

How Far Does BERT Look At: Distance-based Clustering and Analysis of BERT’s Attention

Yue Guan^{1*†}, Jingwen Leng^{1*†}, Chao Li^{2*†}, Quan Chen^{2*†}, Minyi Guo^{2*†}

^{*}Shanghai Jiao Tong University

[†]Shanghai Qi Zhi Institute

¹{bonboru, leng-jw}@sjtu.edu.cn

²{lichao, chen-quan, guo-my}@cs.sjtu.edu.cn

Abstract

Recent research on the multi-head attention mechanism, especially that in pre-trained models such as BERT, has shown us heuristics and clues in analyzing various aspects of the mechanism. As most of the research focus on probing tasks or hidden states, previous works have found some primitive patterns of attention head behavior by heuristic analytical methods, but a more systematic analysis specific on the attention patterns still remains primitive. In this work, we clearly cluster the attention heatmaps into significantly different patterns through unsupervised clustering on top of a set of proposed features, which corroborates with previous observations. We further study their corresponding functions through analytical study. In addition, our proposed features can be used to explain and calibrate different attention heads in Transformer models.

1 Introduction

With the rapid development of neural network in NLP tasks these year, the Transformer (Vaswani et al., 2017) that uses multi-head attention (MHA) mechanism is one recent huge leap (Goldberg, 2016). It has become a standard building block of recent NLP models. The Transformer-based BERT (Devlin et al., 2018) model further advances the model accuracy by introducing pre-training and has reached the state-of-the-art performance on many NLP tasks.

Beyond the general explanation which attributes the effectiveness of BERT model to its capability of long-range dependency and contextual embeddings, more detailed analysis of the MHA mechanism in BERT still remains an active research topic (Jain and Wallace, 2019; Serrano and Smith, 2019; Ethayarajh, 2019; Michel et al., 2019). Michael et al. (2020); Brunner et al. (2020); Coenen et al. (2019) explore the linguistic information or importance of BERT by inspecting hidden state embeddings. Wu et al. (2020); Roy et al. (2020) further investigate possible attention mechanism designs.

Several previous works on attention interpretability have analyzed the function and behavior of attention heads. Jawahar et al. (2019) finds that attention heads within a same layer tend to function similarly, and vast redundancy could be found in the attention heads. Clark et al. (2019) provides the study on relating the behavior of each head to linguistic merits such as dependency parser. Kovaleva et al. (2019) manually annotates attention heatmaps and trains a CNN to classify attention heads. Given these various analytical findings, from an algorithmic view, we are not aware of any research that provides a clear and reliable method to classify the attention heads without extra human intervention.

In this work, we provide a simple set of features that can be used to reliably disentangle the attention heads into different categories through clustering method. Different from heuristic analytics that looks at single inputs, our proposed attention head features are extracted corpus-wide and is thus more comprehensive in its discovered patterns. Interestingly, our algorithmically discovers patterns that match well with previous study (Kovaleva et al., 2019). On the other hand, our method could easily be scaled up to large datasets and is generalizable to other similar multi-head-attention-based models. With the use of the proposed clustering method, we further conduct empirical experiments to identify the important categories and parts of attention. The results could explain previous MHA interpretability findings in terms of a distance view.

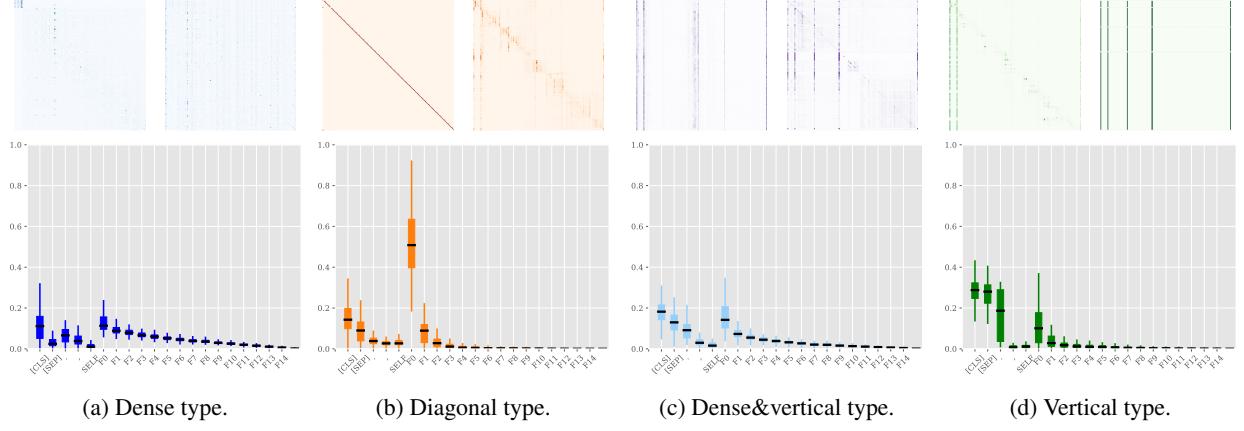


Figure 1: We extract the distance features and perform the K-means clustering of 384 attention heads in the BERT-large model. Top: two examples in each attention type. Bottom: the box-plot of 21-dimensional distance features in each type.

2 Background and Motivation

The BERT model (Devlin et al., 2018) includes multiple layers of Transformer (Vaswani et al., 2017) encoders. Its core component is multi-head attention mechanism (MHA). Given a sequence of input embeddings, the output contextual embedding is composed by the input sequence with different attention at each position. The attention weight is calculated as following,

$$\text{Attention}_{(Q, K)} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \quad (1)$$

where Q, K are query and key matrix of input embeddings, d_k is the length of a query or key vector. Multiple parallel groups of such attention weights, also referred as attention heads, make it possible to attend to information at different positions. BERT uses special tokens to encode the different input and output formats uniformly for downstream tasks, such as `[CLS]` and `[SEP]`.

Given the great success of BERT-like models, researchers start to explore the working mechanisms of multi-head attention component. One common approach is to visualize the attention weight matrix (Voita et al., 2019; Clark et al., 2019; Kovaleva et al., 2019), which often finds that attention heads in BERT demonstrate several specific patterns. For example, some attention heads show a dominant stripe pattern on diagonal direction while some others show a dominant vertical stripe pattern. There are also attention heads with a relatively homogeneous distribution of attention weights. In this work, we refer this patterns as dense, vertical or diagonal with some samples in the top part of Fig. 1.

However, these previous works rely on simple heuristic rules or human annotation, which makes them not scalable and error-prone. In this work, we propose a scalable and efficient unsupervised learning approach that automatically clusters different attention heads, which lets us systematically study their different roles in NLP tasks.

3 Distance Feature

Based on the attention pattern findings in previous works, we design a distance feature that enables unsupervised clustering of attention heads. For each input token, we first accumulate its attention weights to the special tokens including '`[CLS]`', '`[SEP]`', '`,`' and '`.`', because the vertical stripe pattern shows strong focus on these tokens. We then accumulate its attention weights to nearby tokens within a window size, which lets us distinguish the diagonal stripe with strong locality and dense pattern with weak locality. For a input sample with length of L , we divide the sample to N windows, each of which has a size of L/N . Eq. 2 calculates the window feature $F(i)$ of i_{th} window.

$$F(i) = \frac{1}{L} \sum_{s=1}^L \sum_{t=iL/N}^{(i+1)L/N} \text{Attention}(s, s \pm t) \quad (2)$$

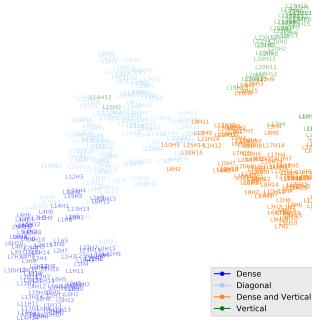


Figure 2: T-SNE visualization of K-means clustering with the distance feature.

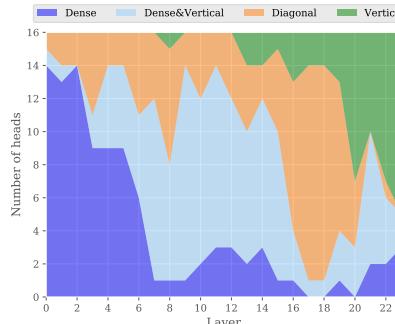


Figure 3: The attention head type distribution across different layers in the BERT-large model.

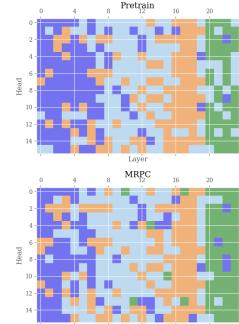
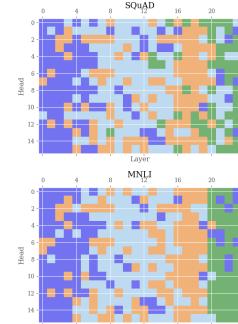


Figure 4: Clustering result on pre-trained language model and models fine-tuned on downstream tasks.



The special token attention and window attention are exclusive and we normalize both them by the input sequence length L . As such, their summation for a give input token remains one.

For an input sample, attention weight matrix from all tokens are averaged to get a low-dimensional representation of an individual head. With 16 windows, we can represent an attention head with a 21 (16 windows + self + 4 special tokens) dimensional feature for further analysis.

4 Clustering

We now describe how we leverage the 21-dimensional distance feature to apply an unsupervised attention head clustering. We study attention behavior in the BERT-large model¹. We randomly pick 1,000 samples from the SQuAD dataset (Rajpurkar et al., 2016) and extract the averaged distance feature across all selected samples, which is used to represent an attention head.

To verify if there exists clustering behavior in BERT, we first use a non-linear dimension reduction algorithm T-SNE (Maaten and Hinton, 2008) to visualize all 384 heads (24 layers \times 16 heads per layer). The result in Fig. 2 clearly demonstrates four clusters. We then use K-means clustering algorithm (Hartigan and Wong, 1979) with 4 clusters and the result is shown as the colors.

Fig. 1 shows two examples (top) and the distance feature box-plot (bottom) for each cluster. We also show the distribution of each type of attention head across different layers in Fig. 3. Our results show that an attention head shows a consistent weight pattern across different input samples and there exists a strong clustering behavior across different layers. We describe the key findings in details.

Dense: This type has dense attention patterns (Fig. 1(a) top). Fig. 1(a) bottom shows that the heads in this type have the highest attention at long range windows and relatively small attention weights to the special tokens. They account for 26.04% of total heads and mainly appear at the beginning layers.

Diagonal: This type has dominant diagonal patterns and a few vertical stripes (Fig. 1(b) top). Its distance features show large values in the short-range windows and medium values in the special tokens (Fig. 1(b) bottom). The short-range attentions can capture the local information while the vertical stripe can carry the global information. They account for 25.56% of total heads and appear at almost all layers.

Dense&Vertical: This type mixes the dense and vertical patterns (Fig. 1(c) top). Its distance features have large values at long range windows and special tokens. They account for 34.11% of total heads and mainly appear at middle layers.

Vertical: This type has mostly strong vertical patterns (Fig. 1(d) top). Its distance features have large values at special tokens, which are much higher than other types. They account for 13.28% of total heads and mainly appear at higher layers.

As a matter of fact, it is possible to exploit the proposed distance feature with other unsupervised clustering algorithms. Our work chooses to use the K-means clustering algorithm as a representative one. With Gaussian mixture model clustering (Améndola et al., 2015), we also get similar results. As such, we choose to use four clusters in the following experiment based on the T-SNE visualization result.

¹We use bert-large-uncased-whole-word-masking-finetuned-squad (Wolf et al., 2019) from <https://huggingface.co/models>.

| | Pre-train | SQuAD | MRPC | MNLI |
|-----------|-----------|--------|--------|--------|
| Stability | 95.98% | 96.26% | 95.98% | 94.37% |

Table 1: Stability of clustering measured following Von Luxburg (2010) on downstream tasks.

Moreover, we find the four clusters are human interpretable. We still get stable clustering results with 3 or 5 clusters, but the patterns can not be distinguished by human-being.

We measure the stability of clustering and generality on a pre-trained language model and downstream tasks including question answering, sentence pair similarity regression and natural language inference. For the aforementioned tasks, we use datasets that include SQuAD (Burger et al., 2001), MNLI (Williams et al., 2018), and MRPC (Dolan and Brockett, 2005) respectively. We also extract the features on fine-tuned BERT-large models accordingly. All of these training settings are inline with Wolf et al. (2019) and Devlin et al. (2018).

We evaluate the stability of clustering following Von Luxburg (2010). The idea is to obtain multiple models with partial training data and evaluate the consistent samples among all fitted models. We randomly split 50% training data to fit clustering models and predict on the whole dataset. The stability of clustering algorithm is calculated as the proportion of input samples that is consistent on all models fitted with the random partial training set split. We report the clustering stability of 10 times clustering in Tbl. 1. Our result demonstrates that over 95% attention heads are clustered stably with a limited amount of data. This shows that the proposed method is generally applicable and scalable.

The detailed results are shown in Fig. 4, which demonstrates similar type distribution among different models. In specific, all the four models share similar head type distribution described in Fig. 3. Empirically, we find many heads have the same cluster after fine-tuning process. However, some heads in the highest layers also change. We suggest that higher layers are more relevant to task specific knowledge while lower layers handles more general linguistic features. Particularly, pre-trained model and SQuAD-fine-tuned model have more dense&vertical types at highest layer, which we will show to be responsible for global information in Sec. 5. This is explained by the fact that question answering and language modeling are more difficult tasks as they use token embeddings from all positions for classification. On the contrast, the other two tasks only use the special token [CLS].

5 Experiments

Based on the attention clustering results, we now study which attention head type(s) is important and what range of information different head types process. In specific, we manipulate the different attention heads in BERT-large model and observe the resulted accuracy impact on SQuAD v1.1 development set. It should be mentioned that all these experiments are directly ablation without fine-tuning on the model parameters.

5.1 Importance of Head Types

We first prune or substitute the attention weight of a particular type to study its importance. We mask the attention value to 0 or substitute it with an uniformed attention value of $1/L$. Tbl. 2 reports the experimental results with different settings.

Firstly, we find that pruning all diagonal heads (3) makes the model totally inoperative, compared with (2,4,5). Pruning all vertical heads directly (5) has less than 1.5% accuracy loss. This indicates that Diagonal type is vital for BERT processing and vertical type is the opposite. On top of this, we further prune two types together with diagonal type always remained. Keeping dense&vertical type (6) together results in a better accuracy. Nextly, we find that substituting the dense type with uniformed attention (8) results in accuracy loss of less than 10%. This indicates that the attention values of dense are trivial. Substituting Vertical type (11) also has an accuracy of over 50%. However, substituting the mixed type dense&vertical (10) has a significant accuracy loss, which indicates its crucial role.

| | id | Dense | Diagonal | Dense&Vertical | Vertical | Acc. | F1 |
|------------|----|-------|----------|----------------|----------|-------|-------|
| Baseline | ❶ | ✓ | ✓ | ✓ | ✓ | 82.85 | 89.68 |
| Prune | ❷ | ✗ | ✓ | ✓ | ✓ | 79.92 | 87.62 |
| | ❸ | ✓ | ✗ | ✓ | ✓ | 4.59 | 7.55 |
| | ❹ | ✓ | ✓ | ✗ | ✓ | 73.33 | 81.92 |
| | ❺ | ✓ | ✓ | ✓ | ✗ | 81.42 | 88.57 |
| | ❻ | ✗ | ✓ | ✓ | ✗ | 75.31 | 84.05 |
| | ❼ | ✓ | ✓ | ✗ | ✗ | 56.87 | 67.62 |
| | ❾ | 1/L | ✓ | ✓ | ✓ | 74.68 | 84.17 |
| Substitute | ❿ | ✓ | 1/L | ✓ | ✓ | 1.27 | 6.07 |
| | ⓫ | ✓ | ✓ | 1/L | ✓ | 9.13 | 11.15 |
| | ⓬ | ✓ | ✓ | ✓ | 1/L | 53.17 | 59.09 |

Table 2: Experimental results of pruning or substituting attention matrix by behavior types. L is the length of input sequence such that $1/L$ is simple uniformed attention.

| | id | Dense intra | Diagonal intra | Dense&Vertical intra | Vertical intra | Acc. | F1 |
|------------|----|-------------|----------------|----------------------|----------------|------|----|
| Prune | ❷ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ |
| | ❸ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ |
| | ❹ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | ❽ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ |
| | ❾ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Substitute | ❿ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ |
| | ⓫ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ |
| | ⓬ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ |
| | ⓬ | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ |

Table 3: Experimental results of ablating attention matrix by distance. The upper part is ablating one type. The lower part is ablating all type but remaining one type.

5.2 Information in Important Heads

After identifying the diagonal and dense&vertical types as the important heads, we continue to study which part of information they are responsible for in terms of attending distance. We constrain attention matrix by sentences and remove the intra-sentence attention or inter-sentence attention. When the intra-sentence attention is removed, the attention head can only access tokens in other sentences and obtains global information. On the contrast, the attention head has only local syntactic or word level semantic information. We remove intra or inter sentence attention of one type to study the impact on the accuracy (❲-❽). Nextly, we keep intra or inter sentence attention of one type to see how much information is carried out by it (❾-⓬).

Tbl. 3 demonstrates the experimental results. Removing intra-sentence attention from diagonal type (❷) leads to a poor accuracy of 19.96%, while removing inter-sentence part (❸) has little interference. This proves that diagonal type focus on short distance information and such information is substantial for its important role in the model. Even with only diagonal head handling intra-sentence information (❽), the model has a descent accuracy of 71.80%. On the contrast, with dense&vertical attending other sentences solely (⓬), the model reaches an accuracy of 75.54%. This result shows that the BERT model accesses to long distance information with dense&vertical mostly. This indicates diagonal and dense&vertical type are responsible for local and global information respectively.

Finally, we draw a general view about the attention behaviors according to our analytic results in Sec. 5.1 and Sec. 5.2. Diagonal type is the most important and they focused purely on intra-sentence processing. Dense&Vertical type absorb information globally from other sentences. The vertical type is insignificant. Attention values of dense type are trivial and is substituted by uniformed attention without large interferences.

6 Conclusion

In this work, we propose a feasible and scalable unsupervised clustering method to classify the attention heads. The distance based feature well captures the observations of previous works and is utilized to conduct stable unsupervised clustering. We believe our method is helpful for attention interpretability study. With such taxonomy, we further apply analytic experiments to explore the function of each behavior according to distance. We are also looking forward to further analysis the behavior and function of variant patterns with probing task datasets (Conneau et al., 2018) and analytic tools (Qiu et al., 2019; Gan et al., 2020) as our next plan. Besides, there are several recent works focusing on the optimization of over-parameterized MHA mechanism (Michel et al., 2019; Kovaleva et al., 2019; Guo et al., 2020). Our results reveal the important types and components of each type empirically. With our explorations, we hope our insights could cast light on novel design of interpretable attention mechanism.

Acknowledgements* We thank the anonymous reviews for their thoughtful comments and suggestions. We would like to thank Zhouhan Lin for his valuable feedback on the clustering methods and suggestions about the evaluation, and thank Jianping Zhang with whom we have inspiring discussion on the interpretability of NLP models. This work was supported by Major Scientific Research Project of Zhejiang Lab (No. 2019DB0ZX01) and the National Natural Science Foundation of China (NSFC) grant (61702328, 61832006, and 61972247). Jingwen Leng and Minyi Guo are corresponding authors of this paper.

References

- Carlos Améndola, Jean-Charles Faugere, and Bernd Sturmfels. 2015. Moment varieties of gaussian mixtures. *arXiv preprint arXiv:1510.04654*.
- Gino Brunner, Yang Liu, Damian Pascual Ortiz, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers.
- John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, et al. 2001. Issues, tasks and program structures to roadmap research in question & answering (q&a). In *Document Understanding Conferences Roadmapping Documents*, pages 1–35.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.
- Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of bert. *arXiv preprint arXiv:1906.02715*.
- Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.
- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. 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)*, pages 55–65.
- Yiming Gan, Yuxian Qiu, Jingwen Leng, Minyi Guo, and Yuhao Zhu. 2020. Ptolemy: Architecture support for robust deep learning. *arXiv preprint arXiv:2008.09954*.
- Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420.
- Cong Guo, Bo Yang Hsueh, Jingwen Leng, Yuxian Qiu, Yue Guan, Zehuan Wang, Xiaoying Jia, Xipeng Li, Minyi Guo, and Yuhao Zhu. 2020. Accelerating sparse dnn models without hardware-support via tile-wise sparsity. *arXiv preprint arXiv:2008.13006*.
- John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.
- Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.
- Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. 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)*, pages 4356–4365.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.
- Julian Michael, Jan A Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. *arXiv preprint arXiv:2004.14513*.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024.
- Yuxian Qiu, Jingwen Leng, Cong Guo, Quan Chen, Chao Li, Minyi Guo, and Yuhao Zhu. 2019. Adversarial defense through network profiling based path extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

- Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2020. Efficient content-based sparse attention with routing transformers. *arXiv preprint arXiv:2003.05997*.
- Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.
- Ulrike Von Luxburg. 2010. *Clustering stability: an overview*. Now Publishers Inc.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. 2020. Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*.

A All Attention Matrices

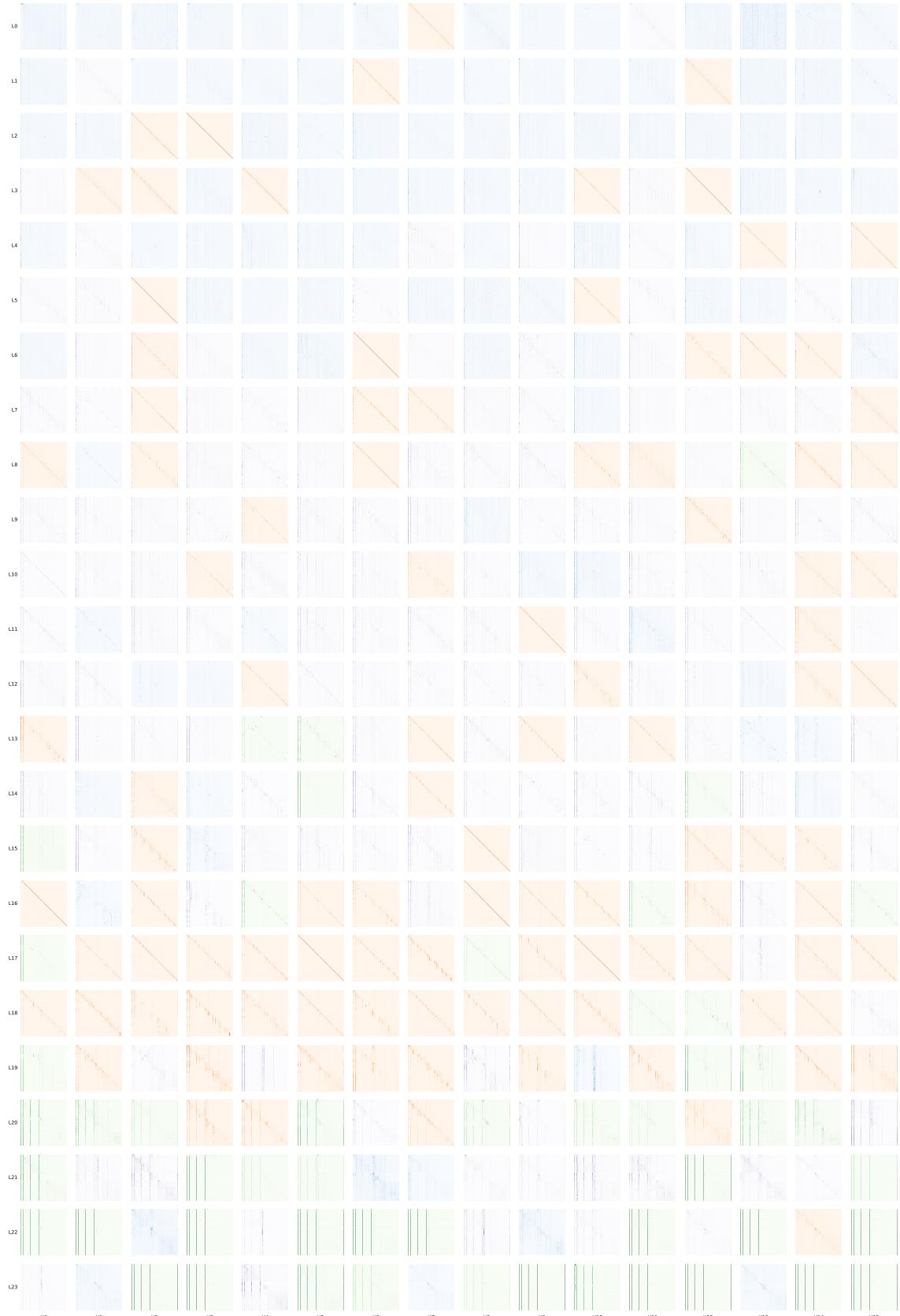


Figure 5: We visualize all attention matrices with a same randomly selected input sample as Fig. 5. Attention types clustered in Sec. 4 are distinguished by colors.