Interpreting Transformers using Spectral Analysis

Stanford CS224N Custom Project

Tulika Jha
tulikaj@stanford.edu
Stanford University

Vishal Mohanty
vmohanty@stanford.edu
Stanford University

Rishu Garg
rishu@stanford.edu
Stanford University

Abstract

This study explores how spectral analysis can be used to interpret transformers. Just as the frequency spectrum gives us the range of frequencies contained by a signal, the frequency spectrum of a neuron's activation should break down information contained in a neuron into different frequencies, where the high-level frequencies consist of shorter-range word-level information and low-level frequencies consist of longer-range sentence-level information. Using spectral analysis as a tool, we examine the differences in information captured by different layers of a transformer and for various NLP tasks.

1 Key Information

- External mentor: Alex Tamkin (atamkin@stanford.edu)

2 Introduction

Interpretation of Transformer Models is a fast-growing research area in the field of Natural Language Processing, with many different approaches and techniques being developed. Due to the extensive application of transformers and their state-of-art performance in domains ranging from NLP to computer vision, there is a growing interest in understanding the inner workings of these seemingly complex models and how their performance can be improved.

Interpreting transformers is crucial for the following reasons:

- Accountability: To successfully exploit the power of transformers in high-stakes applications like healthcare and finance, it becomes increasingly important to understand how these models are making decisions. If we don't understand how the models are arriving at certain decisions, it can be difficult to hold them accountable for their outputs.
- Transparency: Understanding how transformers work can also help to increase transparency in artificial intelligence more broadly. By making these models more interpretable, we can help to build trust and ensure that people are comfortable with the way that AI is being used in different applications.
- Fairness: One of the biggest challenges with AI is ensuring that models are fair and unbiased. Interpreting transformers can help to identify biases and other issues in these models, allowing us to address them before they cause harm.
- Performance: Finally, interpreting transformers can help us to improve the performance of these models. By understanding how they work and where they are making errors, we can refine models and make them more effective for a wide range of applications.

Overall, interpreting transformers is essential for ensuring that these models are used effectively and ethically in different applications. As transformers become more ubiquitous and more complex, it becomes increasingly important to develop new tools and techniques for interpreting them.

3 Related work

Since they were first introduced in 2017 [1], transformers have seen massive successes in a wide range of application areas. The need to understand these models and their variants has led to the development of a variety of interpretability tools and techniques. In 2019, Clark et al studied how the attention heads in BERT specialize to learn unique syntactic information [2]. For example, they show how some attention heads learn to specialize in extracting information broadly (have attention weights evenly distributed across all tokens) while some others specialize in attending to special tokens such as period "." and separator "[SEP]" tokens.

Work has also been done on the front of developing visual applications that allow the user to interactively understand how transformers learn information. InterpreT [3] stores the hidden layer outputs for a set of test examples. It then analyzes these to generate plots, including plots that depict attention weights between tokens in the final transformer layer. The VL-InterpreT [4] tool extends this idea to interpret multi-modal transformers, analyzing attention heads consisting of vision-to-vision, language-to-vision, vision-to-language and language-to-language attention components.

The idea of using spectral analysis to interpret transformers is fairly new. In 2020, Alex et al introduced the concept of applying spectral analysis to study the frequency distributions of the activation values of neurons in a transformer. In that, they apply Discrete Cosine Transform (DCT) [5] across a slice of the activation values in a given hidden layer to identify the weights of the frequency distributions that constitute the particular hidden layer. Higher frequencies imply large variation in the activation values in neurons, while lower frequencies imply gradual change in the neuron activation values. In terms of conceptual understanding, higher frequencies imply more granular information - for example, word level understanding for a task like speech level tagging (word to part-of-speech tagging), whereas lower frequencies indicate longer range information such as document level information for a task like topic classification.

The authors demonstrate how tasks such as topic classification, dialogue-speech acts classification and parts-of-speech tagging can benefit from using low-pass, band-pass and high-pass filters (*spectral filtering*). These filters essentially only allow information from certain frequency bands to pass, thus, eliminating information from frequencies that might not be helpful for a given task. For example, the task of topic classification would require longer-range features (document-level information), thus, using a low-pass filter that would only allow lower frequencies to pass would improve the transformer's performance (as shown in [6]). For speech-tagging, the authors have used high-pass filters that allows only high frequencies to pass through the network, thereby capturing more of word-level information that works well for speech-tagging. This BERT+Prism layer is shown to perform better than vanilla BERT for the appropriate tasks.

4 Approach

The main contribution of [6] is to show how linguistic information can be captured via spectral filtering and leveraged to improve the performance of models trained for certain NLP tasks. In this project, we build from the idea of how spectral analysis can be used for understanding how range-level information is captured and go from here to dive deeper into understanding the inner workings of transformers. The main contributions of our work are outlined below.

- Using NLP tasks of our own, namely, text classification and masked language modelling (MLM), we attempt to validate the hypothesis presented in [6], expecting text classification to learn more lower frequency information and MLM to learn more higher frequency information.
- Within the task of text classification, we attempt to study frequency contribution at a much finer scale (token-wise contribution) and propose a method to find out the most important words and phrases from input sequences.

Discrete Cosine Transform (DCT) [5] is the main tool on which the interpretation of transformers in this work is based. As in [6], DCT is applied to slices of word embeddings across neurons. For a given sequence of word embeddings $v_0, v_1, \cdots, v_n$ of length n, where each embedding is $d-$dimensional, we take the slice $v_0[i], v_1[i], \cdots, v_n[i]$ (across the i^{th} neuron) and find its DCT. The DCT of an $n-$dimensional sequence $[x_0, x_1, \cdots, x_n]$ is also an $n-$dimensional sequence, where the k^{th} term

is given by

$$X_k = \sum_{i=0}^{n-1} x_i cos\left[\frac{\pi}{2n}\left(2i+1\right)k\right] \quad \text{for} \quad k = 0, 1, \ldots, n-1 \tag{1}$$

The binning strategy used in [6] has been adopted for the first part of our experiments - to validate hypothesis regarding expected greater high-frequency content for MLM, and expected greater low-frequency content for text classification. The binning scheme is shown in Table 1. It maps DCT indices to token ranges and their corresponding scale in terms of document-level, paragraph-level, sentence-level, clause-level and word-level information. The scale of a particular frequency is revealed by its *period*: the number of tokens it takes to complete a full cycle. Later parts of the experiment use a finer binning strategy, defined later in the document.

Frequency bin	**DCT indices**	**Token range**	**Information scale**
Low (L)	$0-1$	$256-\infty$	Document level
Mid-Low (ML)	$2-8$	$32-256$	Paragraph level
Mid (M)	$9-33$	$8-32$	Sentence-level
Mid-High (MH)	$34-129$	$2-8$	Clause-level
High (H)	$130-511$	$1-2$	Word-level

Table 1: Binning strategy to accumulate frequency contribution at various scales

5 Experiments

This section provides details about the experiments we perform on transformers trained for the following NLP tasks - text classification and masked language modelling (MLM). The section is broadly divided based on these two tasks, with each section reporting experimental details pertaining to that NLP task.

5.1 Text Classification

The task is that of classifying reviews to the labels 0, 1, 2, 3, and 4. The labels are indicative of sentiment value, with 4 being the most positive and 0 the most negative. Though the task is that of text classification, it's learning procedure is much like a model trained for sentiment analysis, where the overall meaning of the sentence or longer-range features are expected to dominate over word-level or shorter-range features.

Data The model uses the Yelp Reviews dataset to fine-tune a pre-trained model for the classification of reviews into the labels {0,1,2,3,4}.

Model The pre-trained transformer DistilBERT is used and is fine-tuned on the Yelp Reviews dataset. The fine-tuning is done for 10 epochs and the default model hyper-parameters are used for training. The *AutoModelForSequenceClassification* provided by HuggingFace is used for loading the DistilBERT pre-trained model with its weights and adding functionality for classification. Classification is performed using the CLS token in the final layer of the transformer, which is responsible for summarizing the whole input sequence.

Evaluation method Using the spectral transform of the different layers of the trained transformer model, we first attempt to validate the following hypotheses.

- **Hypothesis 1:** Since this is essentially a task of sentiment analysis (but with multiple classes), the overall meaning of the review, how positive or negative it is should determine its class label. Thus, we should see longer sentence-level features to dominate in the trained model.
- **Hypothesis 2:** As we go down the layers of the transformer, the contribution from longer-range features should increase and the contribution from shorter-range features should decrease.

After validating or disproving the above hypotheses, we would like to go a step further to answer questions such as:

- What range of information (in terms of the order of tokens) contribute most to text classification? The answer to this would be explored using certain test inputs.
- What information is provided by the peaks in the spectrogram of the transformer? This is explored in Section 5.3.

Results A 200-worded positive review (see Appendix) is used for evaluating the model. The outputs of all the layers in the forward pass are stored. Since the maximum sequence length for the model is 512, we have 512-dimensional sequences after taking the spectral transform for each neuron. Instead of representing all 512 frequency weights, binning is used to divide the percentage compositions of frequencies into 5 bins - Low (L), Mid-Low (ML), Mid (M), Mid-High (MH), and High (H) according to the DCT indices specified in Table 1. Figure 5 in the Appendix shows the frequency distribution for the 7 hidden layers of the transformer. The following observations are made from the figure.

- Information that is most captured by every layer in the transformer is in the High frequency range, with tokens of the order of 1-2, which means that information contained by single words or a pair of adjacent words is more important than longer sentence-level information. This **disagrees with Hypothesis 1** to some extent, since the H bin is found to be contributing most to the classification task. The variation of frequency composition by layers is shown in Figure 1.
- While the percentage weight of the L bin goes up from 0.05 in the 1^{st} layer to 0.19 in the final transformer layer, the frequency contribution of the H bin decreases monotonically from 0.67 to 0.47. This observation **validates Hypothesis 2**.

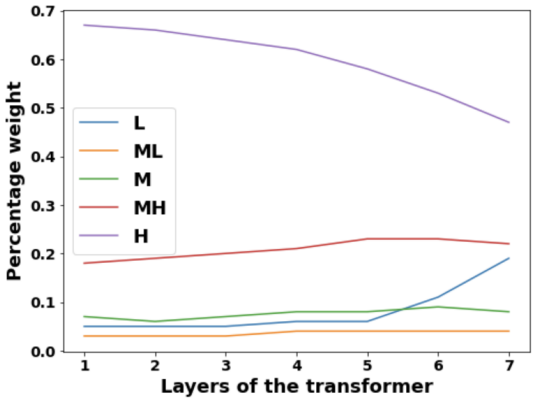
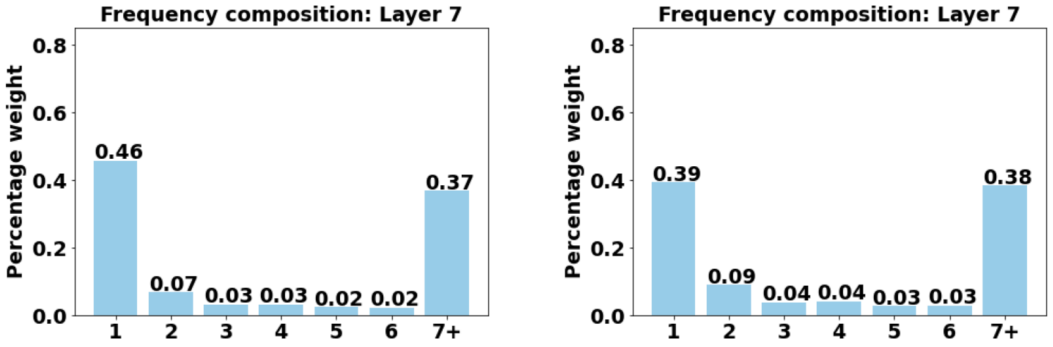


Figure 1: Variation of frequency composition with transformer layers

Frequency bin	DCT indices	Token range
1	$171 - 511$	1
2	$102 - 170$	2
3	$73 - 101$	3
4	$57 - 72$	4
5	$46 - 56$	5
6	$39 - 45$	6
7+	$0 - 38$	$7 - \infty$

Table 2: A finer binning strategy (used for input sequences of less than 7 tokens)

To what extent are frequencies important within the H-bin? Although the figures disprove the first hypothesis, they provide an interesting result. Most of the information that the transformer uses to perform sentiment-based text classification comes from the H bin, i.e., token range of 1-2. This is likely because the presence of single positive or negative words can have a huge impact on the



(a) Input: "awful service"; Predicted label: 0

(b) Input: "awfully good service"; Predicted label: 4

Figure 2: Analyzing higher frequency range

sentiment of the review. To investigate this further, we analyze the spectrum of the last layer with finer granularity for the inputs "awfully good service" and "awful service". The frequency bins are now divided according to the following time periods (tokens) - 1, 2, 3, 4, 5, 6, 7+, which would allow us to see exactly how information at the level of 1 token, 2 tokens, etc. would influence our model. The corresponding DCT index divisions are shown in Table 2. The frequency distribution of the last layer of the transformer with this binning scheme is shown in Figure 2. For the negative example, we see that information with time period $= 1$ token contributes most to classification (about 46%), which alludes to the significant contribution that the word "awful" must have had. For the example "awfully good service", which is basically the same sequence of tokens with additional tokens for "ly" and "good", we see a 7% contribution drop from the 1-token bin. From this experiment we conclude that, depending on the input sequence, the transformer model has the ability to pick out varied single-token or multiple-token information. Also, the effect of contrasting words like "good" and "bad" result in the transformer learning less from the dominant, 1-token bin, and learning more from combinations of words or tokens.

The overall effect of positive and negative words is so strong for the case of text classification that even a neutral to slightly positive review like "I was feeling terrible about the constant dripping of my kitchen tap, when I finally decided to call the plumber. He fixed the tap in no time." is predicted as 0 by the model due to the presence of the word "terrible". When this word is replaced by a meaningless token such as "xyz", the predicted label is 1.

5.2 Masked Language Modelling

Data For fine-tuning the model for MLM, the *wikitext-2-raw-v1* subset of wikitext is used. The tokens in the input sequence are randomly masked and we use the original input (without masking) to calculate the error while training.

Model The DistilBERT pre-trained transformer is fine-tuned on the wikitext for 3 epochs and default model hyper-parameters are used for training. We use the *AutoModelForMaskedLM* model provided by HuggingFace. For tokenizing, we use *AutoTokenizer* from HuggingFace.

Evaluation method A wikipedia page is chosen and around 512 consecutive words are randomly picked up from it. This input sequence is then masked at random and tested on the fine-tuned MLM model. Similar to the task of text classification, the hidden layers are stored in the forward pass during evaluation of the input sequence and DCT is applied on these activations. We attempt to validate and/or disprove the following hypotheses for MLM.

- **Hypothesis 1**: Since in masked language modelling, a word depends more on the words around it and lesser on words farther away, we should expect to see more concentration on higher frequencies.
- **Hypothesis 2**: Compared to text classification, MLM should be more locally dependent, i.e should give more weight to higher frequencies from the DCT transform.

Results The plots with frequency composition of all frequency bins (binned in accordance with Table 1) as shown in the Appendix in Figure 6. The variation of frequency composition across the layers of the transformer is shown in Figure 3. For MLM, there is greater weight on the high frequency bins, especially the H-bin, which implies that high frequencies, in particular, information content from 1-2 token ranges are dominant. This implies high composition of word-level information, which is expected since to predict masked words most likely depend on the words immediately around them. This **validates Hypothesis 1**. On comparing with text classification, we see that MLM places more weight on higher frequencies which is expected as text classification requires broader context for sentiment analysis. This **validates Hypothesis 2**.

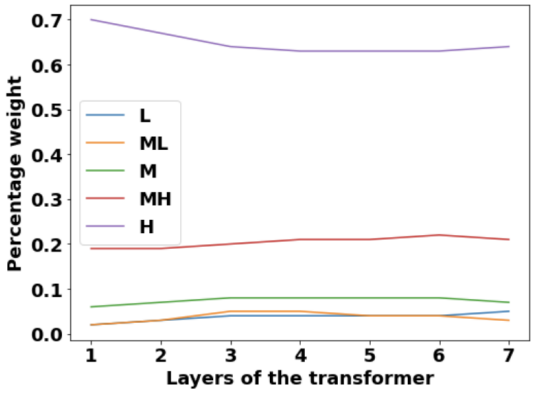


Figure 3: Variation of frequency composition with transformer layers

5.3 Mapping spectrum peaks to word importance

This section analyzes how peaks in the spectrum can give meaningful insights into the words and phrases that are most important for classification. We first see how frequency can be mapped to token ranges. This is a tricky task because the function that relates the two is not linear. For an input sequence of length N, the time period T is related to the DCT index k in the following way.

$$T = \frac{N}{2k} \tag{2}$$

For the particular case of text classification considered before, the input sequence length is set at $N = 512$. Thus, the frequency component k in the DCT spectrum corresponds to a time period of $\frac{256}{k}$ tokens. The variation of time period with DCT index for this task is shown in Figure 4. The mapping of lower DCT indices like $0 - 1$ result in token ranges of $256 - \infty$, which is not much useful in learning one-to-one mappings. The area of the curve that can be exploited to learn such one-to-one mappings is the approximately linear portion between DCT indices 2 and 10 and all the indices beyond 10, i.e, $k \geq 10$ (since their corresponding time periods would be $T \leq 25$, which would make it easy to map DCT indices to tokens exactly).

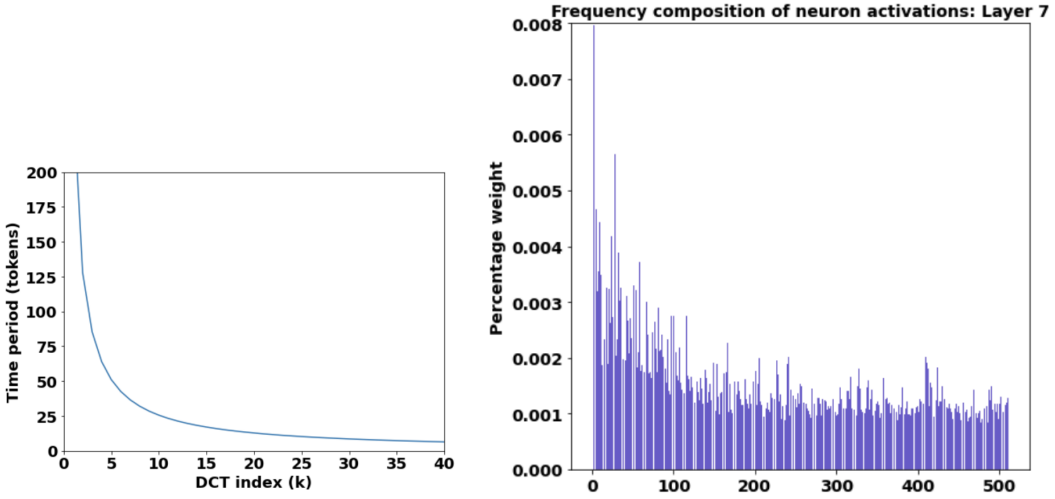
In mapping spectral peaks to tokens (rather than token ranges), we exploit the fact that the last layer in the model uses the CLS token for the final classification, so token ranges are directly mapped from token range (time period) to their index in the input sequence.

We now analyze the frequency spectrum (shown in Figure 4) for text classification when evaluated on the 200-worded example given in the Appendix. The results for a few peaks in the $1 \leq T \leq 100$ region are presented below.

- **Peak at $k = 28$:** The corresponding time period is $T = 8.827$ (since k is 0-indexed, $T = 256/(28 + 1)$). This corresponds to the tokens at indices 8 and 9, which represent "exceptional room". The next token in the sequence is "service". Thus, we see how these tokens "exceptional room service" caused a peak in the frequency spectrum, since they must have been strongly contributing to classification in the positive sentiment direction.
- **Peak at $k = 18$:** The corresponding time period is $T = 13.47$ ($T = 256/(18 + 1)$). This corresponds to the tokens at indices 13 and 14, which represent "look no". The next token

in the sequence is "further". Thus, the phrase "look no further" is found to be contributing strongly to classification.

The reason this finding is so interesting is because it provides for a way to understand the importance of phrases and words in the input sequence. Perhaps it is this correlation of frequency spectrum to token importance that has resulted in the successes of spectral-based attention methods in the literature.



(a) Variation of time period (tokens) with DCT index

(b) Frequency spectrum of the final layer: Text classification

Figure 4: Mapping spectral peaks to word importance

6 Conclusion and Future work

In this work, we explored how the frequency transform of neuron activations can be used to understand the inner workings of transformers. The Discrete Cosine Transform is used to validate some hypotheses about transformers trained for the tasks of Text Classification and Masked Language Modelling. Transformers fine-tuned for the task of Text Classification are expected to show a higher concentration of low-frequency information, i.e., longer-range information such as sentence and paragraph-level. On the other hand, Masked Language Modelling is expected to show higher concentration of the high frequency bin, corresponding to shorter, word-level information. While the hypothesis for text classification was rejected, it led to an interesting finding. Performing a finer spectral analysis revealed most of the classification weight coming from the DCT range corresponding to a time period of 1 (token range of 1). This indicated the strong dependency of the presence of standalone positive and/or negative words, and a weaker dependency on combinations of words, for Text Classification.

The next part of the work focused on mapping peaks in the frequency spectrum to important words and phrases. It was shown that a one-on-one mapping from the frequency spectrum to token indices in the input sequence was possible for a certain portion of the Token range vs. DCT index plot. A couple of examples were then presented which showed how peaks in the spectrum corresponded to phrases such as "exceptional room service" and "look no further", indicating the importance of these input tokens for classification.

An extension to the present work would be to develop an interactive tool like InterpreT [3] that would allow the user to play with custom input sequences and find the corresponding frequency distributions. From those spectral peaks, it could then assign scores to tokens in the input sequence, indicating their importance for the text classification task.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[3] Vasudev Lal, Arden Ma, Estelle Aflalo, Phillip Howard, Ana Simoes, Daniel Korat, Oren Pereg, Gadi Singer, and Moshe Wasserblat. Interpret: An interactive visualization tool for interpreting transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 135–142, 2021.

[4] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21406–21415, 2022.

[5] K. R. Rao, Patrick Yip, and Vladimir Britanak. *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic Press, Inc., USA, 2007.

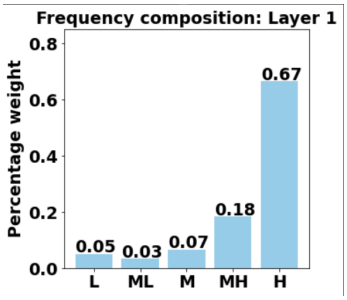
[6] Alex Tamkin, Dan Jurafsky, and Noah D. Goodman. Language through a prism: A spectral approach for multiscale language representations. *CoRR*, abs/2011.04823, 2020.

7 Appendix

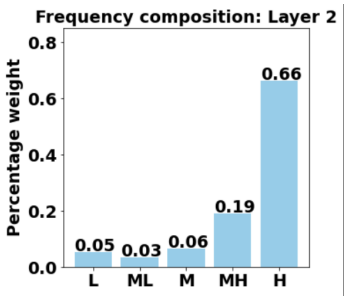
7.1 Text Classification

Test input sequence (200 words) If you're looking for an exceptional room service experience, look no further than this hotel! I recently had the pleasure of staying here and was blown away by the level of service provided by their room service team. First of all, the menu selection was fantastic. There were plenty of options for breakfast, lunch, and dinner, as well as a great selection of snacks and desserts. The food itself was delicious and prepared to perfection. The presentation of each dish was beautiful, and it was clear that great care was taken in every aspect of the dining experience. What really stood out to me, though, was the level of attention and care given by the room service staff. They were friendly, professional, and went above and beyond to ensure that every request was fulfilled. Even when I had a last-minute request, they were quick to accommodate and make sure that I had everything I needed to enjoy my meal. Overall, I can't recommend this hotel's room service enough. If you're looking for a truly exceptional dining experience from the comfort of your room, this is the place to be. I'll definitely be staying here again on my next trip to the area!

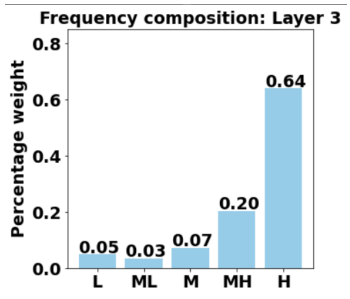
Frequency distributions of neurons for different layers of the network:



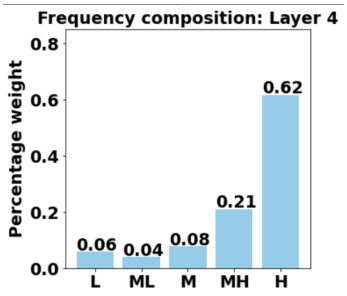
(a) Layer 1



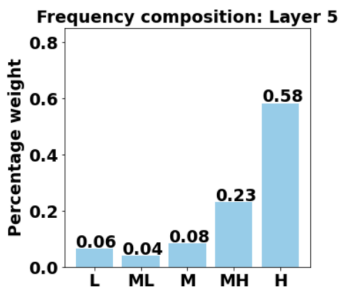
(b) Layer 2



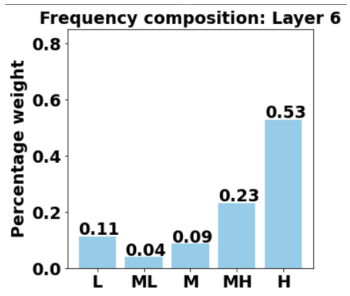
(c) Layer 3



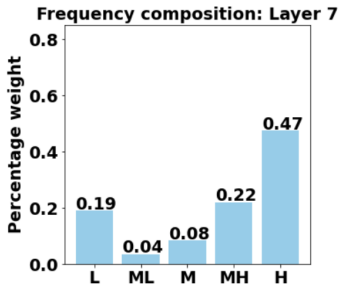
(d) Layer 4



(e) Layer 5



(f) Layer 6

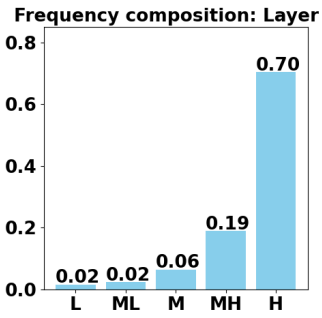


(g) Layer 7

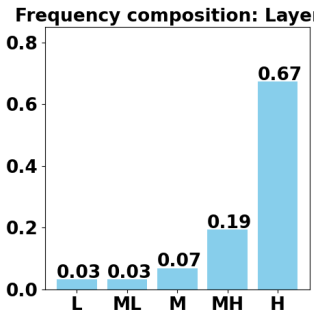
Figure 5: Text classification

7.2 Masked Language Modelling

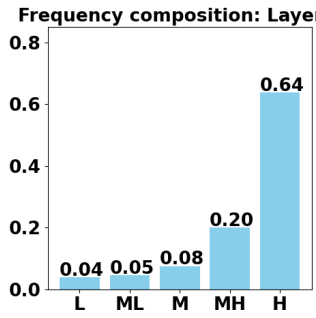
Frequency distributions of neurons for different layers of the network:



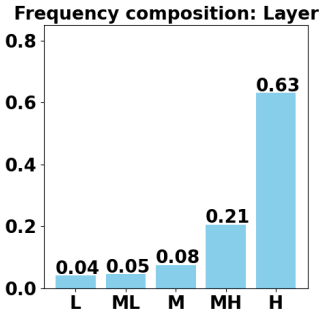
(a) Layer 1



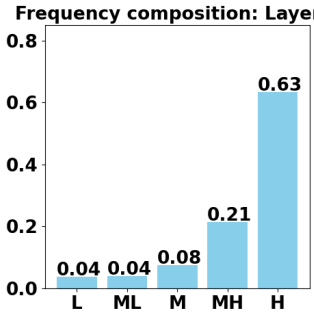
(b) Layer 2



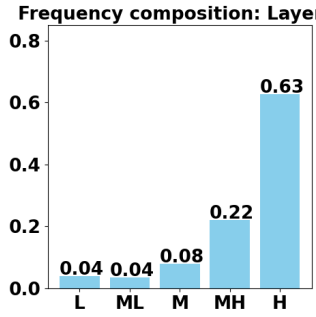
(c) Layer 3



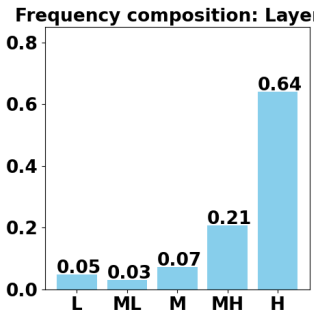
(d) Layer 4



(e) Layer 5



(f) Layer 6



(g) Layer 7

Figure 6: Masked Language Modelling