Bilingual SAS: Evaluating subset selection algorithms for contrastive self-supervised learning on text data

Garvit Rajkumar Pugalia **Joo Wan Lim** **Jie Shao** **Zikun Gan**
{garvitpugalia, lim213, jieshao, zkunbruin08}@ucla.edu
Department of Computer Science, UCLA, Los Angeles, CA, 90024.

Abstract

Self-supervised learning (SSL) aims to learn feature representations from a large corpus of unlabelled data, which can be fine-tuned for downstream tasks. Within the context of text data, SSL can often suffer from "bad" examples, raising the question of selecting a quality subset of the data. In this paper, we translate and apply one such subset selection algorithm, SAS, from the image domain to the text domain for a downstream prediction task. As part of this context switch, we attempt to improve the SAS algorithm by incorporating Spanish translations' embeddings into the selection process. We show that Bilingual SAS consistently performs better than SAS in downstream evaluation on a reduced version of the AG News dataset; however, some of the results point to errors in the training process, which need to be further explored. We also show that GPT-4 based augmentation within contrastive training is significantly better than easy data augmentation techniques such as synonym replacement and random swapping, improving downstream performance by an average of 2%. The code and results can be found here.

1 Introduction

Self-supervised learning (SSL) is an important tool as it allows models to learn high-quality feature representations from unlabelled data. Within the context of text data, the importance is magnified due to the prevalence of the exact problem that SSL solves: insufficient resources or availability of well-annotated data in comparison to a large corpus of unlabeled data. Many papers have explored this research area. In fact, some of the most commonly used language understanding models such as BERT and GPT were trained in a self-supervised setting. [Devlin et al., 2019] While there have been some attempts to improve the data-efficiency of contrastive self-supervised learning within NLP, they focus on modifying the contrastive learning framework rather than selecting the "best" data points.

By selecting a subset of the unlabelled text data, we can intuitively improve data efficiency by reducing the overall computation, but we can also remove "bad" examples from the dataset that might negatively impact our learned representations. However, the process of selecting a "good" subset of data is not trivial. One such algorithm, SAS, attempts to maximize the expected augmented similarity when choosing data points for the subset. [Joshi and Mirzasoleiman, 2023] It estimates the augmentation similarity using a proxy model, and then picks examples that best represent their class in terms of alignment (to similar examples) and divergence (from dissimilar examples). This algorithm led to a performance boost in the learned embeddings' downstream evaluation within the image domain, outperforming a randomly selected baseline subset by over 3% on average. We attempt to apply this algorithm to the text domain in a similar fashion. Consequently, we attempt to improve the algorithm for the text domain by incorporating both English and Spanish embeddings' in the SAS algorithm.

For the evaluation of the original SAS algorithm and our bilingual SAS algorithm, we train embeddings on top of a pre-trained 'BERT-small' model using an unlabeled dataset of news articles i.e.

Preprint. Under review.

AG News, and evaluate the downstream text classification accuracy of those learned embeddings using the labeled equivalent of the dataset. For the contrastive learning process, we compare two different forms of augmentation: easy data augmentation (synonym replacement, random swapping) and GPT paraphrasing. Finally, we conduct ablation studies to further evaluate the efficiency of the SAS algorithm and its usability within the text domain.

2 Related Work

Due to the recency of the SAS algorithm, there has been no similar investigation into the application of SAS to text domain or subset selection within the context of self-supervised learning. Some researchers have explored the idea of selecting representative subsets for supervised tasks [Attendu and Corbeil, 2023] and for self-supervised learning in closely related domains such as automatic speech recognition (ASR) [Azeemi et al., 2023], however, the application of the original paper's algorithm to our text-based classification task is novel to this paper.

In terms of data augmentation for NLP, Wei et al. first introduced the idea of easy data augmentation (EDA) techniques to boost performance of text-based classification tasks. [Wei and Zou, 2019] We incorporate the synonym replacement and random swapping methods from this paper into our contrastive learning framework. Other researchers, in recent times, have also demonstrated the benefits of using large language models such as GPT for data augmentation via paraphrasing, which we incorporate as well. [Balkus and Yan, 2023]

3 Problem Formulation

Since we are trying to apply the SAS algorithm to a new domain, we will be incorporating new methods and models into our experiments; however, the overarching problem structure remains the same as the original paper, which didn't assume domain-specific information.

In summary, we have some unlabelled dataset $X = \{x_i\}_{i \in V}$ of $n = |V|$ training examples drawn i.i.d. from an unknown distribution. This data has some unknown, underlying latent classification that separates certain examples from others. Therefore, we assume that each example belongs to one of K latent classes.

The output of contrastive learning over such a dataset is an encoder f that can meaningfully represent any input i.e. it takes in x_i and returns an embedding $f(x_i)$ that captures features of the input. To do this, we train the encoder to maximize the agreement between an example and its positive pairs (i.e. the same example augmented), and minimize agreement between an example and its negative pair (i.e. a different example augmented). This is captured in the following InfoNCE loss:

$$\mathcal{L}_{cl}(V) = -\mathbb{E}_{i,j \in V}\left[\mathbb{E}_{x_1, x_2 \in A(x_i), x^- \in A(x_j)}\left[\log \frac{e^{f(x_1)^T f(x_2)}}{e^{f(x_1)^T f(x_2)} + e^{f(x_1)^T f(x^-)}}\right]\right]$$

where $A(x)$ is a set of augmented versions of example x.

We evaluate the performance of contrastive learning on downstream tasks. More specifically, we attach a linear classifier head to the learned encoder f, and train and evaluate it with labeled data.

$$g_f^l(x) = \underset{k \in [K]}{\arg\max} \left(W f(x) + b\right)_k$$

Our goal is to find a smaller subset $S \subseteq V$ of the data, such that the encoder f^S generated through contrastive learning on the subset S leads to a similar downstream linear classification performance as using the whole dataset f^V. As demonstrated in the original paper, formally, we are trying to minimize the difference in downstream linear classifier error ξ of the subset encoder f^S and the full-data encoder f^V.

$$S^* = \underset{S \subseteq V, |S| \leq r}{\arg\min} \left| \xi(g_{f_S}^l(V)) - \xi(g_{f_V}^l(V)) \right|$$

Algorithm 1 *SAS*: finding **S**ubsets that maximize the expected **A**ugmentation **S**imilarity

Input: Subset size B, proxy model f_p
Output: Subset S
$\{V_1, ..., V_K\} \leftarrow$ approximate latent classes (Sec. 4.5)
for all $V_k \in \{V_1, \cdots, V_K\}$ **do**
  for all $i, j \in V_k$ **do**
    $s_{i,j} = \langle f_p(\boldsymbol{x}_i), f_p(\boldsymbol{x}_j) \rangle$
  end for
  $S_k \leftarrow \{\}$
  $r_k \leftarrow \frac{|V_k|}{|V|} \cdot B$
  $F(S_k) = \sum_{i \in V_k \setminus S_k} \sum_{j \in S_k} s_{i,j}$
  while $|S_k| \leq r_k$ **do**
    $e \leftarrow \text{argmax}_{e \in V_k \setminus S_k} F(e|S_k)$
    $S_k \leftarrow S_k \cup \{e\}$
  end while
end for
return $S = \{S_1 \cup \cdots \cup S_K\}$

Figure 1: SAS algorithm from the original paper.

4 Method

A large part of our methodology remains the same as the original paper's, and that should be referenced for a detailed look at the SAS algorithm and its validity. For the purposes of our research, we provide our design from a practical application standpoint, which includes summaries of the original paper's approach.

4.1 Latent classification

As described in the problem formulation, our training dataset is unlabelled; however, our subset selection algorithm, SAS, requires a separation of the classes represented in the dataset. To overcome this issue, we apply the recommendation set by the original authors. We will assume that we have access to a super-small set of labeled data points, around 1% of the total dataset, and we will train a small BERT model on this labeled dataset in a supervised manner. This small model will then be used to "guess" labels or latent classes for the entire dataset, which will be used in the subset selection process.

4.2 Subset selection

In the experiment, we will compare three subset selection algorithms: baseline random, SAS, and bilingual SAS, and evaluate 20%, 40%, 60%, and 80% subsets of the dataset for learning high-quality representations.

4.2.1 SAS

The SAS algorithm from the original paper, as shown in Figure 1, needs to be modified in two ways for the text domain. The original paper used pre-trained ResNet models to map examples into an embedding space, and then calculated similarity based on Euclidean distance. For our task, we will analogously use a pre-trained BERT model to embed the text data, and calculate similarity using the cosine similarity metric. This metric is widely used in NLP, and works well with high-dimensional text embeddings regardless of length of text. The rest of the algorithm follows as before. We will use a greedy approach to continuously select the best data points i.e. the points with highest expected augmentation similarity until we have selected our subset for each latent class.

4.2.2 Bilingual SAS

Within the text domain, multilingual embeddings and representations of text in other languages are popularly used in training. Consequently, as an augmentation of the original SAS approach, we will attempt to incorporate Spanish translations of our text data into the selection process. The algorithm

doesn't change, however, instead of only using the cosine similarity of English text embeddings, we equally weight the cosine similarities of both: the English embeddings of a text (to other English embeddings) and the Spanish embeddings of a text (to other Spanish embeddings). These Spanish embeddings are generated using a pre-trained ALBETO model, which is a lightweight BERT model trained exclusively on Spanish text.

This should improve the robustness of our selection process as we intuitively select examples that are (a) best aligned to examples in their class and, (b) well-diverged from examples of other classes, across languages.

4.3 Contrastive learning augmentations

As per the problem formulation, these subsets of data are used within a contrastive self-supervised learning framework to learn feature representations or embeddings. We use the contrastive InfoNCE loss, which incorporates augmentations of data points into the learning process. This allows us to further experiment with augmentation techniques to analyze the impact of SAS within our domain.

4.3.1 Easy data augmentation

For the first form of textual data augmentation, we will consider a subset of easy data augmentation (EDA) techniques, namely synonym replacement and random swapping. As the names suggest, this form of augmentation will rely on two actions: (a) substituting a certain percentage of words with synonyms, and (b) randomly swapping a certain number of words within the text. For synonym replacement, we want to preserve the semantics of the text; however, in comparison to sentiment analysis, we can be looser on the restrictions within a text classification task. Therefore, we will randomly sample from the top-5 synonyms, as generated by finding words with similar 'word2vec' word embeddings.

4.3.2 GPT augmentation

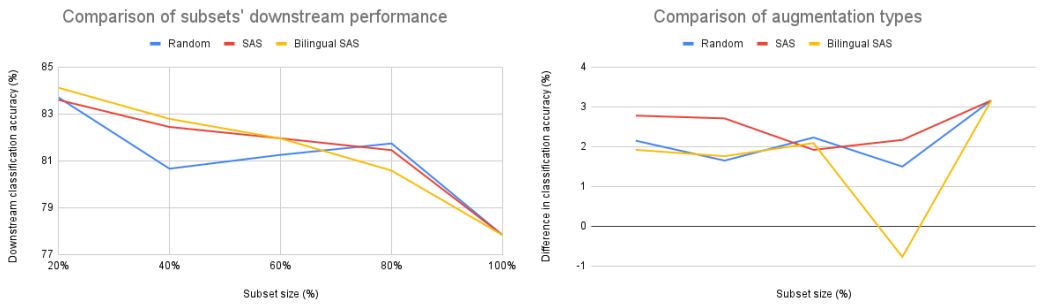
As an improvement to easy data augmentation, we will attempt to use a large-language model (GPT-4) to paraphrase the text, which in itself will be considered an augmentation of the original text. We will use zero-shot learning, and simply provide GPT-4 with the text and the prompt: "Rewrite the text in two different ways: text". These two paraphrased texts will form the two positive augmented versions of the data, as used in the contrastive learning framework.

4.4 Downstream evaluation

Finally, once we have an encoder, trained via contrastive self-supervised learning, we will evaluate it on a downstream text classification task. We will freeze the learned encoder's parameters and attach a one-layer linear head that takes in the embeddings of the last state of the encoder as input, and returns the text classification as the output. We will train this encoder plus linear head model with the labeled version of our training data, and evaluate the accuracy on the test data. Intuitively, if a subset of data improves performance at the downstream task, it must be a good representation of the dataset that leads to linearly separable embeddings for different classes.

5 Experiments

For our experiments, we use a reduced version of the AG News dataset, that contains 3000 news article texts, categorized into four topics. As evident in the problem formulation, we ignore these categorizations for the self-supervised learning process, but use them for downstream evaluation. We use BERT-small as the proxy model for English texts, and AlBETO-tiny as the proxy model for Spanish texts. The encoder function learned through contrastive learning is built on top of BERT-small embeddings of the text i.e. the text is initially represented by its BERT-small embeddings and then the representation is fine-tuned over 100 epochs through contrastive learning. We compare the encoder functions learned by using different subsets of the data by training the downstream linear classifier for 5 epochs with the labeled training data and running prediction on the AG News test data. We also compare the two forms of augmentation used within the InfoNCE loss computation. Finally, we conduct some ablation studies to understand our results.



(a) (b)

Figure 2: (a) A comparison of the downstream classification accuracy for contrastive learning encoders based on different subset selection methods and subset sizes. (b) A comparison of the downstream classification accuracy for contrastive learning encoders using different augmentation types. The vertical axis represents the absolute difference in accuracy between GPT and EDA, so a positive value implies GPT performed better than EDA.

5.1 Downstream performance of different subsets

There are a few observations and concerns regarding the downstream performance of the three subset selection methods, as shown in Figure 2a.

Firstly, on average, the subsets selected by SAS and Bilingual SAS lead to better downstream performance than the random subset. However, the impact seems to be larger with smaller subset sizes, and the difference in performance reduces as the size of the selected subset is increased. Intuitively, as we select larger subsets, the selection policies tend to converge together and the resulting subset and its downstream performance converge as well. The second result from the experiments shows that Bilingual SAS can be a potential improvement on SAS for text setting. For small subset sizes, Bilingual SAS consistently leads to better downstream performance than both random and SAS selections. The final result, and one that we will explore further in the ablation studies, is quite counter-intuitive. Using smaller subsets of the data leads to better embeddings and downstream performance than using larger subsets of data. This is starkly different from the image domain where the performance of the full dataset was considered the goal when using a subset. A possible explanation for this can be the presence of "bad" examples in the dataset. Another reason could be the inherent randomness of text data and its representations, as compared to the structured form of image data. Furthermore, this result might be a consequence of using BERT embeddings as the proxy model for subset selection as these embeddings might not directly represent a property of the input state for the text data.

We will address some of these concerns in our ablation studies.

5.2 Impact of augmentation type

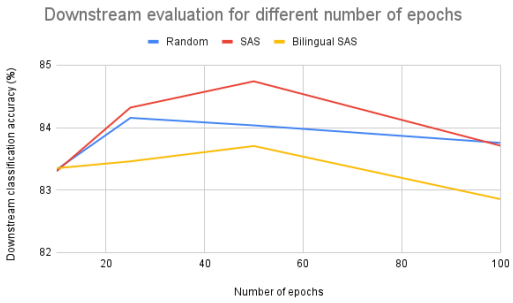
From our experiments around the different augmentation types within contrastive learning, we observe that using GPT-4 paraphrasing as augmentation leads to an average 2% increase in downstream performance compared to easy data augmentation. The plot in Figure 2b shows this exact result for different subset sizes and across subset selection methods, as there is a positive difference between GPT and EDA performance across the different setups.

5.3 Ablation studies

The results from our experiments raise concerns about our training procedure, as we are seeing the worst performance when using the full dataset for contrastive learning.

5.3.1 Analyzing the contrastive training process

We evaluate the encoder at different epochs during the contrastive learning process for our 20% subset to further analyze our experimental results. Figure 3a surprisingly shows that the training process is oscillating the downstream performance over time, which is a big concern. We could address



(a)

Type of subset selection	GPT augmentation	EDA
Random	84.37	81.75
SAS	84.45	82
Bilingual SAS	84.82	81.83

(b)

Figure 3: (a) A comparison of the downstream classification accuracy for contrastive learning encoders for different number of epochs. (b) A comparison of the downstream classification accuracy for contrastive learning encoders when using TF-IDF vectorization as the proxy model.

this by reducing the learning rate or introducing a decaying scheduler; however, it could also be a consequence of using pre-trained BERT embeddings as an initial embedding state.

5.3.2 Using a simpler proxy model

Instead of using BERT-small as a proxy model, we will use a much simpler TF-IDF representation that directly generates a vector representation from a set of texts without pre-training. In this way, the embedding and consequently the expected augmented similarity will be directly based on a property of the input without the added complexity of BERT. Once the subset has been selected based on these simpler properties, we can train on top of BERT embeddings and evaluate the encoder. These results are shown in Figure 3b and unfortunately, we do not see any significant difference. Based on these ablation studies, the experiment needs to be revisited with a more sophisticated contrastive training framework.

6 Conclusion

Our experiments have led us to several key conclusions regarding the use of SAS within the text domain. We found that SAS is not particularly effective with text data, which could be attributed to: (a) the non-structured nature of text compared to structured image data, (b) the use of highly representative BERT embeddings as the baseline, and (c) unsophisticated training framework for contrastive learning. We need to conduct further experiments to narrow down the cause for this supposed lack of translation of SAS for NLP tasks.

Another conclusion we can derive from this research is that incorporating bilingual embeddings into the SAS process enhances the robustness of selection and improves downstream outcomes, especially for small subset sizes. This can be expanded to incorporate multilingual embedding spaces to accurately identify the most representative examples across languages.

Most notably, we showed that GPT paraphrasing as augmentation consistently outperformed basic techniques like synonym replacement and random swapping within the contrastive learning framework.

In terms of future research, we want to focus on improving the contrastive learning framework such that we can build on top of highly representative embeddings such as BERT. This can be small modifications such as more sophisticated learning rate scheduling, but another direction can be to improve the efficiency of the code to allow for larger datasets and ultimately more concrete results. Another research direction would be to build our own embedding model architecture on top of a simpler text representation such as TF-IDF, instead of relying on BERT. However, since BERT is an industry standard, the focus will be on using it as the baseline for further improvement.

Contributions

- Garvit Rajkumar Pugalia: Designed the initial research topic of applying SAS to text domain, and improving SAS with bilingual embeddings. Worked on the experiments' design and implementation, including writing code for SAS, Bilingual SAS, and GPT augmentation. Actively contributed to all parts of the report and presentation, especially the Problem Formulation, Methods and Experiments.
- Joo Wan Lim: Actively contributed to all aspects of the report and presentation, especially the Introduction and Conclusion. Participated in the brainstorming process for the project's concept. Worked on the experiments' design and implementation, including writing code for augmentation for text (EDA).
- Jie Shao: Actively contributed to all aspects of the presentation and report, and participated in the brainstorming process for the project's concept.
- Zikun Gan: Actively contributed to all aspects of the presentation and report, and participated in the brainstorming process for the project's concept.

References

Jean-Michel Attendu and Jean-Philippe Corbeil. Nlu on data diets: Dynamic data subset selection for nlp classification tasks, 2023.

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. Towards representative subset selection for self-supervised speech recognition, 2023. URL https://openreview.net/forum?id=4wXotzMJ7Wo.

Salvador V. Balkus and Donghui Yan. Improving short text classification with augmented data using gpt-3. *Natural Language Engineering*, page 1–30, August 2023. ISSN 1469-8110. doi: 10.1017/s1351324923000438. URL http://dx.doi.org/10.1017/S1351324923000438.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Siddharth Joshi and Baharan Mirzasoleiman. Data-efficient contrastive self-supervised learning: Most beneficial examples for supervised learning contribute the least, 2023.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.