SleepTalk: Textual DeepDream for NLP Model Interpretability

Stanford CS224N Custom Project
Staff Mentor: Angelica Sun

David Yue
Department of Computer Science
Stanford University
davidyue@stanford.edu

Finsam Samson
Department of Computer Science
Stanford University
finsam@stanford.edu

Abstract

We propose SleepTalk, a technique for improving interpretability of pre-trained NLP models. Greater intepretability of black box neural networks is imperative for mitigating bias, developing trust in implemented models, and advancing intuition for better transfer learning. Thus, SleepTalk provides an approach to gain human-interpretable data on the learned representations of neurons in large neural networks. We also augment SleepTalk to be used on the adjacent task of unsupervised textual style transfer; synthesizing output text from just a content reference input and a style reference input. We assess the interpretations of SleepTalk and its behavior at different layers and qualify its resulting outputs of pre-trained NLP models. Finally, via results from SleepTalk, we suggest similarities betwen NLP neural network models and layers of the human brain's temporal and parietal lobe, structures critical to the formation of thoughts.

1 Introduction

Neural networks have become ubiquitous in applications of NLP, but they are often characterized as black-boxes due to lack of understanding in their decision-making under the hood. As the applications of deep learning become more widespread in everyday life, understanding of these models becomes crucial for practitioners, researchers, policymakers, and all stakeholders – especially with respect to the domains of bias, ethics, and human-computer interaction. We want the ability to interpret a neural network at the fundamental level; with interpretability, we can mitigate bias, develop trust in applied models, and advance our intuition on better transfer learning from prior networks.

Lack of interpretability in a neural network stems in part from its structure. Because neural models are composed by stacking or layering large quantities of individual nodes, or neurons, we lack a centralized decision-making unit. Instead, each node might be performing some high-level information extraction or synthesis, and only working together are they able to produce meaningful output. To gain insight on the internal functioning of neural networks, we desired to study the behavior of individual neurons in the network.

A technique to understanding specific neurons is to examine the learned representation of a node. Specifically, we can deduce what kind of input maximally activates, or excites, a neuron in the network [1]. In this way, by producing the input that activates the node the most, we acquire intuition on the decision-making of that specific neuron; for example, in a classification network, the optimized input will tell us what feature that node is looking for when outputting a value.

Recent work in computer vision has applied DeepDream techniques utilizing activation maximization (AM) methods to provide better intuitive visualization as well as understanding of the inner workings of intermediate layers and neurons in deep neural networks [2]. These techniques have facilitated feature visualizations to help better visually interpret what individual classifier neurons in CV models

are doing [3]. However, since DeepDream can not be immediately applied to NLP tasks, we propose a new technique, SleepTalk, to improve interpretability of NLP models.

Our SleepTalk approach is applied to existing large pre-trained models to yield better intuition on their decision-making. We then augment the SleepTalk technique and apply it to the adjacent (sometimes considered downstream) task of textual style transfer. We evaluate the results of the SleepTalk method quantitatively and qualitatively.

2 Related Work

Different methods for assessing interpretability of neural networks have been attempted with varying success. Some have studied implicit interpretability by examining the workings of intermediate modular units, while others have studied the neural model decision-making as an end-to-end process. We use these approaches as baselines or inspiration for our technique's design and study.

In deep learning models for computer vision applications, one approach of visual interpretation is examining activation maximization in classifier convolutional neural networks [4] [5]. In these instances, a specific model neuron is selected, and all input data is fed into the model individually. The input data that maximally activates the selected neuron would be considered the nearest neighbor of the learned representation of the convolutional neural network [6].

While this method produces an optimized input that is human-understandable and typically captured from a real-world distribution, the search space of this approach is limited to the test feature set. As the test feature set is only a small subset of the the complete input feature space, this constraint prevents holistic understanding of the model's behavior on data that is out-of-distribution of the test set, such as adversarial attack inputs [7].

Thus, back-propagated AM techniques in computer vision have attempted to iteratively optimize the input rather than perform a nearest neighbor search over input data [4] [8]. To this end, a trained model's weights are first frozen; then, a randomly sampled input is fed through the model. The input is then iteratively optimized via back-propagation in order to maximize the selected neuron's activation. This expands the input's domain to the entire feature space, instead of being limited to input data.

Both the nearest neighbor AM and the iterative back-propagated AM have yielded successful visualizations of learned representations in computer vision models. However, AM applied to NLP tasks faces many challenges, not limited to lack of continuous inputs and NLP's use of non-stationary data (as opposed to computer vision input images which are usually not auto-regressive).

The DeepDream technique for images has functioned as an extension of traditional AM approaches, focusing on not only better interpreting existing models but leveraging this understanding for generative tasks [9]. DeepDream applies iterative back-propagated AM to trained convolutional neural network classifiers, and selects a class neuron in the last softmax layer (or another layer based on output target) [10] [11]. This way, DeepDream can optimize an input image to adopt qualities of a specific class.

In the separate realm of textual style transfer, most work has performed transfer in the domain of a specific textual attribute as a supervised task [12] [13]. In prior work, one attribute of text is selected (ex. formality, persuasiveness, empathy) and the content text is then translated from unstyled to styled; in this approach, style transfer becomes an NMT task, ie. given an input string, output the string with the attribute imbued. However, this is an unscalable solution due to the requirement of large amounts of parallel corpus, which typically does not exist as there are no prior installments of styled and unstyled text with the same content.

3 Approach

3.1 Methods

To improve interpretability, we draw inspiration from both DeepDream and AM and extend this idea to NLP models. However, DeepDream can not be immediately applied to NLP models, due to three critical nuances between visual data and textual data. Notably, first, visual images exist in continuous data manifolds, whereas words or subwords, are discretized tokens. Second, textual data

interpretability is hinged on completely preserving semantic meaning as well as fluency, whereas visual representations are more tolerant to disruptions in structure. Finally, NLP models usually have sequential inductive biases whereas CV models are inductively biased toward locality; thus the auto-regressive nature of NLP data requires preservation of semantics at every timestep.

We propose SleepTalk, a new technique to provide feature-level understandings of NLP models. We demonstrate the technique on existing large pre-trained NLP models suited for a variety of tasks, including next sentence prediction, sentence classification, and masked language modeling, to provide general intuitions on the decision-making of the models behind-the-scenes. We then augment SleepTalk to a modified technique as the basis for an adjacent NLP task, textual style transfer. Our proposed method applies unsupervised style transfer with only a necessary content reference input and a style reference input, as opposed to traditional NMT which requires large quantities of parallel corpus data for training.

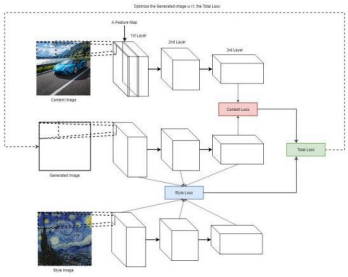

Figure 1: In computer vision style transfer, a style reference is fed through a neural network and the intermediate layer's information is captured. This style information is then used for back propagating the content reference image's inputs until it reaches the desired style. This can not be directly used on text because text input is not continuous. Thus, we must capture the style metrics of the text in an encoding of intermediate layers. Figure from Ganegedara et al.

3.2 Baselines

Direct study of naive AM or DeepDream on pure NLP models remain largely unstudied. Additionally, the interpretability of both models and the techniques to understand them are difficult to quantify. Thus, there does not exist a widely adopted baseline. Therefore we set the baseline of our technique in the adjacent task of textual style transfer. We use Rao et al.'s 2018 GYAFC dataset and NMT style transfer method as our baseline [14]. In addition, we quantitatively and qualitatively assess the interpretability performance of SleepTalk on large pre-trained models.

3.3 SleepTalk

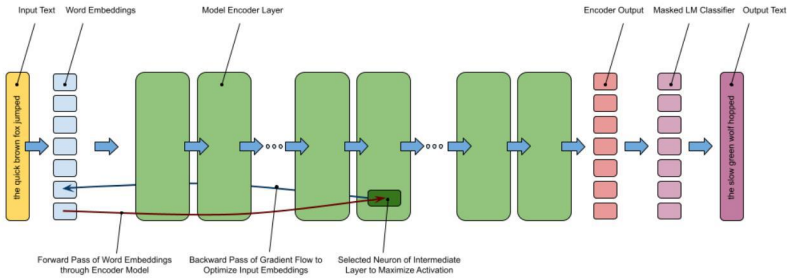

Figure 2: Our SleepTalk model selects a neuron in an intermediate layer to activate. The weights of the model are fixed, and then the input embeddings are optimized via gradient ascent in order to increase that neuron's activation. After optimization is complete, the input word embeddings are then fed through the model to the BERT MLM classifier, where the words are then converted to an output text.

Our SleepTalk approach is applied to fully-trained or pre-trained language models. In naive AM approaches, we choose a neuron η from the model and forward-feed an input ξ. Let the activation of η be $h^{\eta}(\xi)$. Thus, in naive AM, we optimize input ξ such that

$$\xi^{*} = argmax_{\xi} h^{\eta}(\xi)$$

which is carried out by gradient descent with update rule

$$x := x + \frac{\partial h^{\eta}(\xi)}{\partial \xi}$$

To overcome the discrete token input problem, we first transform text to word embeddings, and then apply AM on the embeddings, which exist on a continuous manifold. However, in order to transform word embeddings back to tokens, it does not immediately make sense to find nearest neighbors, as oftentimes optimized word embeddings may not be close to actual tokens–in fact, quite far from embeddings of words in the lexicon. Thus, we then feed our optimized word embeddings through a Masked LM model (in our case pre-trained BERT), but with a classification requirement on every word. The reason for this design is because the Masked LM model for BERT is trained to closely reconstruct it's original word embedding input as output.

3.4 Style Transfer via SleepTalk

We augment SleepTalk to perform unsupervised style transfer. That is, rather than training an NMT model on parallel data and transferring a specific style attribute, we implicity encode the style in a style reference text. With both a style reference text and a content reference text as input, we can optimize the inputs such that the intermediate layers of the model are "close" to the intermediate layers when style reference and content reference are used as input.

We capture intermediate layer feature maps and optimize the input on a multi-part loss function that penalizes the input from straying from the input content and rewards it for adopting intermediate feature maps similar to the style reference's layers. In other words, for a layer l let q be a neuron of that layer. We have h_q^l for the activation of that neuron, which we calculate for the input g and the content reference c. We also have G_q^l, the Gram matrix of that neuron's values, which we calculate for the input and the style reference s.

$$L = \frac{1}{2} \sum_{q} (\psi_{content}(h_q^l(g) - h_q^l(c))^2 + \psi_{style}(G_q^l(g) - G_q^l(s))^2)$$

We utilize ψ_{style} and $\psi_{content}$ as hyperparameters to weight the content and style preservation. Our loss function implements the Gram matrix because it acts as a "weaker" constraint compared to absolute disparity (which content loss is measured by). This facilitates a positive transfer of style implicitly encoded into the style reference, while forcing content preservation. The ψ_{style} and $\psi_{content}$ hyperparameters further allow control of style and content adoption with granularity.

Code for our SleepTalk approach, unsupervised style transfer, and project scaffolding is written by us.

4 Experiments

4.1 Data

We studied SleepTalk on text data from a variety of domains. For interpretability studies, we used select sentences from the Wikipedia Corpus and the Book Corpus [15] [16]. For textual style transfer, we used Rao et al.'s 2018 GYAFC dataset [14]. This dataset has parallel text for the informal to task. However, in our unsupervised style transfer task we require the use of both a style reference text and a content reference text. We utilize content and style reference texts from book data we selected–for most of our unsupervised explorations, we used style references from Shakespeare's A Midsummer Night's Dream and content references from JK Rowling's Prisoner of Azkaban; we also apply unsupervised style transfer to other data from the Book Corpus.

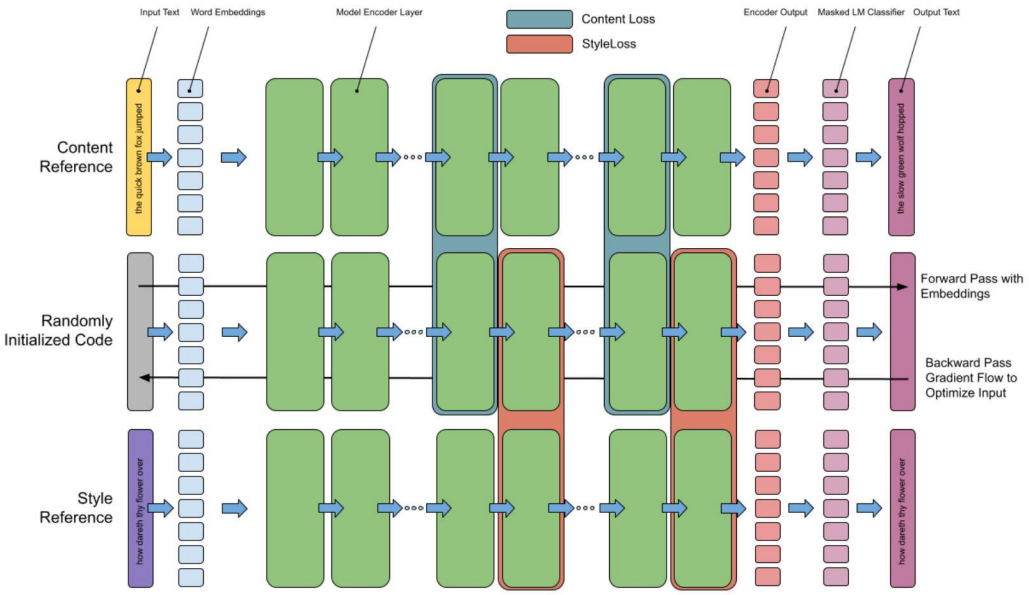

Figure 3: The augmented SleepTalk for unsupervised textual style transfer requires a content reference and a style reference. The content and style are implicitly encoded in the intermediate layers when they are forward fed through the model. We begin with a randomly initialized code and optimize it so that it's intermediate layers in the model are similar to the content and style references' intermediate layers (we use the Gram matrix for the style loss of the layers). In practice, we selected layers 8 to 12 for both style and content loss.

4.2 Evaluation method

We first studied the results of SleepTalk as a means for interpretability. We manually assessed the fluency of SleepTalk at different iterations to examine the preservation of lucidity as SleepTalk progresses. This way, we can understand if the "subconscious" learned representations of the model have developed intuition on syntactic structure. We also studied the frequency of SleepTalk's dreamed words to see if the model had greater affinities to specific kinds of tokens. With this evaluation, we assessed the pre-trained model's decision-making process and its relationship with indicator words in the lexicon.

In our further exploration, we measured contextual drift by analyzing the Euclidean distance of words after iterations of SleepTalk. In this scenario, we examined the holistic movement of embeddings as the model progresses from it's original input text. This provides insight on directional patterns in the SleepTalk approach. Qualitatively, we visualized the embedding at particular iterations via t-distributed stochastic neighbor embedding (t-sne).

We additionally studied the role of different intra-layer neurons and content drift. We measured the BLEU score between the SleepTalk output with the original input text to understand the change in semantic content. We then compared the BLEU score dropoff curve for nodes within the same layer. We also compare BLEU score dropoff curves for nodes in different layers. Because transformer layer have multiple heads of attention that are interchangeable without loss of generality (prior to pretraining), we hypothesized that neurons in the same layer would exhibit similar dropoff curves whereas different layers would produce different dropoff curves.

4.3 Experimental details

For our interpretability studies, we performed SleepTalk on the Bidirectional Encoder Representations from Transformers (BERT) model, specifically the BERT-Base-Uncased variant of the model. The pretrained BERT-Base-Uncased model is composed of 12 layers of encoders stacked together, with each layer comprising 768 hidden units and 12 attention heads. BERT-Base-Uncased contains

approximately 110M parameters. For each input, we ran our SleepTalk system for 400 iterations with unit learning rate and no learning rate decay. On a 64-token input string, the process took approximately 20 minutes per input.

In our unsupervised style transfer task, we used the BERT-Base-Uncased model, selecting layers 8 to 12 as the loss layers for both style and content. We made this design choice as earlier layers in the model are typically capturing broader features and we did not want these high-level attributes to overpower the human-interpretable reference attributes. We set ψ_{style} and $\psi_{content}$ both as 1 to have equal weighting of style and content. We ran the augmented SleepTalk for style transfer for 200 iterations per input. On a 64-token input string, the process took approximately 20 minutes per input.

4.4 Results

In our initial interpretability studies, we performed SleepTalk to understand learned representations of the pre-trained models. For one example, we studied layer 12 at the activation of the first node in the first attention head, and we began with an input text from Harry Potter and the Sorcerer's Stone. As we performed SleepTalk, we noticed that the optimized input slowly changed as the node activation was maximized. The quality of the text, ie. the fluency and syntactic meaning of the text, decreased as SleepTalk was performed. Although the text did not retain English fluency, it still provided intuition on the behaviors of individual neurons. The interesting results of the interpretability study demonstrated that SleepTalk provided a viable avenue to further understand the behavior of neural networks.

Iteration	SleepTalk Optimized String	BLEU Score
Original	it was the unicorn all right and it was dead harry had never seen anything so beautiful and sad its long slender legs were stuck out at odd angles where it had fallen	1.0
20	. . was the unicorn all right and it was dead harry had never seen anything so beautiful and sad its long slender legs were stuck out at odd angles where it had fallen .	0.9103429040065263
40	. it was the god all right and it was dead harryd never seen anything so beautiful and sad its long black legs were stuck out at the angles where it had . .	0.5884678099624038
60	. lords the almighty all right and it was dead harryd never seen anything sogn and sad its long slender legs were stuck out at odd angles where it had . .	0.659958302850984
80	. how was the almighty all right and it was dead hed never seen anything so awful and sad its long slender legs were stuck out at odd angles where it had . .	0.6225705543415939
100	. goddamns the almighty all right and it was dead she had never seen anything so damn and by its long slender legs were stuck out at odd angles where it had . .	0.5852432074546792
120	. . holy the almighty as right . it was dead he had never seen anything so damn and . its long slender legs were stuck out at odd angles and it had and .	0.43273125052358746
140	. hell holy lord almighty as tongue , it was dead he had never seen anything so damn and . its long slender legs were stuck out at odd angles and it had and .	0.4234197579236933
160	the . . godsmenheart and god the mad he had never seen . so damn and . its long his legs and stuck out at odd places where it had a .	0.14868720326332424
180	. how holy lordsef tongue ,h the damned kate had never seen . so damn and . s long and legs and , out at odd spot , it had a .	0.0
200	.all lord and . twomac andhard asbas pd not . , so damn and a he , he , tip , a , and something and it , and .	0.0

Figure 4: Performing SleepTalk on BERT-Base-Uncased for a neuron in layer 8. We see qualitatively that the optimized text string has gradual contextual drift. Part of human-interpretability shows that when this neuron is activated at high degrees, words relating to the Bible, or some kind of religious text, begin to appear. This helps us understand the learned representation of that neuron in the model. We also see that the contextual drift causes the BLEU score to decrease over iteration.

We then studied the effect of the intermediate layer being maximized, as well as the specific intra-layer neuron selected. We measured content preservation and drift by examining the BLEU score of the SleepTalk output with the initial input text. We discovered that for neurons in the same layer, the BLEU dropoff when performing SleepTalk exhibited similar curves, and that there was no clear patter amongst neurons. This is consistent with our hypothesis, as intuitively, the different nodes in the transformed architecture in the same layer see the exact same prior, and thus before training are indistinguishable without loss of generality.

As we had initially expected, the BLEU score for different layers demonstrated different dropoff curves. Crucially, over different input texts from many domains, we noticed a common pattern:

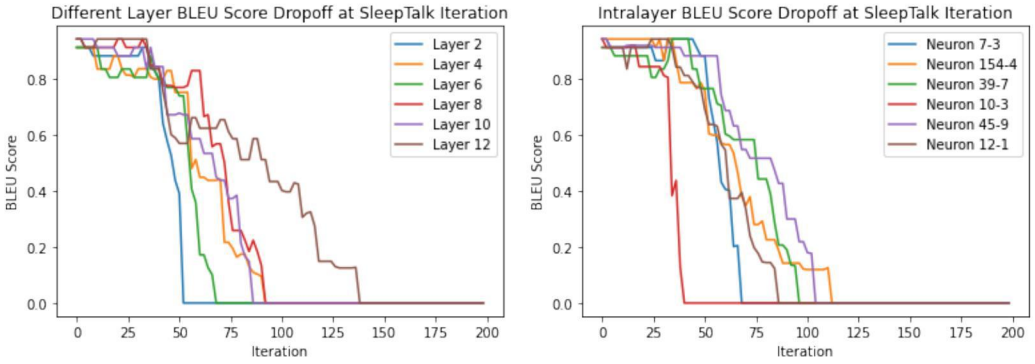

Figure 5: When studying content drift via BLEU score of outputs compared to input as reference, we saw that individual neurons in the same layer did not exhibit a consistent patter. However, when comparing neurons across layers, we found that upstream layers typically had faster dropoffs. We believe that this is because upstream layers learn higher level features, and so changes in earlier layer activations create larger tolerances for content constraints.

performing SleepTalk on earlier layer (ie. upstream in the model) caused BLEU score dropoff to happen quickly. For example, layer 2 typically saw BLEU score drop to 0 around iteration 52. Comparatively, layer 12, had a much longer tail in BLEU dropoff, reaching 0 around iteration 140. The difference in dropoff curves was most obvious for extreme layers, such as layer 1 and 2 or 11 and 12.

We performed t-distributed stochastic neighbor embedding to study the directional movement of the subword vectors. We found that as SleepTalk iterations increased word vectors spread out more; we have two orthogonal hypotheses for this behavior: 1) every iteration of SleepTalk caused drift by injecting entropy into the output, and 2) SleepTalk forced different input tokens to move in different directions since the learned representation of the neuron is multifaceted and constrained to a singular high-level function.

In our unsupervised text style transfer, our results demonstrated that pure unsupervised transfer has difficulty producing human-understandable results. As opposed to visual style transfer which can tolerate changes in content, textual style transfer requires complete preservation of the content reference. Our explorations demonstrated that any change in adopting features from the style reference would immediately drop similarity with the content reference. To over come this problem, we increased the weight of the content loss; however, when the weight was large enough to prevent loss in content, the style loss was not significant enough to transfer style attributes.

5 Analysis

SleepTalk's approach as a means for interpretability studies demonstrated some interesting results. One is the gradual change in semantic meaning as SleepTalk progressed. For example, "odd angles" slowly changed to "odd places" and then "odd spot." The change in content is a drift that slowly deviates from the original text. Similar to human thought processes, words jump to neighboring tokens of close semantic meaning.

As evidenced in Table **??**, another case of this kind of drift is the word "unicorn" becoming "god," perhaps because they have similar embeddings (both describing supernatural beings). The word "god" then became "almighty", and then "lord." Also interestingly, not only were individual tokens changed, but groups of words also drifted. For example, "god" led to the string "lords the almighty" to appear in the output, and then which later became "holy the almighty." In some instances, the semantic meaning was changed quite quickly over iterations. For example, the word "beautiful" changed to "awful." In other cases, the embedding drift causes changes in the words that were not human-interpretable.

In terms of model interpretability, SleepTalk provided intuitions on the behavior of specific nodes in the the model. For example, neuron 1-1 of layer 12 drifted to words relating to religion, exhibiting

words like "god," "almighty," "lord," "holy," and "hell." Because these dreamlike changes occur as the activation of the neuron is slowly maximized, we can hypothesize that the specific neuron has learned representations dealing with words that are divine or supernatural. Furthermore we can connect this with the pretraining process–possibly because the Bible or other religious text were in the pre-training dataset and the model developed those representations through the pre-training process. Another example is neuron 1-1 of layer 8. The output drifted to include words like "song," "music," and "sing." In this case, we can hypothesize that the neuron had a learned representation that dealt with words relating to music.

Measuring the BLEU score dropoff, we can see that upstream layers have quicker drops. We hypothesize that this is because upstream layers are learning broader level features, and thus changes in upstream layer activations create larger changes for content constraints. In this way, the model's output can have greater variability with fewer iterations.

For our unsupervised textual style transfer task, SleepTalk's outputs lacked human interpretability. We belive this is because text input requires perfect preservation of content and semantic meaning. However, we quickly realized that in our unsupervised approach, preserving content in the intermediate layer activations was mostly mutually exclusive with attaining similarities in style layer activations. Thus, any small attempt to inject style attributes immediately disrupted the content and semantic meaning. The added challenge is that text data exists as discrete tokens (unlike visual data which resides on continuous data manifolds), and thus there is a barrier to shifting the data to adopt a style on a discretized landscape.

In visual DeepDream, created images were noted to resemble imagery similar to psilocybin- and LSD-induced hallucinations [17]. In this way, it was believed to point to a functional similarity between convolutional neural networks and specific physiological layers of the visual cortex. In the same way, we see similarities between the outputs of SleepTalk and sentences that might be formed when one is under the influence, impaired, hallucinating, or sleeping. Thus, we believe this suggests a functional resemblence between NLP neural network models and layers of the human's temporal and parietal lobe, which are critical to forming thoughts.

6 Conclusion

We developed SleepTalk, a technique to better understand the decision-making of black box NLP models. Our SleepTalk approach allowed for illustrating learned representations of specific neurons in large pre-trained networks. By understanding the learned representations of individual nodes, we gain insight on the overall behavior of the model's decision process. We studied the content drift and directional movement of the SleepTalk technique. In addition, we augmented SleepTalk as a method of unsupervised text style transfer. We demonstrated that the SleepTalk technique provided general intuitions on specific nodes, layers, and models, and helped improved interpretability of neural networks by providing human-understandable words that capture a cross-section of the learned representation.

This work provides a solid first step in assessing model interpretability in the NLP domain, as well as attempting textual style transfer in an unsupervised way. A current limitation of the work is that generated outputs did not retain human fluency, which may be necessary in the future for perfectly interpretable data. We believe that future avenues of work that should be studied are preserving fluency in the process, using SleepTalk to transfer text style without loss of content, and applying SleepTalk on other model architectures.

SleepTalk provides a pivotal tool for practitioners, researchers, and policymakers to better interpret NLP models, thus providing insight on model behavior. In turn this facilitates better mitigation of bias, fosters trust in deployed models, and extends our intuition on improved transfer learning. In addition, the results of SleepTalk suggest functional similarities between neural networks used in the NLP domain and layers of the parietal and temporal lobe, parts of the human brain that are imperative to thought formation. The SleepTalk approach will improve human understanding of artificial neural networks and help researchers safely understand and unlock the power of NLP models.

References

[1] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016.

[2] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them, 2014.

[3] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015.

[4] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks, 2016.

[5] Jianyu Wang, Zhishuai Zhang, Cihang Xie, Vittal Premachandran, and Alan Yuille. Unsupervised learning of object semantic parts from internal states of cnns by population encoding. *arXiv preprint arXiv:1511.06855*, 2015.

[6] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[7] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016.

[8] Zhiming Zhou, Han Cai, Shu Rong, Yuxuan Song, Kan Ren, Weinan Zhang, Yong Yu, and Jun Wang. Activation maximization generative adversarial nets, 2018.

[9] Emily L. Spratt. Dream formulations and deep neural networks: Humanistic themes in the iconology of the machine-learned image. *ArXiv*, abs/1802.01274, 2018.

[10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015.

[11] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models, 2021.

[12] Alexey Tikhonov, Viacheslav Shibaev, Aleksander Nagaev, Aigul Nugmanova, and Ivan P. Yamshchikov. Style transfer for texts: Retrain, report errors, compare with rewrites, 2019.

[13] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: A simple approach to sentiment and style transfer, 2018.

[14] Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. 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 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[15] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[16] Ludovic Denoyer and Patrick Gallinari. The wikipedia xml corpus. *SIGIR Forum*, 40, 06 2006.

[17] Will Xiao and Gabriel Kreiman. Xdream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLOS Computational Biology*, 16(6):e1007973, Jun 2020.

A Appendix (optional)

The authors would like to thank their TA mentor, Angelica Sun for all of the guidance and support. The authors would like to thank the CS224n course staff for instruction.

In our t-sne analysis, we found that the directional movement of token embeddings caused them to become more spread out as iterations increased. We had two orthogonal hypotheses for this observed phenomenon. First, was that each iteration of SleepTalk added entropy to the output (ie. breaking initial patterns of the input since the optimization goal is an acivation). Second, we believe that since the learned representation of each neuron is multifaceted, then different tokens moved in different directions in the high-dimensional space.

Below, our t-distributed stochastic neighbor embedding shows the directional movement of the word vectors as the SleepTalk iterations are performed. We iterations of SleepTalk injecting entropy to the output and moving the tokens in different directions in the high-dimensional space.

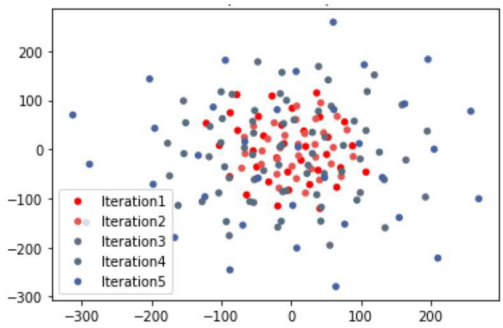

Figure 6: The t-sne visualization of input word embeddings changing over iterations of SleepTalk. We see that as iterations progress, the word vectors deviate further away from the origin and further away from each other (in the projected low-dimensional space)