COVID - 1 9 . <0x0A> The __concentration __of __women __in __in secure __employ ment __and __c ations , __means __their __jobs __are __more __vulner able __and __less __likely __to __have __ade qu __in __the __inform al __economy . __In __Australia , __women __are __more __likely __to __work __cas _flex ibility _to _undert ake _car ing _respons ib ilities . <0x0A> Occ up ational _gender _seg reg __un emp loyment __during __COVID - 1 9 __based __on __their __over re presentation __in __more __

__is __seeing __large __numbers __of __lay - offs , __including __in __ret ail , __hospital ity __and __tou

Entropy: 5.8438

penalty that we were able to offset by slightly increasing the model size. Similarly to SAEs, the features learned by our SAE-like components appear to encode interpretable concepts such as "words related to biology and evolution" (left figure) or "words related to work and employment" (right figure). In addition to interpreting the outputs of the fully trained model, the sparsity of the activations in the SAE-like components also allows tracking how these features develop and gradually specialize during training.

成果についてより詳細な情報を提供しているWebページ、発表論文などの情報:n/a