Controllable Style Generation Through Activation Adjustments

Gregory Hunter
Columbia University
geh2129@columbia.edu

Abstract

Controllable text generation has been a problem for Large Language Models (LLMs) since their inception. As they have become more impressive in their ability to generate human like text, the need to control this generation has become an issue of great importance because controlling style of writing is an integral part of human level text generation. While state-of-the-art LLMS can create sentences, paragraphs, or more of generated text that holds some consistency of subject from the prompt, they cannot control for style consistently. I designed a new model, the Style Adjustment Transformer (SAT)[1], designed to control the style of the generated text without prompt design, causal response optimization Frans (2021), or beam search. The SAT model adjusts the activations of the a pre-trained LLM (in this case GPT-2 (Radford et al., 2019)) toward a style of the users choice. I demonstrate that the SAT model is able to steer a neutral prompt to fit the selected style. The context information (style) will be embedded using RoBERTa (Liu et al., 2019b) so generation in a style will be possible by sampling from a given styles' embeddings. While outside the scope of this project, the SAT architecture may provide the ability for a model to generate in styles that it has not been trained on by moving in the embedding space toward the untrained style.

1 Introduction

In recent years Large Language Models (LLMS), for example GPT-like models Radford et al. (2019) Brown et al. (2020) have shown promise at generating coherent text, but controlling their outputs has proven difficult. Controlling the style of a text that an LLM generates is important because style is an integral part of human level text generation. The style in which one person writes a passage of text for one particular audience could be completely different from the way another author communicates the same content and information for a different audience. For example two different news articles might report the same event with different sentiment or ”voice”. Similarly, two authors writing for the same newspaper will still write different text since they are different people with different personalities and backgrounds. While state-of-the-art LLMs can create sentences, paragraphs, or even longer passages of generated texts that hold some consistency of subject from the prompt, these same LLM have no built-in mechanism for controlling the style of that generated text. The goal of this work is to create and train a model, the Style Adjustment Transformer (SAT), to show proof of concept of controlled style generation without the need to retrain an entire LLM or design a prompt that loosely controls the generation.

To-date, three main techniques have been proposed to control text generation in LLMs: 1) prompt design 2) fine-tuning, or 3) beam search with a heuristic. Each has limitations. Prompt design (Reynolds and McDonell, 2021) allows only mild generation control. Control is reduced as the output gets longer, and meanwhile, the prompt designer must craft a new prompt to elicit each separate desired style. Fine-tuning requires training the model on a particular style–each style requiring a different fine-tuned model. Further, fine-tuning often suffers from hallucinated and out of place context from the fine-tuning dataset. Beam search with a heuristic (Meister et al., 2020) is limited by the heuristic being used. This is often a classifier that similarly must be trained on a dataset that is from a particular style, each style requiring it's own classifier.

Other more advanced methods like Frans (2021) require many forward passes of a GPT like model and rely on the LLM classifying its own generated

[1] https://github.com/GregoryEHunter/SAT_MODEL

style which is often inaccurate. Retraining a LLM from scratch to consider style in it's pretraining task would be extremely expensive and possibly hinder the unsupervised nature of its training.

The SAT model is a transformer that utilizes cross attention to relate the context embeddings to the logits of a LLM, for the purposes of this proof of concept it only utilizes the last decoder block logits of the language model. Neither the weights of the model used to create the context embedding or the weights of the LLM are changed. The SAT requires only one trained model, is inexpensive to train, requires only one forward pass at inference time, and might allow the generation of unseen styles due to the learning of a map between context embeddings and LLM activations, making it a more robust approach to style controlled text generation. To my knowledge, all previous research that has attempted to control generation without re-training an entire LLM (Cao and Wang, 2021) (Jin et al., 2020b) has sought to control both the content and the style of the generation simultaneously; And none have attempted to control style with an architecture similar to the SAT.

The novel contributions of the this work are as follows:

1 Style control is achieved by learning a ”mapping” between two pretrained models. To the best of my knowledge this type of ”mapping” has never been attempted.

2 There is limited work in controlling style in free-form generation.

3 This approach allows for the possibility of unseen style generation if the context embedding and activation space mapping is sufficiently sampled.

2 Related Work

This work is closely related to two ongoing areas of research in NLP: Text Style Transfer and controlled LLM generation.

2.1 Text Style Transfer

Text Style Transfer is automatically editing a text into a particular voice or style. Many works relating to Text Style Transfer have been compiled in ”Deep Learning for Text Style Transfer: A Survey” Jin et al. (2020a), updated in Deep Learning for Text Style Transfer Github. Text style transfer has two goals, the preservation of information (not relevant for this work) and style transfer.

Much of the relevance of Text Style Transfer to this work has to do with the disentanglement of latent style from text and the use of stylized text data that does not directly relate to the model's task. Jin et al. (2020b) generates stylized headlines (Humorous, Romantic, or Click-Baitey) for articles by sharing parameters between a decoder trained on neutral newspaper headlines and an encoder trained on unrelated stylized text. The pattern of generating text via a decoder while influencing that generated text via a style encoder model is a common theme in both Text Style Transfer and LLM generation. This work also uses this paradigm.

To the best of my knowledge the most similar approach to this work is Cao and Wang (2021), where the final logits of BART (Lewis et al., 2020), a denoising autoencoder, are altered during inference based on a style classifier. Similar to Cao and Wang (2021) the model proposed in this work also modifies the logits of a transformer based model, however my logit adjustments are mediated by a dedicated model instead of being a byproduct of a modified loss function.

Riley et al. (2020) and Subramanian et al. (2018) modify the style of a text by conditioning denoising autoencoders on a style vector. The difference between Riley et al. (2020) and Subramanian et al. (2018) and this work is that their style vector is input into the decoder in parallel with the text (i.e. as an input to the entire network) while this work uses the style vector to adjust the activations internal to the decoder network. Riley et al. (2020) and Subramanian et al. (2018) then train the entire decoder network while I train only the adjustment layer.

2.2 Controlled LLM generation

The schema outlined in Prabhumoye et al. (2020) for the controllable text generation process and the subsequent breakdown of all five of their postulated modules is very robust and while not directly applicable in the implementation of this work still provides enlightening insights. I found the generalization to multiple approaches very insightful, this also led me to John et al. (2019). In John et al. (2019) they successfully show the disentanglement of content and style in the latent space, in this work I would like to adjust the style latent space. Though the method for disentanglement and goal is different in this work when compared to John et al. (2019). The goal of this work is similar to the goal of casual response optimization in Frans (2021) in that I also try to ”extract knowledge” (Frans, 2021) from GPT. However, whereas Frans (2021) focuses on prompting GPT-3, getting multiple generations, and then re-prompting GPT-3 with a question to control style, this work trains the SAT model to learn a mapping of the style space between two pretrained models.

3 Data

I am using the IMDb Movie Reviews Dataset (Maas et al., 2011), which can be found on the HuggingFace datasets hub, and raw text from ”On the Origin of Species” by Charles Darwin along with ”The Great Gatsby” by F. Scott Fitzgerald.

3.1 IMDb Movie Reviews Dataset

The IMDb Movie Reviews Dataset (Maas et al., 2011) consist of 50,000 reviews labeled as positive or negative. Only highly polarizing positive and negative reviews are used i.e reviews with scores $\leq$ 4 out of 10 for negative and $\geq$ 7 out of 10 for positive. It is split into 25 thousand training examples and 25 thousand test examples. Both the test and train sets contain 12500 positive and negative reviews. The mean length of each review in the train set is approximately 1325 words. To pre-process the data all of the HTML is removed and GPT-2 logits and RoBERTa (Liu et al., 2019b) context embeddings are produced from a entire review. The RoBERTa embeddings are than averaged. Both the context embedding and logits are stored to later be inputted into the SAT. The following is an example excerpt from the The IMDb Movie Reviews Dataset.

> Alright this was quite a sensitive little number but I can’t help thinking I’ve seen it before. Reminds me of another VCA film I saw at Poitier called ”THE OTHER DAYS OF RUBY RAE” Also had specks of ”Welcome to the Dollhouse” and ”Ratcatcher” and Lynne Ramsay in it’s execution. Which is not to say that they’re not tasteful references...just that they feel very modern and very fashionable...which makes me feel like this is closer to advertising (as an approach in style and story) than the work of an original and authentic auteur to come. The cinematography is just...too perfect for my liking. Too coral filter (or charcoal) for my liking...too archly framed 12mm. Therefore the entire impression left me a little distant – beware of art that proclaims itself too readily! The french (they are a conservative bunch) seemed to buy it as did the jury however... but Cannes short film selection is notoriously conservative compared to it’s feature selection although I wonder what’s been happening in the last few years.

3.2 Origins and Gatsby

Additionally raw text versions of both ”The Great Gatsby” by F. Scott Fitzgerald (Fitzgerald, 1925) and ”On the Origin of Species” book by Charles Darwin (Darwin, 1859). The reasoning behind using these two books was because they are different genres and from different times. Consideration was made into using a larger portion of Project Gutenberg (Gutenberg, 1971) books but unfortunately computing resources and time did not allow for this. To pre-process both books the text is truncated to 384 token sequences, cleaned to remove any trailing or leading weight space or returns, and inputed into both a fine-tune version (Face, 2020) of Micrsoft’s MPNet (Song et al., 2020) to produce context embeddings and GPT-2 to produce logits. Excerpts from ”The Great Gatsby” and ”On the Origin of Species” follow respectively.

> The practical thing was to find rooms in the city, but it was a warm season, and I had just left a country of wide lawns and friendly trees, so when a young man at the office suggested that we take a house together in a commuting town, it sounded like a great idea. He found the house, a weather-beaten cardboard bungalow at eighty a month, but at the last minute the firm ordered him to Washington, and I went out to the country alone. I had a dog — at least I had him for a few days until he ran away — and an old Dodge and a Finnish woman, who made my bed and cooked breakfast and muttered Finnish wisdom to herself over the electric stove.

In regard to animals, much fewer experiments have been carefully tried than with plants. If our systematic arrangements can be trusted, that is, if the genera of animals are as distinct from each other as are the genera of plants, then we may infer that animals more widely distinct in the scale of nature can be crossed more easily than in the case of plants; but the hybrids themselves are, I think, more sterile. It should, however, be borne in mind that, owing to few animals breeding freely under confinement, few experiments have been fairly tried: for instance, the canary-bird has been crossed with nine distinct species of finches, but, as not one of these breeds freely in confinement, we have no right to expect that the first crosses between them and the canary, or that their hybrids, should be perfectly fertile.

4 Methods

4.1 SAT Model Variant Overview

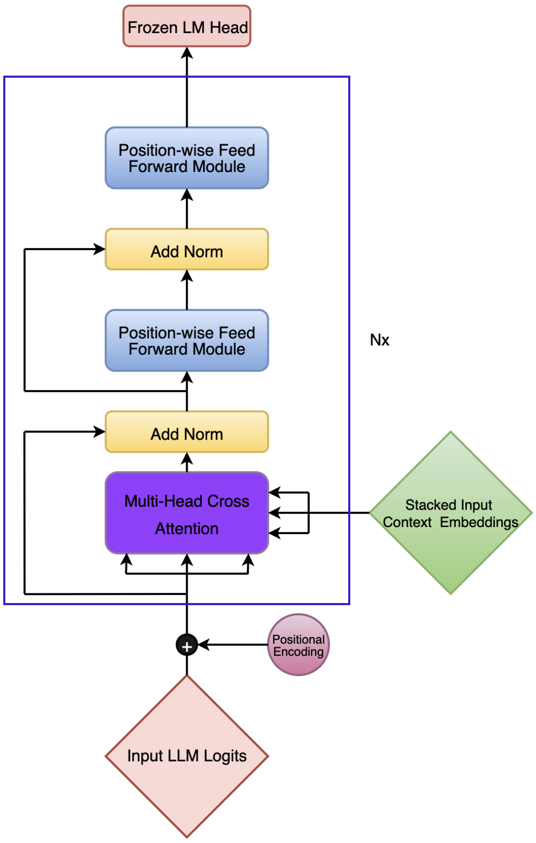

Frozen LM Head
Position-wise Feed Forward Module
Add Norm
Position-wise Feed Forward Module
Nx
Add Norm
Multi-Head Cross Attention
Stacked Input Context Embeddings
Positional Encoding
Input LLM Logits

Figure 1: SAT Variant Architecture

Figure 1 broadly outlines how the Style Adjustment Tranformer (SAT) uses embeddings of context sentences to adjust the inputs of the final layer of a LLM. I refer to these final layer inputs as logits. To save compute, I pre-calculate all context Embeddings and logits. Additionally the head of the LLM is frozen. LLM logits can be precomputed because causal masking is used during generation. Precalculating the embeddings and logits is equivalent to freezing the weights of the two input models. The SAT adjusts the LLM logits, learning a map between the context embedding space and the ”activation space” of the LLM’s decoder blocks, and adjusting the output of the LLM toward a particular style.

This paper experimented with two Context Embedding models: RoBERTa (Liu et al., 2019b) and a fine-tuned version (Face, 2020) of Micrsoft’s MPNet (Song et al., 2020). The LLM model being used is GPT-2 (Radford et al., 2019). The GPT-2 variant being used is 1.5 billion parameters; The Roberta variant is 354 million parameters; and MPNet, 110 million parameters. The 1.5 billion parameter version of GPT-2 was chosen both because of it’s successful generalization to other tasks Radford et al. (2019) and because it was the largest model that could be used within the compute limits for this project. RoBERTa was chosen because it has been shown to be successful at various NLU tasks (Tarunesh et al., 2021). Due to RoBERTa’s lack of pre-traning on the NSP (Next Sentence Prediction) task, I also experimented with the fine-tuned MPNet from (Liu et al., 2019b). This MPNet model has relatively high performance given it’s fairly low parameterization (sbert.net, 2020).

The SAT model shown in Figure 1 is made up of: a positional encoder, Multi-Head Cross-Attention modules, Feed-Forward modules, and normalization layers. The positional encoder is used to reintroduce positional information into the first Multi-Head Cross-Attention module. The Multi-head Cross-Attention is the heart of the model that finds similarities between the context embeddings and the logits. The Feed-Forward modules are position wise and made up of two linear layers separated by a ReLU layer. Skip connections and Layer Normalizations are standard components of transformer based models Vaswani et al. (2017) and are thus also included.

4.2 Training and Compute

All reported experiments used a model variant with two stacked SAT Multi-Head Cross Attention blocks. The parameterization of this stacked model variant was 105,838,400. Each model was trained on an Nvidia A100 for ten epochs. Two versions of the SAT model were produced: One trained on context embeddings from the ”On the Origin of Species” and ”The Great Gatsby” (The ”Author model”) and the other trained on context embeddings of positive and negative movie reviews (the ”Sentiment model”.) After optimization, training time of the Author model was 2 minutes and 30 seconds for 10 epochs. The Author model used the fine-tuned MPNet (Liu et al., 2019b) to get context embeddings, whereas the model trained on the positive and negative movie reviews used an average of RoBERTa (Liu et al., 2019b) word embeddings. Both models used 100 randomly sampled context embeddings during training. Context embeddings were also split into training and test set to prevent information leakage. An RAdam optimizer was used to avoid model divergence and increase performance (Liu et al., 2019a). As the SAT model takes two inputs, the context embeddings and LLM logits, during training a given example's GPT-2 logits are inputted along with 100 stacked randomly sampled context embeddings from the style of the GPT-2 logits being considered. During generation 100 stacked randomly sampled context embeddings of the desired style are inputted with the GPT-2 logits of the prompt and any tokens that have been inferred.

4.3 Baselines

As baseline models against which to compare the SAT model, I Fine-tuned four versions of GPT-2. Each model was trained on one of the four styles (Origin, Gatsby, positive reviews, and negative reviews.) All GPT-2 weights were trained. Each was also trained on an NVidia A100 for ten epochs with the same dataset splits as the SAT models. An Adam optimizer with warmup was used to train the fine-tuned models (Liu et al., 2019a) and HuggingFace's standard GPT-2 pre-training was used from Wolf et al. (2019). In addition to the finetuned GPT-2 models, I also generate passages using the base GPT-2 without any Fine-tuning. The final baselines against which I compare the SAT generations are randomly sampled excerpts from each of the differently styled datasets. These excerpts came from development data and were the same length as the generations from both the fine-tuned models and the SAT models.

4.4 Generation

For the SAT models Top-k sampling was used during generation. So the k, 3 in this case, tokens with the highest probability were identified. Then the probabilities of the k tokens were divided by their sum and the next token was sampled from the new distribution. Top-k sampling was chosen so that I could look at the most probable tokens during debugging. At inference 100 stacked randomly sampled context embeddings from a dev set were used so the inputted stacked context embedding size matched the size used during training. For the Fine-tuned models and GPT-2 without fine-tuning Top-p sampling was used with a probability mass of .5. Neither generation method used temperature.

5 Experiments

As I had developed multiple versions of the SAT model, a limited experiment to conclude what model was best was conducted. I will briefly describe this process in the Model Iteration subsection. In addition an evaluation dataset was created and used to both conduct a human study and be automatically evaluated. After a model variant was selected, an SAT model was trained for both of the two style datasets, Origins and Gatsby and IMDb Movie Reviews. Both SAT models being trained on two styles. Each SAT model produced 100 generations for each style, so 400 SAT generations in 4 styles from 2 models. Using the same prompts each Fine-tuned model produced 100 generations each, so 400 Fine-tuned generations in 4 styles from 4 models. Additionally, 100 generations were produced by GPT-2 without any Fine-tuning and 400 excerpts were extracted directly from the datasets. All generations and excerpts were 50 tokens long. In total there were 900 50 token generations from the models and 400 excerpts resulting in 1300 50 token passages. The prompts for all models were 100 randomly selected bigram sentence starts that were present in both ”On the Origin of Species” and ”The Great Gatsby”. An evaluation dataset consisting of 1300 passages in a random order was then created and used to conduct both a human and automatic evaluation.

5.1 Model Iteration

From the beginning I planned to use Cross attention to relate the context embeddings to a LLM's logits. Model performance and divergence was a problem until I utilized an RAdam optimizer (Liu et al., 2019a). Initially cross attention was built from scratch but when I made the choice to use multiple attention heads I utilized pytorch's (Paszke et al., 2019) built in MULTIHEADATTENTION class because it is highly optimized and capable of being adapted for cross attention. After the model was training and generating reasonably I attempted to stack blocks of the model. Stacking a block that utilized Multi-head Self-Attention after the initial Multi-head Cross-Attention block led to significant repetition and decreased performance. This is a point of interest and will be addressed in future work. In total Fourteen distinct SAT models were made. Though significant changes were made internally to each of these variants during experimentation. Of these, 5 were evaluated by training on a portion of the Authors dataset and evaluated using a logistic regression classifier. The dataset was split randomly with a different seed from the version in Automatic Evaluation to avoid train test leakage. Mild performance improvement was observed when two cross attention blocks were stacked and positional Encodings were added to the GPT-2, so a variant of the SAT model with two Multi-Head Cross-Attention blocks each using 8 attention heads was chosen.

5.2 Human Evaluation Setup

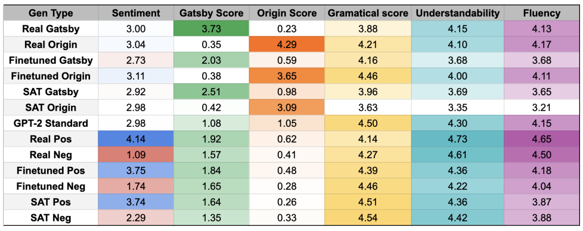
To conduct the human evaluation I distributed the evaluation dataset to 30 individuals. In total 15 participated. Each participant received a google doc with directions asking them to answer 5 questions for as many of the passages as they could. The document also included two passages taken from ”The Great Gatsby” and ”On the Origin of Species”. The passages were selected to not include character names or other subjectively obvious indicating content that the models could produce without actually writing in a particular style. The chosen passages are shown as examples of ”The Great Gatsby” and ”On the Origin of Species” in the Data section. As each generation was 50 tokens and did not necessarily end at a natural stopping point, participants were asked to not consider sentences that did not terminate when answering. The 5 questions were: ”Rate the sentiment of the text (0 = negative 5 = positive)”, ”How similar is this passage to the style of: The Great Gatsby (0-5)”, ”How similar is this passage to the style of: On The Origin of Species (0-5)”, ”Is this passage: Grammatical? (0-5)”, ”Is this passage: Understandable? (0-5)”, ”Is this passage: Fluent? (0-5)”. 560 passage's questions were fully answered each participant filling out an average of 37.3. Annotators were expected to understand how to score sentiment and how grammatical the generation or excerpt was. However, to clarify how to score how understandable or fluent a passage was, annotators were directed ”Please base your score for understandable on how easy it was to understand the text and please penalize the fluency score when the text is discursive, has disconnected ideas, unclear syntax, out of place words and characters, or repeats itself.”. The minimum number of annotations from a participant was 15 and the maximum was 100. The range of the passages annotated for each type of model and excerpt was 37-49. I prioritized having more passages annotated so no inter-annotator agreement was done in this human evaluation because of the relatively small number of participants and large amount of classes to label (13).

5.3 Automatic Evaluation Setup

The automatic evaluation was done using 3 classifiers: a logistic regression classifier trained on positive vs negative movie reviews, a logistic regression classifier trained on ”The Great Gatsby” vs ”On The Origin of Species”, and a HuggingFace sentiment classifier. The logistic regression classifiers were trained on sentences from the four styles, two styles each. To avoid training the logistic regression classifiers on sentences that had been extracted and put into the evaluation dataset, any sentences in the train data that overlapped with the excerpts were removed. Unigrams, Bigrams, and Trigrams were extracted and a countvectorizer was fit and used to vectorize the training corpora. The excerpts and generations were then transformed with the countvectorizer previously fit to the training corpus. The logistic regression classifier trained on positive vs negative movie reviews and the HuggingFace sentiment classifier were used to predict and score the real positive and negative IMDB review excerpts, and both the positive and negative IMDB review fine-tuned model generations, and the sentiment SAT models positive and negative generations. The logistic regression classifier trained on ”The Great Gatsby” vs ”On The Origin of Species” was used to predict and score the real Gatsby and Origin excerpts, and both the Gatsby and Origin fine-tuned models generations, and the Gatsby/Origin SAT models Gatsby and Origin generations. I also calculated accuracy scores and generated confusion matrices.

6 Results

6.1 Human Evaluation Results



Gen Type	Sentiment	Gatsby Score	Origin Score	Gramatical score	Understandability	Fluency
Real Gatsby	3.00	3.73	0.23	3.88	4.15	4.13
Real Origin	3.04	0.35	4.29	4.21	4.10	4.17
Finetuned Gatsby	2.73	2.03	0.59	4.16	3.68	3.68
Finetuned Origin	3.11	0.38	3.65	4.46	4.00	4.11
SAT Gatsby	2.92	2.51	0.98	3.96	3.69	3.65
SAT Origin	2.98	0.42	3.09	3.63	3.35	3.21
GPT-2 Standard	2.98	1.08	1.05	4.50	4.30	4.15
Real Pos	4.14	1.92	0.62	4.14	4.73	4.65
Real Neg	1.09	1.57	0.41	4.27	4.61	4.50
Finetuned Pos	3.75	1.84	0.48	4.39	4.36	4.18
Finetuned Neg	1.74	1.65	0.28	4.46	4.22	4.04
SAT Pos	3.74	1.64	0.26	4.51	4.36	3.87
SAT Neg	2.29	1.35	0.33	4.54	4.42	3.88

Figure 2: Human Evaluation Table Averages

6.1.1 Human Evaluation Table

Unless otherwise noted all result comparisons from the human evaluation were statistically significant at the 99% level using a one sided t-test. All results will be reported as averages of the scores a given model received for a given question. The Human Evaluation Table in Figure 2 can be used as reference. In Figure 2 the more blue a cell is in the sentiment column the more positive humans identified that model's generation to be, the more red it is the more negative humans identified it to be. For all other columns the darker the color gets the higher the score is.

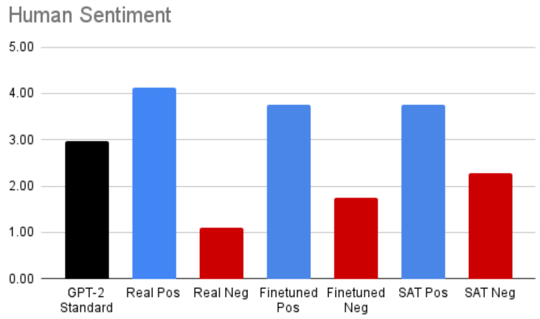


Figure 3: Models and Excerpt Human Evaluated Sentiment

Figure 3 is a bar graph showing the human evaluated sentiment for the 561 passages annotated for sentiment. The bars for the models and excerpts that were supposed to be negative are shown in red, the bars for the models and excerpts that were supposed to be positive are shown in blue and the bar for GPT-2 without any fine-tuning is shown in black. There were 37 annotations for real positive excerpts, 44 annotations for real negative excerpts, 44 annotations for fine-tuned positive generations, 46 annotations for fine-tuned negative generations, 47 annotations for SAT positive generations, and 48 annotations for SAT negative generations. All Gatsby and Origin of Species generations and excerpts were within .14 of 3 (a neutral score) and will not be reported. Annotators were asked to rate each sentence from 0 to 5, where 0 is very negative and 5 is very positive. Here ”SAT Pos” and ”SAT Neg” are one SAT model trained on two styles. As expected, the Real excerpts outperform both the fine-tuned and SAT models for both positive and negative with the average positive excerpt score at 4.14 and the average negative excerpt score at 1.09. Interestingly the SAT positive generation (the Sentiment SAT model inputted with positive sentiment) performs almost as well as the fine-tuned positive model with scores of 3.74 and 3.75 respectively. Here the SAT Sentiment model negative generations do however do worse than the fine-tuned model with a average score of 2.29 vs 1.75 where lower is better (more negative.) Still, overall the human annotators identified the Sentiment SAT model generations as negative. GPT-2 without any fine-tuning performed right around neutral at 2.98 as expected.

6.1.2 Human Evaluated Gatsby and Origin scores

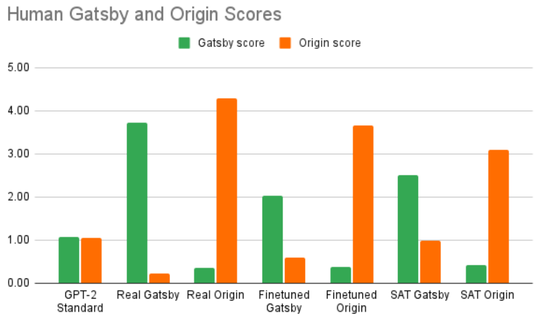


Figure 4: Models and Excerpt Human Evaluated Gatsby and Origin scores

Figure 4 shows the average human evaluated scores for how much a given generation or excerpt was like either ”The Great Gatsby” or ”On the Origin of Species” based on both the passages shown in the Data section and a preexisting knowledge of the two books. 560 generations and excerpts were labeled by humans for this question. The sentiment related excerpts, fine-tuned models, and sentiment SAT model are not shown in this graph, however the scores comparisons are statistically significant based on a one tailed t-test. The real positive excerpts scored 1.92 for Gatsby and .62 for Origin, the real negative excerpts scored 1.57 for Gatsby and .41 for Origin, the fine-tuned positive model generation scored 1.84 for Gatsby and .48 for Origin, the fine-tuned negative model generation scored 1.65 for Gatsby and .28 for Origin, the Sentiment SAT model using positive context embedding scored 1.64 for Gatsby and .26 for Origin, and the the Sentiment SAT model using negative context embedding scored 2.29 for Gatsby and .33 for Origin. While these scores are not for generations and excerpts that should score high in either Gatsby or Origin, the Gatsby scores seem to be slightly higher than expected with an average of 1.66. GPT-2 without any fine-tuning had a Gatsby score of 1.08. While this is not a major concern it is worth noting as the fine-tuned Gatsby generations and the SAT model inputted with Gatsby context generations only have a Gatsby score of 2.03 and 2.51 respectively. Here though the SAT model with Gatsby context does outperform the fine-tuned Gatsby model baseline. Focusing back Figure 4, the orange bars indicate the Origin scores for a particular type of generation or excerpt and the green bars indicate the Gatsby score for a particular type of generation or excerpt. The relevant model annotations count are: 40 annotations for real Gatsby excerpts, 48 for real Origin excerpts, 37 for fine-tuned Gatsby generations, 37 for fine-tuned Origin generations, 49 for the SAT model with Gatsby context, 43 for the SAT model with Origin context, and 43 for GPT-2 standard (no fine-tuning.) Real Gatsby excerpts perform the best on the Gatsby score (3.73) and real Origin excerpts perform the best on the Origin score (4.29) as expected. The Gatsby score for the Gatsby/Origin SAT model generations when Gatsby context is taken outperforms the fine-tuned Gatsby baseline however neither was identified to be strongly in the style of Gatsby (2.51 vs 2.03.) The Origin score for the Gatsby/Origin SAT model generations when Origin context is taken is slightly outperformed by the Origins fine-tuned baseline (3.09 vs 3.65.) GPT-2 Standard scores low in both Gatsby and Origin scores, 1.08 and 1.05 respectively.

6.1.3 Human Evaluated Grammatical, Understandability, and Fluency

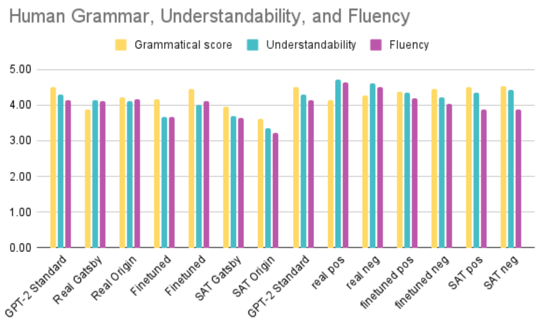


Figure 5: Human Grammar, Understandability, and Fluency Scores

Figure 4 is a graph showing the Grammaticality, Understandability, and Fluency Scores. The SAT model very mildly underperforms all baselines as can be seen in Figure 2, However the SAT model using Negative context at inference outperforms all other baselines in grammaticality. This is unexpected as the participants seem to have scored it higher than the real excerpts. These scores are less statistically significant however, with a one sided t-test is only statistically significant at a 90% level. The mean for the Grammar score is 4.24, the mean of Understandability score is 4.15, and the mean of the Fluency score is 4.02. Across all three of these scores the SAT Gatsby/Origins model underperforms the mean with both context inputs. Yet the Sentiment SAT model outperforms the mean and the best in grammar when using negative context with a score of 4.54.

6.2 Automatic Evaluation Results

6.2.1 Accuracy scores

The logistic regression classifier trained on Gatsby vs Origin identified 98% of the real excerpts from ”The Great Gatsby” and ”On the Origin of Species” correctly indicating it was trained correctly. For the fine-tuned models it classified 93.5% of the generations made by the models as the style they were trying to generate in. For the SAT model it identified 81.5% of the generations made using a styles context embedding as that style.

The logistic regression classifier trained on Negative vs Positive IMDb reviews identified 85% of the real excerpts from the reviews correctly. While this is not as high as the score that the other Gatsby vs Origins classifier achieves, the logistic regression classifier is still reasonably accurate. For the fine-tuned models it classified 77.5% of the generations made by the models as the style they were trying to generate in. For the SAT model it identified 68.5% of the generations made using a styles context embedding as that style.

The HuggingFace sentiment classifier identified 79.5% of the real excerpts from the reviews correctly. This score is lower than the score my classifier achieved. As the HuggingFace sentiment classifier was not directly trained on the movie reviews this makes some sense. Because of the lower score, I will conduct all further automatic evaluation experiments using the logistic regression movie review classifier.

Here both SAT models underperform their two fine-tuned counterparts. However the SAT models are still capable of multi-style generation and and the models are significantly more efficient to train than fine-tuning GPT-2.

6.2.2 Confusion Matrices

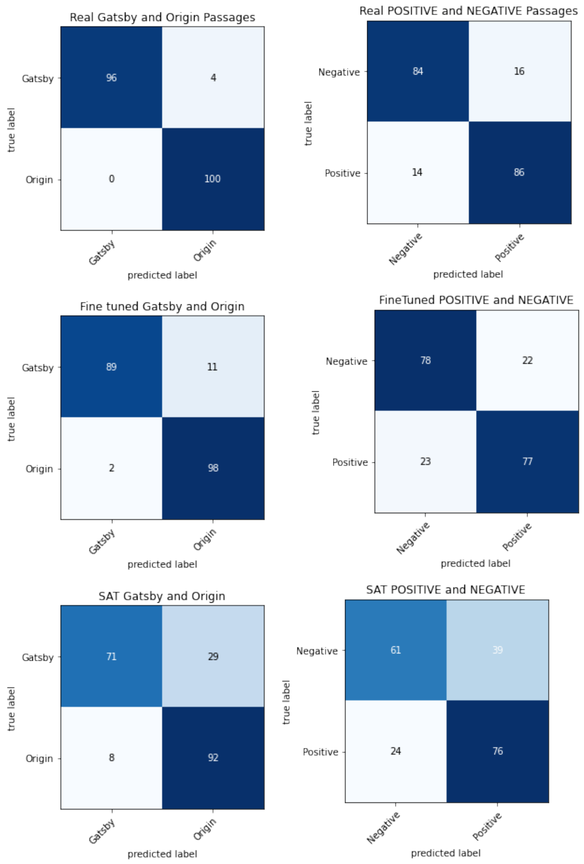


Figure 6: Confusion Matrices

Figure 6 shows the confusion matrices for the two logistic regression classifiers. It is worthy of note that throughout the Gatsby vs Origin confusion matrices there is a large imbalance between the Gatsby false positive rate and the Origin false positive rate. This is likely because there are more sentences in ”On The Origin of Species” (3941) than there are in ”The Great Gatsby” (2439). This imbalance biases the classifier toward labeling the style of the input as ”On the Origin of Species”.

6.2.3 Human Evaluation and Automatic Evaluation correlation

The Pearson’s R correlation coefficient between the regressor’s predicted probabilities for negative and positive sentiment and the human scores when asked to rank passages as negative and positive is strong at 0.65. The correlation coefficient between the regressor and the human labeled scores was moderate for Gatsby. When only considering excerpts from Gatsby and generations that are meant to sound like Gatsby, the correlation between the classifier’s probability and the human scores was 0.37. The correlations between regressor and human scores for Origin passages was close to zero at -0.05. While the lack of correlation for the Origin values and lower correlation for Gatsby values is surprising, this result is understandable as the questions asked of the humans were significantly different than the task given to the classifier. The classifier is asked to either classify something as Gatsby or Origins whereas the humans are asked to rank how much a given passage is like Gatsby or Origins. It is conceivable that one passage might have some similarities to both books. In the future, better human study design and using a classifier that is trained on Gatsby/Origins vs everything else will likely see these correlations go up. The difference between the Gatsby/Origin correlation and the sentiment correlation is likely because the sentiment question asked to the human and the task given to the classifier are more similar.

Only considering automatic and human scores on real excerpts from The Origin of Species still yields a correlation of 0.09 not much better than only considering SAT generation (0.08). When only considering Origin fine-tuned, however, the correlation is -0.13. The correlation is actually highest for Gatsby when only considering SAT generations at .4780 vs .2813 for real excerpts and .2097 for fine-tuned generations. The correlations across real, fine-tuned, and SAT and the human scores are all similar for the negative vs positive logistic classifier.

6.2.4 Human Evaluation and Automatic Evaluation Overlap

When all passages that humans labeled are classified by the logistic regression positive vs negative review classifier, and human scores are considered to be negative when below 2.5 and positive when above 2.5, there is a 68.9% overlap between the automatic evaluation and the human evaluation. Additionally, in an attempt to measure ”disagreement” I used the Gatsby vs Origins logistic classifier's probability of both Gatsby and Origins and divided the human score by 5 for all entries filled out by humans to get the scores between 0 and 1. I then found the average of the absolute value of the difference between the probabilities and the adjusted human scores. I then multiplied these averages by 5 in an attempt to quantify the ”disagreement” between the Gatsby/Origin human scores and the classifier's scores in terms of the human evaluation. This ”disagreement” value across all entries for Gatsby was 1.709 and 1.823 when only considering Gatsby excerpts and attempted generations. For Origin the value across all entries was 1.75 and 1.26 when only considering Origin excerpts or generation attempts. This ”disagreement” value going up for Gatsby when only considering relevant generations and excerpts was interesting given that humans frequently miss-scored Real Gatsby excerpts and had the tendency to misclassify Unrelated passages as Gatsby, 2. Interestingly while the ”disagreement” was lowest for Gatsby when only considering real excerpts at 1.2169, the SAT generations had significantly less ”disagreement” than the fine-tuned generations at 1.69 and 2.64 respectively.

7 Error Analysis

Subjectively the most common errors with the SAT model were: Repetition, Reversion back to GPT-2 Standard, human error, and occasional leakage between the styles.

7.1 Repetition

In the example in the next sentence the model repeats itself when trying to generate a positive review. “How do you know this movie is the worst movie ever made? Well, it's not the worst movie ever, it's the worst movie ever made. It's not even the worst movie ever, it's the worst movie ever made. The only thing worse”. The model switches back and forth between positive clauses and negative clauses. Examples like this seem to occur due to a mixture of Top-k sampling and the IMDb review dataset. They only occur, as far as I know, with the SAT trained on the positive and negative movie reviews. While the majority of the words that the model proposes are in the style of the given context embeddings, after a common word like ”movie” is produced, among the following top-k tokens will be a word of the opposite style. If this opposite-style token is selected, then the model will start to repeatedly contradict itself. I think the possibilities are: 1) The style is just leaking because the model is trained on both styles or 2) The other style's token has now been fed back into GPT-2 which increases the probability of that style to occur in the next token but the SAT is adjusting the logits back so the probabilities just end up being somewhere in the middle of the two styles. This in turn causes the waffling repetition. 3) Search errors: Due to time contraints I only experimented with greedy generation algorithms. Beam search generation algorithms could look ahead and realize that a particular token reduces the final probability of the generated sequence. To investigate the waffling generations, I will in future work put the logits of GPT-2 that are fed to the SAT through the frozen LM head, softmax them, and then also look at the output probabilities of the SAT model to see if the SAT is pushing back against the increased probabilities of the other style. Additionally I would like to explore other paths without using a heuristic to see if it is just a problem with the Top-k search method.

Repetitions also can occur when the model is not contradicting itself and is generating in the correct style like the example in the next sentence where the context is provided for a negative movie. “If I had seen this film when I was a little girl, I would have been so disappointed in myself. I was so disappointed in myself that I wouldn't have believed that a movie could be so bad. I was so disappointed in myself that I would hav”. The generation is negative but it seems to get stuck on the phrase ”so disappointed”. This is likely to be a problem stemming from Greedy sampling.

7.2 Human Error

Sometimes the human evaluators were not paying much attention to the passage they were evaluating. For instance the next sentence was scored as a 5 out 5 (maximum positive) by an annotator. “And it's not as if I have to be a fan of this movie to like it. It's just the fact that the plot line is just plain ridiculous. I mean, the movie starts out with the most ridiculous and implausible plot line I've seen". While the first sentence could imply the passage is positive the rest of the passage is pretty negative.

7.3 GPT-Standard reversion

In the example in the next sentence, although the context sentences supplied to the SAT model were from Gatsby, the model seems to revert back to more standard unstylized GPT-2 generations: "No one will dispute the fact that thousands are killed in drone attacks, and as the Middle East and Yemen have gotten more dangerous, so have drones. But to dispute the fact is willful blindness to the truth, which is that drones are an indispensable tool for the". The reason this sentence sounds like modern unstylized text is likely because I am using GPT-2 XL in this experiment. I have noticed during finetuning and with previous tests with the SAT model that, because I am training SAT on smaller datasets and because I am currently using a proof of concept model, using the biggest possible model does not really make since. However, I will assess it in the future like the repetition error by putting the GPT-2 logits through the LM head, softmaxing the output, and checking the SAT model probabilities (after the LM head) to see if the model is properly adjusting the logits. I will need to start keeping track of top-k tokens and store LLM logits and SAT outputs as well.

7.4 Style leakage

Sometimes, the SAT model simply produce text in the opposite style. In this first example, the context was sentences from Gatsby, but the style is more reminiscent of The Origin of Species: "There is no doubt that many species, naturalised through mans agency, have spread with astonishing rapidity over new countries, and have added to the numbers of the inhabitants, both in the Old and New Worlds." Meanwhile, in this second example, the context is Origin of Species, but the style leans toward The Great Gatsby: "But I have been urged not to do this thing, not to speak to this man, not to see this man. It would ruin my own reputation, which would thus become just as much at risk as his own; I cannot let it happen." Since the SAT model is trained on sentences from two completely different styles, it is understandable that one style might bleed through into the other.

8 Conclusions, Limitations and Future Work

8.1 Conclusions

In this paper I have proposed and implemented a new type of model (SAT) that can be used to generate at least two styles when provided with those styles' context embeddings. I have shown that the SAT model successfully allows the user to direct the style of the generated output text. Both human and automatic evaluations clearly differentiate between passages generated with the SAT model when different context embeddings are provided. While the SAT model does not yet outperform models which are finetuned toward one particular style, one single SAT model can output text in more than one style (which an individual finetuned model cannot do), and training an SAT model is significantly less computationally expensive than full finetuning. This work also found that the correlation between the human and automatic evaluation of at least some styles (e.g. positive vs negative) are well correlated. This correlation will permit future experiments with the SAT model to quickly assess the quality of the generated styles without resorting to repeated time-consuming human evaluations.

8.2 Limitations

The long term goal of this work is to train one SAT model on a variety of different styles, and then generate text in not only the styles that were in the training data, but also similar styles not directly found in the training data. This work instead focuses on the more limited scenario of training and testing on 2 highly contrasting styles at a time (although preliminary experiments training on multiple different styles are promising.)

This work is also limited by the available compute. Models larger than GPT-2 simply did not fit in RAM on the machines I had available. Meanwhile, recently released models like ChatGPT, which are several orders of magnitude larger than GPT-2, are able to modify the style of the text they generate without expliciting training for context (but they are also much more expensive to train and run).

8.3 Future Work

I am excited to continue to explore and evolve the architecture laid out in this work and continue to conduct research into light weight control of transformer based architectures. Right now I see the next steps as follows:

1 Increase the amount of styles the SAT model is capable of generating by using larger datasets.

2 Iterate On the SAT and SAT like architectures as some of the more promising architectures were not able to be included in this paper

3 Explore a similar architecture with a model pretrained on summarization

4 Work on content preservation with style control

5 Continue to work on the idea of mapping the ”embedding and activation space”

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Shuyang Cao and Lu Wang. 2021. Inference time style control for summarization. *CoRR*, abs/2104.01724.

Charles Darwin. 1859. *On the Origin of Species*. John Murray.

Hugging Face. 2020. Fine-tuned mpnet.

F.Scott Fitzgerald. 1925. *The Great Gatsby*. Charles Scribner’s Sons.

Kevin Frans. 2021. To extract information from language models, optimize for causal response. *kvfrans.com*.

Project Gutenberg. 1971. Project gutenberg.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020a. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020b. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5093, Online. Association for Computational Linguistics.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019a. On the variance of the adaptive learning rate and beyond. *CoRR*, abs/1908.03265.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. Best-first beam search. *CoRR*, abs/2007.03909.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *CoRR*, abs/2102.07350.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David C. Uthus, and Zarana Parekh. 2020. Textsettr: Label-free text style extraction and tunable targeted restyling. *CoRR*, abs/2010.03802.

sbert.net. 2020. Sentence transformer comparison.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *CoRR*, abs/1811.00552.

Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. 2021. Trusting roberta over BERT: insights from checklisting the natural language inference task. *CoRR*, abs/2107.07229.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.