Fantastic Semantics and Where to Find Them: Investigating Which Layers of Generative LLMs Reflect Lexical Semantics

Anonymous ACL submission

Abstract

Large language models have achieved remarkable success in general language understanding tasks. However, as a family of generative methods with the objective of next token prediction, the semantic evolution with the depth of these models are not fully explored, unlike their predecessors, such as BERT-like architectures. In this paper, we specifically investigate the bottom-up evolution of lexical semantics for a popular LLM, namely Llama2, by probing its hidden states at the end of each layer using a contextualized word identification task. Our experiments show that the representations in lower layers encode lexical semantics, while the higher layers, with weaker semantic induction, are responsible for prediction. This is in contrast to models with discriminative objectives, such as mask language modeling, where the higher layers obtain better lexical semantics. The conclusion is further supported by the monotonic increase in performance via the hidden states for the last meaningless symbols, such as punctuation, in the prompting strategy.

1 Introduction

GPT-like large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023) have recently demonstrated impressive performance on various understanding and generative tasks, shifting from the pretraining-then-finetuning approach employed by BERT-like models (Zhao et al., 2023). However, existing research (Ethayarajh, 2019) suggests that the contextual representations of GPT-like models exhibit subpar performance in downstream tasks, struggling to fully capture the semantic nuances of words. This discrepancy raises a crucial research question: To what extent and through which layers do LLMs encode lexical semantics?

Previous research on intermediate layer representations in BERT has revealed important linguistic information, including its hierarchy. For instance, BERT encodes surface features at the bottom, syntactic features in the middle, and semantic features at the top (Jawahar et al., 2019). However, contextual representations in LLMs have received less attention due to structural differences and challenges, as illustrated in Figure 1. Firstly, LLMs employ a decoder-only strategy, which restricts their ability to access only preceding context during inference. Consequently, LLMs struggle to differentiate between homonymous meanings of words such as "bank" in the case of "the bank of the river" and "the bank to save money," due to the shared left context "the". Furthermore, LLMs are trained to predict the next token, resulting in varying degrees of comprehension of historical and predictive contexts across layers (Wang et al., 2023a; Voita et al., 2019). In contrast, BERT focuses on masked word restoration through mask language modeling (MLM), where both understanding and prediction processes are targeted for the same word.

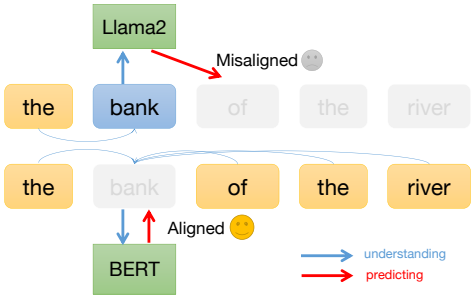

Figure 1: Key Differences between BERT and Llama2 Language Models. Blue and red lines indicate the information flows of understanding and predicting.

Given these observations, we hypothesize that GPT-like LLMs encode lexical semantics in lower layers while making predictions, potentially leading to the forgetting of information related to current tokens in higher layers. This hierarchical behavior suggests a dynamic interaction between understanding and prediction in generative LLMs, as indicated by the view of information flow in recent studies (Wang et al., 2023a; Voita et al., 2019).

To validate our hypothesis, this study delves into the examination of lexical semantics in LLMs by analyzing how the hidden states at each layer reflect word meanings. In particular, we investigate the understanding of lexical semantics in the popular open-source LLM, Llama2 (Touvron et al., 2023), utilizing the word in context benchmark (Pilehvar and Camacho-Collados, 2019). We employ various input transformation and prompting strategies to fully utilize the contextual information. The results suggest that lower layers of Llama2 capture lexical semantics, while higher layers prioritize prediction tasks. These findings offer practical insights into determining which layers of hidden states to utilize as representations of the meaning of the current word in GPT-like LLMs.

2 Related Work

2.1 Interpretability of language models

Interpretability of LLMs can be categorized into mechanistic (bottom-up) and representational (top-down) analysis (Zou et al., 2023). Mechanistic interpretability focuses on translating model components into understandable algorithms for humans, typically by representing models as computational graphs and identifying circuits with specific functions (Olsson et al., 2022; Geiger et al., 2021; Wang et al., 2022). On the other hand, representational analysis abstracts away lower-level mechanisms and explores the structure and characteristics of representations. Probing, an effective approach in top-down interpretability, can be classifier-based or geometric-based. Classifier-based probing trains additional classifiers for specific proxy tasks, including syntactic analysis (Hewitt and Manning, 2019), semantic roles (Ettinger, 2020), named entity recognition (Wang et al., 2023b), and world knowledge (Petroni et al., 2019). These linguistic features have demonstrated a rich hierarchy, spanning from lower layers to higher layers (Jawahar et al., 2019). Geometric probing without additional classifiers, examines the properties of the representational space itself. For example, difference vectors, obtained by subtracting base vectors, can detect linguistic features such as scalar adjective intensity (Garí Soler and Apidianaki, 2021) and stylistic features (Lyu et al., 2023). Furthermore, methods from the view of information flow indicate that models with autoregressive objectives (Voita et al., 2019) and specifically LLMs (Wang et al., 2023a) gather information in shallow layers while making predictions in deep layers.

2.2 Representations of Lexical Semantics

Lexical semantics, the study of word meanings, is a prominent field in both linguistics and computational research. Linguistics offers rich descriptive entries, known for their high dimensionality, contextual modulation, and discreteness (Petersen and Potts, 2023). Early rule-based models, including the Generative Lexicon (Pustejovsky, 1998) approach, used discrete feature representations. In contrast, neural models represent words as compact continuous vectors to avoid arbitrary feature selection. Static vector models, such as word2vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), and fastText (Mikolov et al., 2018), provide unified representations for all word occurrences. To distinguish word meanings in various contexts, especially for polysemous words, researchers have developed context-sensitive representations. Notable models include Elmo (Peters et al., 2018), BERT (Kenton and Toutanova, 2019), and the GPT family (Radford et al., 2019; Brown et al., 2020). While LSTM-based Elmo and transformer-based models offer bidirectional context around the target word, the GPT family focuses solely on context preceding the query word as a generative model. Large language models (LLMs) (Touvron et al., 2023; Brown et al., 2020) follow training mechanisms similar to GPT and have shown competitive performance via prompting engineering (White et al., 2023) compared to BERT-like models, e.g., in lexical tasks like word sense disambiguation (Kocoń et al., 2023) and named entity recognition (Wang et al., 2023b). Our research emphasizes evaluating the quality of representations in LLMs to enhance interpretability, rather than focusing on prompting strategies.

3 Experimental Design

3.1 Probing

We leverage the Word in Context (WiC) dataset as a proxy task for exploring lexical semantics (Pilehvar and Camacho-Collados, 2019)[1]. This well-structured benchmark presents a binary classification challenge - determining whether identical words convey the same meaning in distinct contexts. Our approach involves utilizing 638 instances from the development set to fine-tune the optimal hyperparameter, and assessing the final performance on

[1] https://pilehvar.github.io/wic/

setting	input
base	the **bank** of the river
repeat	the bank of the river the **bank** of the river
prompt	The bank in this sentence:"the bank of the river" means in one word **:**

Table 1: An example to show different input formats in three settings. **Bold** token positions are used as hidden states h_i of target words.

1400 instances from test set. We evaluate results based on accuracy and calculate accuracy separately for instances with different parts of speech.

3.2 Settings and Models

For a given word w within context c, Llama2 extracts hidden states $h_i \in \mathbb{R}^D$ across each of its 32 layers, where D is 4096 in Llama2. The cosine similarity of w in paired contexts (c^a, c^b) is calculated as s_w^{ab}. Subsequently, sentence pairs are classified as true if s_w^{ab} exceeds a threshold γ, and false if it falls below γ. The optimal γ is determined through development dataset, with distinct values potentially assigned for each layer to accommodate varying similarity ranges. The optimal values of γ are listed in Appendix A.1. To address potential anisotropy in the embedding space, we employ standardization across samples following prior research (Ethayarajh, 2019).

We employ different input variants for Llama2. The **base** setting uses the original context c with lexical representations h_i at the target position. Since w cannot access the context behind it in this setting, we repeat the original context and obtain h_i in the second context, ensuring all information is left of w. This configuration is referred to as **repeat**. Another setting is inspired by the prompting strategy proposed in the paper (Jiang et al., 2023). Here, we modify the context c as: *The w in this sentence: c means in one word :*. Then, we calculate the representation from the position of the last token, i.e., the final colon **:**, as h_i and we denote this as **prompt**. An example is provided in Table 1.

In order to compare autoregressive generative models with bidirectional models, we conduct experiments on BERT-large[2], which consists of 25 layers, a hidden dimension of 1024, and 336M parameters. Additionally, we consider other word-level contextualized embedding methods, such as WSD (Loureiro and Jorge, 2019), Context2vec (Melamud et al., 2016), and Elmo (Peters et al., 2018), as mentioned in the dataset paper (Pilehvar and Camacho-Collados, 2019)[3].

[2]https://huggingface.co/bert-large-uncased

Method	All	Noun	Verb
Human	80.0	-	-
Random	50.0	-	-
WSD	67.7	-	-
BERT_large†(23)	67.8	69.1	67.6
BERT_large (22)	**71.0**	**70.7**	**71.5**
Context2vec	59.3	-	-
Elmo	57.7	-	-
Llama2_base†(6)	60.9	62.6	62.1
Llama2_base (12)	62.9	63.9	62.5
Llama2_repeat†(8)	64.6	66.5	63.6
Llama2_repeat (7)	68.4	70.3	65.5
Llama2_prompt†(23)	68.9	68.9	69.5
Llama2_prompt (20)	69.2	69.2	69.5

Table 2: Overall accuracy (%) on the WiC test set. †indicates methods without anisotropy removal. The numbers in brackets after the model name indicate the number of layers for achieving the best performance.

4 Results and Analysis

Table 2 presents the overall performance. Llama2, as a generative model, achieves comparable results to bidirectional and non-regressive BERT models, outperforming non-transformer models like Elmo. This suggests that LLMs have the potential for word-level understanding, even though it is not explicitly trained for this capability. As expected, the prompting strategy achieves the highest accuracy among all the Llama2 variants. This approach incorporates downstream tasks into the generative process during LLM training and has proven to be popular and effective in addressing both intermediate and high-level tasks in the LLM era (Zhao et al., 2023). However, prompting relies on the choice of prompts and may not directly reveal the model's internal understanding. On the other hand, our repeat strategy demonstrates comparable performance to prompting and significantly outperforms the base version (with a 5.5 advantage gap). This simple yet effective transformation strikes a balance between information accessibility and prompting robustness.

In terms of parts-of-speech, nouns generally exhibit higher accuracy than verbs, as evidenced by a 4.8 advantage gap in Llama2_repeat. This observation aligns with previous studies that have

[3]It is important to note that we reproduce the result of BERT-large, which is relatively higher than the reported performance in the dataset paper.

concluded that verbs are more challenging to disambiguate (Barba et al., 2021).

Effectiveness of Anisotropy Removal. In Table 2, we compare methods with and without anisotropy removal (marked by †). The results consistently demonstrate the advantage of methods with anisotropy removal, suggesting that the representation space may collapse into a smaller cone space, as indicated by previous work (Ethayarajh, 2019). This also offers a simple and practical approach for calculating similarity in the embedding space.

Figure 2: Layer-wise accuracy for different settings and models. Star shows the best value.

Trends Across Layers. Figure 2 illustrates the layer-wise dynamics in two settings for Llama2 and also BERT_large. We observe non-monotonic trends for Llama2 across layers: both the base and repeat initially increase in lower layers before decreasing in higher layers. Consequently, optimal performance is achieved at lower layers when utilizing the hidden states of the target word as the default choice. This suggests that lower layers in LLMs encode lexical semantics, offering both a practical insight and a pathway for interpreting LLMs. Moreover, the trend contrasts with bidirectional BERT_large model, which obtains the best performance in higher layers. This highlights a difference between these two architectures: BERT concentrates on its current word across the layers while Llama2 aims for next token prediction.

Balancing Understanding and Prediction. To explore the balance between lexical understanding and predictive capability in Llama2, we computed the accuracy using representations of the previous token before the target (referred to as repeat_prev)

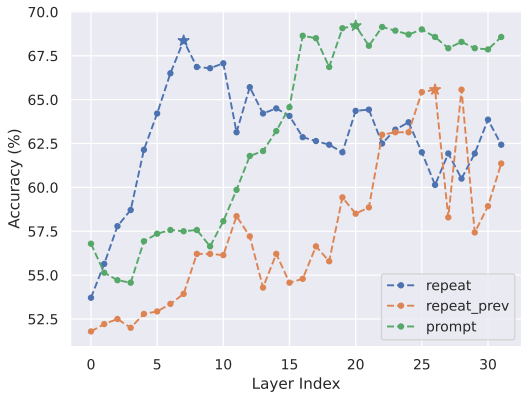

Figure 3: Layer-wise accuracy of Llama2 representations (repeat and prompt setting), as well as the previous token in the repeat setting (repeat_prev). The increasing trends observed in repeat_prev and prompt accuracies, as well as the non-monotonic trend observed in repeat accuracy, suggest that while the understanding ability may be weakening, the predictive ability is improving.

in the repeat setting. It is important to note that we opted for the repeat setting instead of the base setting, given that the base setting is constrained by incomplete information access. Furthermore, we conducted a comparison with the prompt setting, as depicted in Figure 3. Despite the fact that the representations do not originate from the correct target word but are anticipated to represent the next word, both repeat_prev and prompt exhibit a monotonic trend and comparable result across the layers. This observation suggests that while the understanding may diminish (as indicated by the inverted-U trend in the repeat setting) as layers go deeper, the predictive ability improves.

5 Conclusion

This study investigates how Llama2's layer-wise representations encode lexical semantics using the WiC dataset. Our experiments reveal that optimal performance is achieved at lower layers for generative tasks, while predictive accuracy improves in higher layers. This suggests that Llama2 prioritizes understanding before prediction as information flows from lower to higher layers. These findings offer practical guidance on extracting representations for lexical semantics tasks in engineering applications and shed light on the interpretability of LLMs due to their impressive performance.

6 Limitations

Probing offers a valuable viewpoint on lexical semantics, but it is still unclear what kind of semantics representations are exactly learned. Bridging the gap between dense, high-dimensional vectors from computational models and discrete, low-dimensional concepts from linguistic conventions remains an important issue to consider.

Another pressing issue is the narrow focus on only English and one large language model, namely Llama2. Different languages and models may yield varying effects on lexical semantic estimation. We anticipate that future studies will refine and complement our findings using a more diverse sample of natural languages and models.

7 Ethics Statement

We do not foresee any immediate negative ethical consequences of our research.

References

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Aina Garí Soler and Marianna Apidianaki. 2021. Scalar adjective identification and multilingual ranking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4653–4660, Online. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138.

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, Florence, Italy. Association for Computational Linguistics.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *arXiv preprint arXiv:2307.16645*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, page 101861.

Daniel Loureiro and Alipio Jorge. 2019. Liaad at semdeep-5 challenge: Word-in-context (wic). In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 1–5.

Qing Lyu, Marianna Apidianaki, and Chris Callison-burch. 2023. Representation of lexical stylistic features in language models' embedding space. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 370–387, Toronto, Canada. Association for Computational Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality.

Advances in neural information processing systems, 26.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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. Association for Computational Linguistics.

Erika Petersen and Christopher Potts. 2023. Lexical semantics with large language models: A case study of English "break". In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490–511, Dubrovnik, Croatia. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. 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)*, pages 2463–2473.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, pages 1267–1273.

James Pustejovsky. 1998. *The generative lexicon*. MIT press.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. 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 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *NeurIPS ML Safety Workshop*.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Appendix

A.1 Optimal Thresholds for each layer

We list the optimal thresholds for each layer in terms of three settings of Llama2, i.e., base, repeat and prompt in Table 3. They are searched according to the best performance in the development set of WiC dataset.

Layer Index	base	repeat	prompt
0	0.95	0.95	0.40
1	0.90	0.90	0.35
2	0.55	0.75	0.25
3	0.70	0.70	0.40
4	0.45	0.55	0.30
5	0.35	0.45	0.40
6	0.35	0.40	0.35
7	0.30	0.35	0.35
8	0.25	0.25	0.35
9	0.25	0.25	0.35
10	0.35	0.25	0.40
11	0.35	0.20	0.35
12	0.30	0.25	0.40
13	0.30	0.35	0.35
14	0.35	0.30	0.40
15	0.45	0.35	0.45
16	0.40	0.40	0.50
17	0.40	0.35	0.55
18	0.45	0.40	0.45
19	0.45	0.35	0.55
20	0.45	0.40	0.55
21	0.45	0.40	0.45
22	0.40	0.35	0.55
23	0.35	0.35	0.60
24	0.40	0.35	0.60
25	0.35	0.30	0.50
26	0.40	0.25	0.50
27	0.40	0.45	0.55
28	0.30	0.25	0.50
29	0.35	0.45	0.50
30	0.35	0.40	0.55
31	0.40	0.35	0.55

Table 3: Optimal thresholds for each layer.