Difficulty Estimation in Natural Language Tasks with Action Scores

Aleksandar Angelov and **Tsegaye Misikir Tashu** and **Matias Valdenegro-Toro**
Department of Artificial Intelligence, University of Groningen.
t.m.tashu@rug.nl, m.a.valdenegro.toro@rug.nl

Abstract

This study investigates the effectiveness of the action score, a metric originally developed for computer vision tasks, in estimating sample difficulty across various natural language processing (NLP) tasks. Using transformer-based models, the action score is applied to sentiment analysis, natural language inference, and abstractive text summarization. The results demonstrate that the action score can effectively identify challenging samples in sentiment analysis and natural language inference, often capturing difficult instances that are missed by more established metrics like entropy. However, the effectiveness of the action score appears to be task-dependent, as evidenced by its performance in the abstractive text summarization task, where it exhibits a nearly linear relationship with entropy. The findings suggest that the action score can provide valuable insights into the characteristics of challenging samples in NLP tasks, particularly in classification settings. However, its application should be carefully considered in the context of each specific task and in light of emerging research on the potential value of hard samples in machine learning.

1 Introduction

While contemporary artificial intelligence (AI) algorithms can be successfully applied to a range of tasks, they need vast amounts of data to be trained on. One of the inevitable problems with such AI systems is the biases (Mehrabi et al., 2021) they often exhibit. Large datasets can contain inherent biases that get amplified when used to train AI models. They often stem from skewed or unrepresentative training data and can result in models misinterpreting or struggling with certain samples. Another substantial issue when dealing with large amounts of data is the expense and sometimes the inability to fact-check the correctness of every data sample (Sukhbaatar and Fergus, 2014), which amplifies the tendency of large-scale datasets to have a significant portion of their examples wrongly labeled. One way of observing both biased and incongruous samples is by analyzing whether they comply with the optimization dynamics. Therefore, it is crucial to have a systematized, robust, and model-agnostic way to pinpoint such samples and observe how exactly they influence the model's performance.

Hard Sample, AS = 33.06, H = 0.01

i feel that he was being overshadowed by the supporting characters

Easy Sample, AS = 0.01, H = 0.005

i feel reassured that if something happened to me my guests would be able to easily get the help they need

Figure 1: Example of easy and hard samples for Sentiment Analysis, together with their Action Score (AS) and Entropy (H). The AS reveals different prediction information compared to Entropy.

In this regard, the metric that this study will explore is called action score, which can be categorized as part of the tools available for understanding model dynamics through the lens of individual samples. This metric has been extensively studied and applied across various computer vision tasks, demonstrating its relevance in assessing model behavior (Arriaga et al., 2023). However, its potential usefulness in the domain of natural language processing (NLP) has yet to be explored. This study aims to fill this gap by applying the metric to sentiment analysis, natural language inference and abstractive text summarization. Tables 2a, 2b and 2c show examples of difficult and easy samples (high and low action scores accordingly) from each of the tasks.

The main research question in this paper is: *Can the action score measure difficulty in natural language processing tasks?*

	Text	Label	Action Score
Hard	i as representative of everything thats wrong with corporate america and feel that sending him to washington is a ludicrous idea	surprise	29.0019
Easy	i feel reassured when i listen to waldmans songs	joy	0.0048

(a) Sentiment Analysis

	Premise	Hypothesis	Label	Action Score
Hard	I read everything I could get my hands on.	There was nothing I couldn't get my hands on.	neutral	41.1284
Easy	She remembered poems that she had learned when she was in high school.	She didn't remember any poems.	contra	0.0166

(b) Natural Language Inference

	Dialogue	Label Summary	Action Score
Hard	Sophie: Whats for dinner mom?Olivia: Tacos and burritos Sophie: wowwww! my favorite please keep it ready will be home in 20 mins Olivia: all is ready dear!!	Sophie is coming home in 20 minutes for the dinner Olivia, her mother, prepared.	19.8895
Easy	Mattie: Will you call me when dad is at home? Ross: Sure Mattie: ty :*	Ross will call Mattie when dad is at home.	0.5993

(c) Abstractive Text Summarization

Figure 2: Examples of action scores. Hard (high action score) and easy (low action score) samples from each task. Each data sample is accompanied by its corresponding target label / summary and its action score.

The existing metrics are either model or task-specific. Furthermore, the majority of those metrics are too computationally expensive and/or require model architecture modifications. The action score is calculated as the accumulated loss over all epochs for each individual sample. The metric utilizes a single parameter that can be relatively easily obtained. In essence, the main contribution of the action score lies in its effective application to a variety of different architectures with minimal additional modifications. The contributions of this paper are the experimental validation that the action score measures sample difficulty in three NLP tasks: Sentiment analysis, Natural language inference, and abstractive text summarization.

This study hypothesizes that the action score, originally developed for computer vision tasks, can be effectively applied to NLP tasks to measure the difficulty of individual samples. It is expected that samples with higher action scores will correspond to more challenging linguistic inputs, such as complex sentence structures, rare words, or ambiguous meanings, while lower action scores will be associated with simpler, more straightforward language samples. The effectiveness of the action score in NLP will be evaluated by comparing it to established difficulty metrics like the predicted entropy and through qualitative analysis of high and low-scoring samples.

2 State of the Art

An important aspect of model evaluation that is often overlooked is the varying difficulty of samples within a dataset. An interesting work (Swayamdipta et al., 2020) on dataset cartography, identifying three distinct regions in datasets: easy-to-learn samples, hard-to-learn samples, and am-

biguous samples. This categorization reveals that the difficulty of samples is not binary but exists on a continuum. Traditional evaluation metrics often fail to capture these nuances, potentially leading to an incomplete understanding of model performance.

An intriguing work (Pleiss et al., 2020) has utilized the insights we can obtain from the model's training dynamics - in a broad sense, this is the model's behavior during the training process - as a potential avenue for developing more generalizable metrics. The Area Under the Margin (AUM) metric (Pleiss et al., 2020) shows promise in identifying mislabeled instances in classification datasets, but its effectiveness in filtering NLP datasets has been questioned (Talukdar et al., 2021). Although the AUM metric can successfully find mislabeled samples, it also removes a significant amount of correctly labeled samples, which results in the loss of a large amount of relevant information.

Interpreting model behavior through the lens of training dynamics allows us to gain insights into both the nature of the dataset and the model's learning process. Samples that are consistently classified correctly with high confidence throughout training likely represent "easy" instances, while samples where the model's predictions fluctuate greatly may represent ambiguous or challenging instances. Samples that are consistently misclassified, even late in training, may represent very difficult instances or potentially mislabeled data.

Several difficulty metrics have been studied for particular NLP tasks. (Bommasani and Cardie, 2020) performed a large-scale evaluation of summarization datasets, introducing 5 intrinsic metrics and applying them to 10 popular datasets. Their findings highlight that data usage in recent summarization research is sometimes inconsistent with the underlying properties of the datasets employed. They also discovered that their metrics can serve as inexpensive heuristics for detecting generically low-quality examples.

In the context of text classification, a study (Mujumdar et al., 2023) identifies difficult samples by analyzing data inputs in the semantic embedding space. The method proves to be an effective way to find difficult samples in 13 datasets. By removing them, trained models achieve better F1 scores (up to 9%). Despite these efforts, the AI research community has yet to develop a truly universal metric that can be applied without major model modifications, that is model- and task-agnostic and that does not add a significant computational overhead.

3 Action Score for NLP Tasks

The action score (Arriaga et al., 2023) is a novel metric designed to quantify the difficulty of individual samples in machine learning tasks. It is based on the principle that samples that do not conform to the optimization dynamics of a model can be considered unnatural or difficult. The action score is calculated by accumulating the loss of each sample overall validation (or training) epochs, resulting in a single scalar value that represents the sample's difficulty. This approach is model-agnostic and can be applied to a wide range of tasks without requiring modifications to the underlying model architecture. The action score is defined as

$$\mathcal{A}(x) = \sum_{n=0}^{N} \mathcal{L}(y, m(x, \theta_n)\Delta n) \tag{1}$$

where $\mathcal{L}$ is the loss function, m is the model, θ are the model parameters at epoch n, and Δn is the training time step (in our case we use an epoch as an optimization step and the step itself is one). Higher action scores indicate samples that were more challenging for the model to learn, while lower scores suggest easier samples. This metric provides a unique perspective on the characteristics of the dataset, the biases of the model, and the potential mislabeled samples by offering valuable insights for improving both datasets and models in various machine learning applications.

For tasks involving sequence outputs, as many NLP tasks like summarization are, we compute the loss for each element in the output and take the average over sequence elements to obtain a single action score that is accumulated over epochs for each sample. Using the average instead of plain sum allows one to obtain an action score that is partially invariant to the output sequence length.

4 Experimental Setup

To investigate the effectiveness of the action score in estimating sample difficulty in natural language processing (NLP) tasks, a systematic methodology was employed, involving the following steps:

1. **Selection of representative NLP tasks:** Sentiment analysis, natural language inference, and abstractive text summarization were chosen as the target tasks for this study. These tasks cover a range of applications and vary

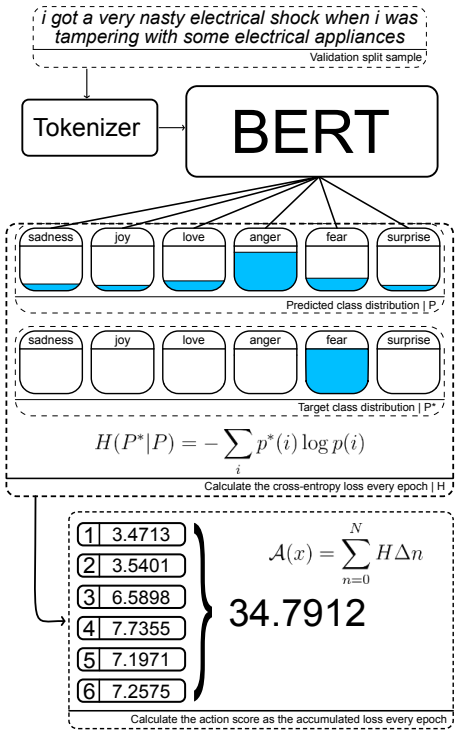


Figure 3: Conceptual description of Action Score computation for a sequential task, with an example of sentiment analysis.

in complexity, allowing for a comprehensive evaluation of the action score's performance.

2. **Implementation of state-of-the-art models:** For each selected task, a state-of-the-art transformer-based model was implemented using the Hugging Face *Transformers* library (Wolf et al., 2020). The models were fine-tuned on task-specific datasets.

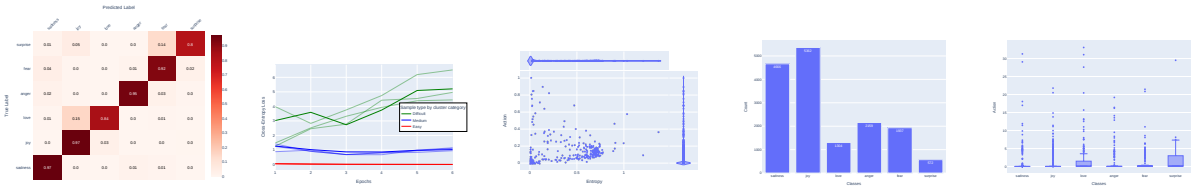
3. **Calculation of the action score and entropy:** During the fine-tuning process, the action score for each validation sample was calculated by accumulating the loss values across all epochs. Additionally, the predicted entropy for each sample was computed to compare with the action score.

The tasks this study will explore were implemented using the tools provided by the *transformers* library (Wolf et al., 2020) developed and maintained by the Hugging Face team. *Transformers* is an open-source library that consists of pre-trained cutting-edge models readily available in a unified API. The Hugging Face platform also hosts a large collection of curated datasets that are easily accessible and integrated into the transformers workflow. The main reasons behind choosing to work with pre-trained large language models (LLMs) are their computational efficiency (leveraging transfer learning and adapting them to different tasks with minimal fine-tuning) and their SOTA performance in a range of NLP tasks that would allow for a rigorous test of the action score's ability to measure difficulty in advanced scenarios.

Sentiment analysis and natural language inference are both text classification tasks, where the goal is to assign a pre-defined label to an input text. For these tasks, we employed the base BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), which contains 110 million parameters, offering an effective balance between performance and computational efficiency. BERT's key innovation lies in its bidirectional nature, enabling it to consider the context from both the left and right sides of each word in a sentence. To adapt BERT for our specific classification tasks, the base model was augmented with a task-specific classification head, incorporating the appropriate number of output labels for each task.

Sentiment Analysis. The sentiment analysis model was fine-tuned on the aforementioned version of BERT using the emotion dataset (Saravia et al., 2018). A total of 24 000 samples were utilized, with 20 000 assigned for training, 2 000 for validation, and 2 000 for testing. A single sample consisted of a text with a mean length of 20 tokens and one of 6 labels (sadness, joy, love, anger, fear, surprise) denoting the sentiment of the text. The model was trained for 6 epochs with a learning rate of 5×10^{-5}. After each training epoch, the validation samples were fed one by one (batch size of 1) to the model and their respective losses and predicted entropies were kept track of. The entropy metric for each sample was obtained during the final validation cycle (it is a single value, not an accumulated number as the action score).

Natural Language Inference. The model for the natural language inference (NLI) task was fine-tuned again on the same variation of BERT using the multi-genre natural language inference corpus (MNLI) (Williams et al., 2018) task from the GLUE benchmark (Wang et al., 2019). Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral). The premise and hypothesis are concatenated with a special token between them. The model was similarly trained for 6 epochs with a learning rate of 5×10^{-5}. Ob-



(a) Confusion matrix (b) K-means Clustering (c) Action vs Entropy (d) Class Distribution (e) Per-class AS

Figure 4: Sentiment analysis task results. (a) normalized confusion matrix on the validation split after the 6th (final) evaluation epoch. (b) clustering of the per-sample loss curves on the emotion dataset on BERT. Each cluster center is shown with its top three closest curves which can be inspected in greater detail in Table 4. On the x-axis are the epochs and on the y-axis is the cross-entropy loss. (c) normalized action score against the entropy for the individual sample points. (d) training split class distribution. (e) per class action score distribution of the validation split.

taining the action score and the predicted entropy are analogous to the sentiment analysis task.

Abstractive Text Summarization. Unlike the text classification tasks, text summarization requires a different approach in terms of a model architecture. The Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020), built on a sequence-to-sequence architecture, is particularly well-suited for tasks involving text generation. This study will utilize the base version of T5 and the SAMSum dataset. The dataset contains about 16,000 messenger-like conversations (created and written down by linguists fluent in English, who were asked to mimic their daily messaging habits) with summaries (a concise brief of the conversation's content written in the third person).

The acquisition of the loss value for each validation sample is similar to the previous implementations. However, the entropy metric was a bit more difficult to obtain. Since there are multiple tokens that are predicted, each has its own uncertainty value. Combined with the fact that the predicted summaries have different lengths, it is impossible to simply add up the individual entropy values for a given prediction as the metric will be quite hard to interpret. A solution to this problem turned out to be taking a simple average over all the tokens in a prediction.

5 Experimental Results

5.1 Sentiment Analysis

The sentiment analysis model achieved a test accuracy of 93%. Although there are better models out there, for the purpose of this study, it is not required to aim for the best. Table 1 shows the five most and least difficult samples according to the action score metric. It is arguable whether the texts are mislabeled as often there could be multiple sentiments in a single piece of text and our model is not configured to predict multiple labels. In figure 4a it can be observed that the model mostly misclassified surprise as fear and love as joy. Surprise and fear are both characterized by heightened arousal and can be triggered by unexpected events. Similarly, love and joy are positive emotions that often co-occur, leading to potential confusion for the model. However, upon reviewing Table 1 more thoroughly, it appears that the highest action score samples are not misclassifications of the surprise-fear and love-joy pairs. The most difficult samples span a range of different emotion pairs, such as love-sadness, sadness-anger, and love-fear. This observation implies that the action score captures a more nuanced aspect of sample difficulty that goes beyond the confusion between specific emotion pairs.

One can look at the classification head's probability distribution and expect to see rather conflicting predictions. This would reflect the model's uncertainty and therefore the entropy metric, but as it can be seen the model is generally certain in its predictions. Nevertheless, the samples with a high action score are hard to perceive semantically and tend to be tricky even for human evaluators. The least difficult samples happen to be classified as joy. One reason for that could be their predominant occurrence in the training data as it can be observed from figure 4d. Another interesting characteristic of the low action score samples is their syntactic structure. Some start with "I feel ..." followed by a particular feeling. Such patterns are straightforward and quite easy for the model to learn. Others, like the 5th example, tend to be more elaborate and require a prompt understanding of the conveyed sentiment. As already mentioned, the class distribution is quite uneven and when comparing it to figure 4e we can see there is a negative correlation

between the number of classes and the mean action score of every class, which suggests that some samples might not be difficult usually, but has failed to generalize over them due to the limited amount in the training data.

Figure 4b depicts the clustered individual sample losses. The three clusters can be interpreted as difficult (green), medium (blue) and easy samples (red). The difficult samples are progressively less conforming with the model dynamics, which in turn suggests that they are either incongruous (mislabeled, outliers) or genuinely hard for the model to learn. Each cluster center is shown alongside the three of its closest curves, which can be observed in Appendix B.1, Table 4. Taking a closer look at the difficult samples we can see that the loss has progressively increased each epoch and the model at no point managed to predict the target label correctly. Medium-difficulty samples are also misclassified, but unlike the difficult samples, they have been correctly predicted during the 3rd-4th epoch. This suggests that the model might be overfitting and a better training strategy can potentially make the model generalize more successfully.

Figure 4c shows the joint distribution of entropy vs action score. It is evident that the two metrics measure different properties of a prediction. Although the majority of the samples are concentrated around the origin (near zero entropy and action), there are a dozen or so samples scoring high on action, but near 0 on entropy. The first couple of examples in Table 1 are a nice illustration of how the action score can capture irregularities that would be missed from the predicted entropy metric.

5.2 Natural Language Inference

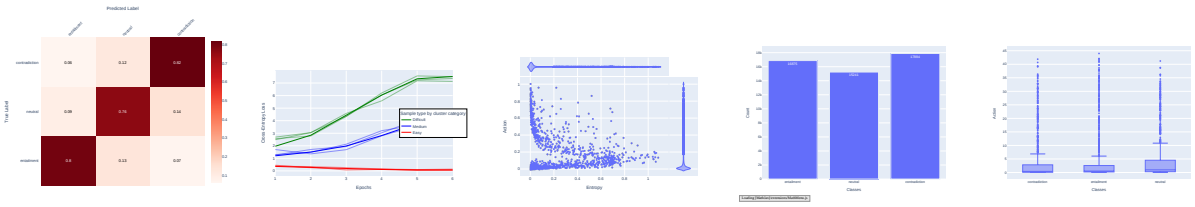
As previously mentioned, NLI falls under the broader umbrella of text classification tasks. One might question the necessity of including a second task of this nature, and such an inquiry would be valid. While NLI shares similarities with sentiment analysis, it presents a substantially higher level of complexity. NLI demands a more nuanced understanding of input characteristics, requiring the model to comprehend and reason about the relationship between two separate text segments. This increased complexity makes NLI an excellent candidate for evaluating the action score's ability to differentiate between tasks of varying difficulty within the same general category. Another important feature is that there is no room for overlapping labels as in sentiment analysis. The premise can be entailing, contradicting the hypothesis or being neutral to it.

The NLI model achieved an accuracy of 79%, although having only three categories. Table 2 presents the top four samples with the highest and lowest action scores for the natural language inference task. A deeper analysis reveals several characteristics typical for difficult samples in the dataset. The most difficult sample is an example of an atypical sentence structure namely that the subordinate clause ("Even though we receive operating funds from the state") being before the main clause ("there are a myriad of additional expenses to be met"). This makes the essence of the premise harder to understand and therefore more likely to be mistaken. The second most difficult sample exemplifies that phrasal verbs ("take out" - "shoot down") and abbreviations ("Vice President" - "VP") can also be hard for the model to make sense of. Interestingly, the third most difficult sample appears to be nonsensical and is likely mislabeled. The easy samples, on the other end, follow a systematic structure where the premise contains a statement ("She remembered...") and the hypothesis straightforwardly states the opposite ("She didn't remember...").

The training split class distribution observed in figure 5d is more evenly spread. This is also reflected in the recall per class in figure 5a and the per class action score distributions in figure 5e. The correlation between the class distribution, the recall and the action score distribution is also evident here and once more reiterates the impact of imbalanced data.

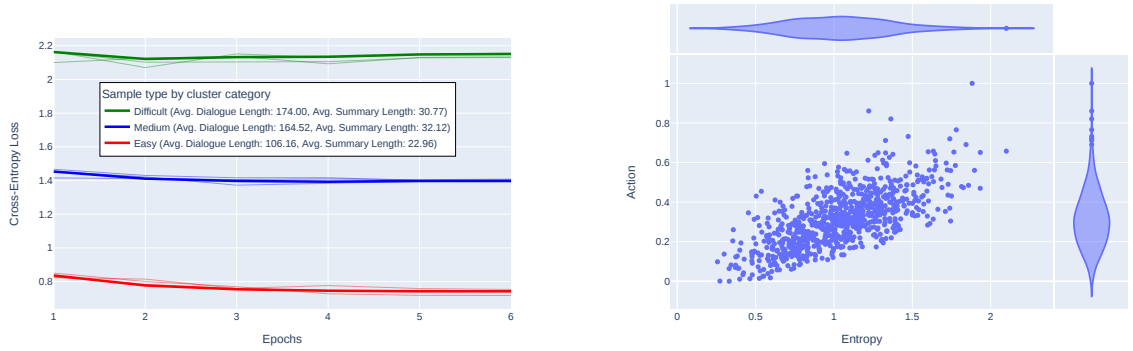
Figure 5b shows the clustered individual sample losses. The three clusters can be interpreted as very difficult (green), difficult (blue) and easy samples (red). It is not directly evident from the figure that the clusters are not of equal size (the red cluster represents the correctly predicted samples, the majority of the evaluation split). Therefore, as the model converges to some generalizable state, the samples that are not conforming to its dynamics tend to stand out. In our case, two distinct groups of difficult samples are formed. Upon a closer look at the three closest curves of each cluster center in appendix B.2, Table 5. We can see pretty much the same properties discussed previously in figure 2 responsible for the action score values. There is no qualitative difference between the difficult and very difficult samples.

Figure 5c represents the joint distribution of the



(a) Confusion Matrix (b) K-means Clustering (c) Action vs Entropy (d) Class Distribution (e) Per-class AS

Figure 5: Natural language inference task results. (a) confusion matrix on the validation split after the 6th (final) evaluation epoch. (b) clustering of the per-sample loss curves on the emotion dataset on BERT. Each cluster center is shown with its top three closest curves, with difficult samples having increasing loss, in contrast to easy samples with decreasing loss. (c) normalized action score against the entropy for the individual sample points, showing how the action score behaves differently than entropy. (d) training split class distribution. (e) per class action score distribution of the validation split.



(a) K-means Clustering (b) Action vs Entropy

Figure 6: Abstractive Text Summarization task results. (a) clustering of the per-sample loss curves on the SAMSum on T5. Each cluster center is shown with its top three closest curves, these curves clearly show three difficulty types (easy, medium, hard). In the legend are specified the average token lengths of the dialogues and the target summaries of each cluster (b) normalized action score against the entropy for the individual sample points, showing how action and model uncertainty are highly correlated.

action score versus the entropy. Once again, most of the data points are saturated near the origin. Here, there is substantially a larger portion of samples with a high entropy score. Nevertheless, it is evident that a significant amount of samples tends to have low entropies and high action scores, which is in support of the claim that the action score captures different properties than entropy.

5.3 Abstractive Text Summarization

The abstractive text summarization task presents a unique challenge compared to the previous classification tasks. It requires the model not only to understand the input text, but also to generate a concise summary that may use different words and phrases than those in the original text. This complexity makes it an ideal candidate for evaluating the action score's effectiveness in more advanced NLP tasks.

The samples that yielded the highest action score 3 tend to have some distinct properties. First, the dialogues are quite lengthy, which in turn can make the predicted summaries hard to match perfectly with the target summary. The model also fails to properly understand and produce phrases such as "break wind" found in the target summary of the first sample. Other samples are just hard either because of the way of interaction or the language used. Worth noting is the fact these samples also have a high entropy.

The samples with low action scores tend to be short and to the point. The target summary pretty much uses the same vocabulary as the dialogue. The easy samples also have a low entropy score which suggests a correlation between the two.

Figure 6a depicts the clustered individual sample losses for the summarization task. The three dis-

tinct clusters can be categorized as difficult (green), medium (blue) and easy (red) to learn samples. Looking closely we can observe the steady increase of the green line and the steady decrease of the red line. We can conclude that the model is learning, but it is very slow compared to the other tasks.

The most striking observation about the model and the task emerges from Figure 6b. Unlike the previous scatter plots, the relationship between the action score and entropy appears nearly linear for this task. This finding breaks the trend observed in earlier tasks where the action score measured properties distinct from entropy. While the action score can still serve as a human auditing tool, its unique value in this specific task may be less pronounced.

5.4 Discussion

The results of this study demonstrate the potential of the action score as a metric for estimating sample difficulty in various natural language processing tasks. In sentiment analysis and natural language inference, the action score effectively identified challenging samples that were often misclassified by the model, despite having low entropy scores.

The abstractive text summarization task, however, presented a different scenario. The nearly linear relationship between the action score and entropy in this task indicates that the action score may not provide as much additional value in identifying difficult samples compared to the classification tasks. It is essential to consider the specific characteristics of each NLP task when evaluating the effectiveness of difficulty estimation metrics.

It is worth noting that recent research (Wu et al., 2022) has proposed alternative approaches to dealing with hard samples in machine learning tasks. DiscrimLoss is a universal loss metric designed to discriminate between hard samples and incorrect samples. This metric suggests that excluding all hard or incorrect samples, as some popular metrics do, can actually degrade the model's performance, as these challenging samples can contribute to the model's generalization ability. The findings of the DiscrimLoss study raise important considerations for the application of the action score in NLP tasks. While the action score can effectively identify difficult samples, it is crucial to carefully evaluate whether removing these samples from the training data is the most appropriate course of action. In some cases, retaining hard samples may actually benefit the model's robustness and generalization capabilities.

6 Conclusions and Future Work

This study has explored the effectiveness of the action score, a metric originally developed for computer vision tasks, in estimating sample difficulty across various natural language processing tasks. The results demonstrate that the action score can provide valuable insights into the characteristics of challenging samples in sentiment analysis and natural language inference, often identifying difficult instances that are missed by other metrics like entropy.

However, the effectiveness of the action score appears to be task-dependent, as evidenced by its performance in the abstractive text summarization task. This finding underscores the importance of considering the unique properties of each NLP task when applying difficulty estimation metrics.

7 Limitations

This research is limited by our selection of evaluation tasks (Sentiment Analysis, Natural Language Inference, and Abstractive Text Summarization), while these tasks are a good representation of commonly used NLP tasks, it is possible that the action score does not measure difficulty in other tasks or behaves differently.

We only evaluated a handful of language models, and we leave detailed comparisons across different models for future work. Our aim is to show that the action score is usable for natural language tasks.

There is no agreement in the literature on how to divide difficulty ratings. In this paper, we use easy/difficult or easy/medium/hard, but these difficulty labels are subjective and motivated by clustering of the action score. In practice difficulty ratings can be divided differently, depending on the task and desired difficulty granularity.

8 Ethics Statement

There are no guarantees on performance and discrimination of different difficulty ratings when using action scores, its performance reflects model biases, so careful data analysis should be performed when assessing and selecting models. We expect that difficulty estimation in natural language tasks can shine a light on different kinds of model and data biases and improve our understanding of how (large) language models work.

References

Octavio Arriaga, Sebastian Palacio, and Matias Valdenegro-Toro. 2023. Difficulty estimation with action scores for computer vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 245–253.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Shashank Mujumdar, Stuti Mehta, Hima Patel, and Suman Mitra. 2023. Identifying semantically difficult samples to improve text classification. *Preprint*, arXiv:2302.06155.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. *Preprint*, arXiv:2001.10528.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Sainbayar Sukhbaatar and Rob Fergus. 2014. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3):4.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Arka Talukdar, Monika Dagar, Prachi Gupta, and Varun Menon. 2021. Training dynamic based data filtering may not work for nlp datasets. *Preprint*, arXiv:2109.09191.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tingting Wu, Xiao Ding, Hao Zhang, Jinglong Gao, Li Du, Bing Qin, and Ting Liu. 2022. Discrimloss: A universal loss for hard samples and incorrect samples discrimination. *Preprint*, arXiv:2208.09884.

A Most and Least Difficult Samples per Task

A.1 Sentiment Analysis

	Text	Label		Metrics	
		Target	Predicted	Action	Entropy
Difficult	i feel that he was being overshadowed by the supporting characters	love	sadness	33.0632	0.0108
	i hate being the party girl because i feel like such a hypocrite because i always hated them	sadness	anger	31.2775	0.0201
	i feel badly about reneging on my commitment to bring donuts to the faithful at holy family catholic church in columbus ohio	love	fear	31.0677	0.5154
	i as representative of everything thats wrong with corporate america and feel that sending him to washington is a ludicrous idea	surprise	sadness	29.5054	0.1263
	im sure much of the advantage is psychological the feeling ive out clevered the competition who are now hopelessly burdened with their big chainring jump	sadness	joy	29.0591	0.1285
Easy	i love to hear from my friends so feel free to leave me a comment	joy	joy	0.0094	0.0050
	i also reply to most comments so please feel free to share your thoughts and let s talk	joy	joy	0.0094	0.0050
	i feel reassured when i listen to waldmans songs	joy	joy	0.0097	0.0048
	i feel reassured that if something happened to me my guests would be able to easily get the help they need	joy	joy	0.0097	0.0048
	i constantly worry about their fight against nature as they push the limits of their inner bodies for the determination of their outer existence but i somehow feel reassured	joy	joy	0.0097	0.0053

Table 1: The five most and least difficult samples in the evaluation substrata of the emotion dataset using BERT fine-tuned for 6 epochs. Each text sample is accompanied on the right with its corresponding target label, predicted label and entropy on the 6th (final) evaluation epoch, and action score.

A.2 Natural Language Inference

	Text		Label		Metrics	
	Premise	**Hypothesis**	**Target**	**Predicted**	**Action**	**Entropy**
Difficult	Even though we receive operating funds from the state, there are a myriad of additional expenses to be met, such as welding equipment for sculpture, pottery wheels for ceramics, and computers for graphics.	The state won't fund welding equipment, pottery wheels or computers.	entail	contra	43.9887	0.0025
	Believing they had only a minute or two, the Vice President again communicated the authorization to engage or take out the aircraft.	The VP thought they only had a minute or two to make the decision, so he told them to shoot any plane down immeditaely.	entail	neutral	42.0983	0.0700
	\\ How \\, how come?	How did you do it, and how come you even wanted to?	entail	neutral	42.0540	0.2205
	Given the predominant share of workers in assembly, organization of work in the sewing room has been the central focus of management attention.	Management attention has been focused mostly on providing good benefits for workers.	contra	neutral	41.8857	0.0029
Easy	She remembered poems that she had learned when she was in high school.	She didn't remember any poems.	contra	contra	0.0166	0.0025
	Consider the following problematic situations that parents recently raised with	Parents didn't bring up any problematic situations.	contra	contra	0.0193	0.0023
	You will soon receive information from the Alumni Association with the details.	The Alumni Association cannot contact you with any information.	contra	contra	0.0195	0.0025
	Here are just some of the services your gift can provide.	Your gift can't provide any service, it's completely useless.	contra	contra	0.0198	0.0024

Table 2: The four most and least difficult samples in the evaluation substrata of the MNLI dataset using BERT fine-tuned for 6 epochs. Each premise-hypotheses sample pair is accompanied on the right with its corresponding target label, predicted label and entropy on the 6th (final) evaluation epoch, and action score.

A.3 Abstractive Text Summarization

	Dialogue	Summary		Metrics	
		Target	Predicted	Action	Entropy
Difficult	Mark: Hey dude, what's up? Harry: Nothing much buddy. How's everything with you? Mark: All good. Yesterday I went to a 7-star Hotel restaurant. Harry: Wow, that's amazing buddy. I missed it. Mark: It's not amazing. Harry: Why, what happened? Mark: When I was there, I really needed to pass gas. Harry: And? Mark: The music was really loud, so I did it. Harry: And? Mark: I realized I was listening to my iPod. Harry: hahaha hahaha	Mark had to break wind, while being in a fancy restaurant. He sought to do it silently. He failed.	Mark went to a 7-star Hotel restaurant yesterday. He missed it because the music was loud and he was listening to his iPod.	24.1177	1.8819
	Paula: hey, it was a great time I spent there, really nice experience. Paula: Do you have, by chance, the names of the bands? Paula: Not that important though Tim: actually these were not bands but just single people, but later I can give you their Facebook profiles. Tim: really nice meeting you too, let's jam more often! Paula: Definitely! Paula: once I'm back in town, I'll let you know;-) Tim: always welcome Paula: thanks, we're in touch	Paula enjoyed jamming with Tim. Tim will send her Facebook profiles of the other people that played.	Paula spent a lot of time there. Tim will give Paula the names of the bands. Paula will let Tim know when she's back in town.	20.8349	1.2220
	Sophie: Whats for dinner mom? Olivia: Tacos and burritos Sophie: wowwww! my favorite please keep it ready will be home in 20 mins Olivia: all is ready dear!!	Sophie is coming home in 20 minutes for the dinner Olivia, her mother, prepared.	Sophie and Olivia are going to have tacos and burritos for dinner. Sophie's favorite will be home in 20 minutes.	19.8895	1.3630
Easy	Frank: i owe you one btw! Judy: haha, you owe me two Frank: okay then, two dates it is. haha Judy: lol	Frank owes Judy two dates.	Frank owes Judy two dates.	0.5828	0.3312
	Mattie: Will you call me when dad is at home? Ross: Sure Mattie: ty :*	Ross will call Mattie when dad is at home.	Ross will call Mattie when dad is at home.	0.5993	0.2718
	Adam: Do you know where Mary is? Lizzy: She went to library with Carl. Adam: Oh, I see... Adam: Thanks!	Mary went to the library with Carl.	Mary went to the library with Carl.	0.8321	0.3979

Table 3: The top three most and least difficult samples in the evaluation substrata of the SAMsum dataset using T5 fine-tuned for 6 epochs. Each dialogue pair is accompanied on the right with its corresponding target summary, predicted summary, action score, and entropy.

B Closest Loss Curves for each Cluster Center per Task

B.1 Sentiment Analysis

	Text	Loss and predicted label per epoch						Target label
		Nº**1**	Nº**2**	Nº**3**	Nº**4**	Nº**5**	Nº**6**	
Difficult	one afternoon	1.4842 sadness	2.5214 sadness	3.3205 anger	3.9363 anger	4.3904 sadness	4.4469 sadness	fear
	i am afraid of my emotions because certain people cause me to feel assaulted by feeling and i just get hammered by their waves as if i am an tempestuous ocean raging and only god knows why	1.2642 fear	2.4620 fear	2.7783 fear	4.4261 fear	4.5439 fear	4.9665 fear	sadness
	someone acting stupid in public	3.9812 sadness	2.8304 sadness	3.7569 sadness	4.7473 sadness	6.1888 sadness	6.5105 sadness	anger
Medium	i started out feeling amazing	1.2512 surprise	0.9103 surprise	0.6835 joy	0.8053 surprise	0.9422 surprise	1.1174 surprise	joy
	i do feels amazing and is an investment for something greater	1.4135 surprise	0.9069 surprise	0.6823 joy	0.8062 surprise	0.8339 surprise	0.9518 surprise	joy
	i went from feeling helpless to powerful	0.8828 fear	0.9912 fear	0.6817 sadness	0.6757 sadness	0.9881 fear	1.1522 fear	sadness
Easy	i cant seem to get passed feeling stunned	0.0571 surprise	0.0236 surprise	0.0112 surprise	0.0050 surprise	0.0036 surprise	0.0036 surprise	surprise
	i feel amazed and surprised when the exact question i am trying to ask	0.0592 surprise	0.0248 surprise	0.0120 surprise	0.0054 surprise	0.0035 surprise	0.0034 surprise	surprise
	i feel that im most amazed still by silent knight which is an instrumental song ala hizaki	0.0550 surprise	0.0214 surprise	0.0079 surprise	0.0047 surprise	0.0040 surprise	0.0040 surprise	surprise

Table 4: The samples corresponding to the top three closest curves to each cluster center (Figure 4b) for the sentiment analysis task. Each sample is accompanied with its corresponding loss and predicted label every epoch and its target label.

B.2 Natural Language Inference

	Text		Loss and predicted label per epoch						Target label
	Premise	**Hypothesis**	№**1**	№**2**	№**3**	№**4**	№**5**	№**6**	
Very difficult	Um, Christmas is coming up pretty soon huh?	Did you get a present for the Christmas party yet?	1.8059 neutral	2.8519 neutral	4.2737 neutral	6.1842 neutral	7.3844 neutral	7.6621 neutral	contra
	And that keeps me, as an adult, one, remembering to pray like a child, and to maintain some of the innocence, which is difficult, of a child.	As an adult I know I can never have the innocence of a child.	2.3766 contra	2.7429 contra	4.3114 contra	6.0187 contra	7.4387 contra	7.4049 contra	neutral
	(I have often wondered why the publishers did not have the nerve to call themselves F**k and Wagnalls.	I thought Fuck and Wagnalls was a good name for the publishers.	2.1916 contra	2.5666 contra	4.4929 contra	6.2313 contra	7.6666 contra	7.7242 contra	neutral
Difficult	In the apt description of one witness, It drops below the radar screen and it's just continually hovering in your imagination; you don't know where it is or what happens to it.	It is hard for one to realize what just happened.	0.8152 neutral	1.6826 neutral	1.5516 neutral	3.1601 neutral	3.0363 neutral	3.4533 neutral	entail
	They want to regain their parents' warmth and approval as quickly as possible.	They really like their parents, so they want to be approved.	1.6376 entail	1.4896 entail	1.4248 entail	3.2634 entail	3.8970 entail	3.4582 entail	neutral
	So, we stayed there.	We were not moving anytime soon.	0.8648 contra	0.9151 neutral	1.5693 contra	3.2045 contra	3.9115 contra	3.4252 contra	neutral
Easy	Five years.	Its been five years since I have been here.	0.3803 neutral	0.1914 neutral	0.1739 neutral	0.1433 neutral	0.0990 neutral	0.0518 neutral	neutral
	I enjoy sharing these small victories with you through my letters.	I like telling you about good stuff.	0.4092 entail	0.2695 entail	0.0974 entail	0.1393 entail	0.1016 entail	0.1255 entail	entail
	Note that this system poses production questions for BMW similar to those faced by apparel suppliers.	BMW has new questions about production that are being faced by the suppliers of apparel.	0.3858 entail	0.2208 entail	0.1140 entail	0.0330 entail	0.0204 entail	0.0220 entail	entail

Table 5: The samples corresponding to the top three closest curves to each cluster center are shown here. Each sample is accompanied with its corresponding loss and predicted label every epoch and its target label.