ESTIMATING SHAPLEY VALUES OF TRAINING UTTERANCES FOR AUTOMATIC SPEECH RECOGNITION MODELS

Ali Raza Syed[1], *Michael I. Mandel*[2]

[1]The Graduate Center, CUNY
[2]Brooklyn College, CUNY

ABSTRACT

Data Valuation in machine learning is concerned with quantifying the relative contribution of a training example to a model's performance. Quantifying the importance of training examples is useful for identifying high and low quality data to curate training datasets and for address data quality issues. Shapley values have gained traction in machine learning for curating training data and identifying data quality issues. While computing the Shapley values of training examples is computationally prohibitive, approximation methods have been used successfully for classification models in computer vision tasks. We investigate data valuation for Automatic Speech Recognition models which perform a structured prediction task and propose a method for estimating Shapley values for these models. We show that a proxy model can be learned for the acoustic model component of an end-to-end ASR and used to estimate Shapley values for acoustic frames. We present a method for using the proxy acoustic model to estimate Shapley values for variable length utterances and demonstrate that the Shapley values provide a signal of example quality.

Index Terms— automatic speech recognition, acoustic model, shapley value, data valuation, data-centric machine learning, structured prediction

1. INTRODUCTION

Training examples do not contribute equally to model performance and some examples can even hurt model performance [1, 2]. Recent research has demonstrated that this observation applies to deep neural networks as well [3, 4]. Data Valuation in machine learning is concerned with principled methods for quantifying the relative contributions of individual examples in a training dataset to the performance metric achieved by a supervised learning model. The original motivation for data valuation is for data markets where individuals or data vendors can be compensated for providing their data [5]. In data-centric machine learning [6], quantifying the importance of data has applications in the curation and selection of training datasets. Identifying high quality data may allow for smaller training subsets that can reduce training times without a significant impact on model performance. Low-valued training data may be indicative of noise or annotation errors. Such examples may be amended to improve model performance or removed from the training set if they make negative contributions to a model's performance. Low values may also signal that the data arise from a distribution different than the target distribution of interest, or that the data are irrelevant for a model's task and thus not pertinent for a given problem. Hence, quantifying the value of data is useful for data curation as a means of ranking data based on the domain or task, and selecting high quality subsets to improve model performance or reduce training times.

The Shapley value [7], arising from economics and cooperative game theory, is an equitable method for allocating rewards among a coalition of game players. In machine learning, it has been used to value training examples based on their relative contributions to a model's performance on a training set [8, 9]. Shapley Values have been employed in federated learning for incentivizing participants to provide data [10], and in image [11], dialog [12], and audio [13] classification tasks for identifying data quality issues.

Modern automatic speech recognition (ASR) systems require large amounts of data with training that is compute- and time-intensive. For this reason, ASR systems stand to benefit from methods for data-efficient training [14] and one way to achieve this is by identifying high quality data. Methods for computing Shapley values have proven useful for classification models that accept fixed length inputs [8, 9]. However, ASR systems perform a structured prediction task of mapping variable length inputs to variable length outputs. To date, there has been no straightforward method for computing Shapley values of training examples for sequential models. Given a sequential ASR model trained on variable length speech utterances, our approach is to determine Shapley values of the constituent acoustic frames. We show that these values are indicative of data quality for an acoustic model. Further, we demonstrate a method for determining the value of speech utterances based on the utility of the constituent frames.

2. DATA VALUATION

Data valuation seeks to quantify the value of a training example based on its relative contribution to model's performance. There are two primary methods used for data valuation: *influence functions* [15] and *Shapley values* [8]. Influence functions determine the influence or value of an example by measuring the change in parameters of a model when an example is given a little more weight than the other examples in a training set. However, computing the influence function requires the model Hessian (second-order derivative) which can be unavailable or difficult to determine for complex models. The Shapley value [7] of an example is its expected marginal contribution to model performance when that example is added to any random subset of the training data. The Shapley value, described further in Section 3, is attractive because it is considered an economically equitable valuation of the data [7, 16]. Empirical results in data valuation [8] have shown that Shapley values are better for quantifying the importance of data for simple models such as Naive Bayes, and complex models such as deep neural network classifiers.

3. SHAPLEY VALUATION

In cooperative game theory, a coalition of players cooperate toward a common goal to earn a reward. Shapley valuation method for fairly allocating rewards to individual players [7] based on their contribution. The resulting allocation uniquely satisfies a set of basic fairness axioms [7]. In machine learning, training examples are viewed as players, and the learning algorithm uses information from the examples (players) to achieve a reward measured by a performance metric on a held-out set, e.g. accuracy. The Shapley value quantifies the contribution of each example toward the performance metric. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the training data and $\mathcal{D}_{\text{eval}} = \{(x_j, y_j)\}_{j=1}^{N_{\text{eval}}}$ be the held-out data used for evaluation. A learning algorithm $\mathcal{A}$ may use as input any subset $S \subseteq \mathcal{D}$ and its performance is measured by an evaluation function $U_{\mathcal{A}}(S)$. The Shapley value $\sigma(x_i)$ of example x_i is defined as the average marginal contribution of x_i when it is added to all possible subsets of the remaining examples, $S \subseteq D \setminus \{x_i\}$:

$$\sigma(x_i) = \sum_{S \subseteq \mathcal{D} \setminus \{x_i\}} \frac{1}{\binom{N-1}{|S|}} \left[U_{\mathcal{A}}(S \cup \{x_i\}) - U_{\mathcal{A}}(S) \right] \quad (1)$$

Shapley valuation is model agnostic and makes no assumptions about the distribution of the training data, $\mathcal{D}$.

Approximating Shapley values. Computing the Shapley value is computationally prohibitive ($\mathcal{O}(2^N)$), thus Shapley values are often estimated using Monte-Carlo (MC) algorithms [8]. The Truncated Monte Carlo (TMC) algorithm [8] allows for more efficient estimation using an early stopping criterion. However, MC methods can be computationally prohibitive for a complex model since they require re-training the model in each iteration. It has been shown that exact Shapley can be computed for k-Nearest Neighbors (KNN) model, using accuracy as the evaluation measure, with quasi-linear complexity. It is possible to take advantage of this fast procedure for complex models by learning a KNN proxy model for the more complex model [17]. This has been empirically demonstrated for deep neural network (DNN) classifiers in computer vision tasks. The proxy model approach takes advantage of the representation learned by the DNN. The neural embeddings are used as inputs, with the classification labels as outputs, to learn a k-Nearest Neighbors model (KNN) as a proxy for the DNN classifier.

4. METHODS

ASR systems perform a structured prediction task of mapping variable length speech to variable length text. We investigate estimating Shapley values for an end-to-end ASR system with an attention-based encoder-decoder architecture [18]. The encoder network learns an acoustic model, while the decoder network learns a language and transition model for outputting sequences of text. An acoustic model performs the classification task of mapping an acoustic unit (frame) to a linguistic unit, e.g. phoneme or grapheme. Thus, we can focus on the acoustic model component to determine the value of acoustic frames for that part of the ASR's function.

4.1. Acoustic Proxy Model

Given a speech utterance, we use the encoder network to map the constituent frames to the learned representation space. These neural features are inputs for a KNN model. We obtain ground truth phonetic labels for the frames by forced alignment, and these labels are outputs for the KNN model. Since the KNN model learns to map the acoustic representation of the frame to its phonetic label, it serves as a proxy for the acoustic model learned by the end-to-end ASR. We evaluate the utility of the proxy acoustic model by determining the frame classification accuracy on held-out utterances. We can then use the proxy KNN model to estimate Shapley values of the acoustic frames. Finally, we test whether the Shapley values have utility for acoustic modeling by evaluating the model's classification performance on held out frames. We sort the acoustic frames in an order determined by their Shapley values. In batches, we incrementally grow the training set by adding examples in the determined order. We learn a KNN for each batch and its performance on a heldout test set. This produces a performance curve for evaluating the Shapley values. We can produce two performance curves to visualize the results when examples are added in a best-first (highest Shapley values first) or worst-first order (lower Shapley values first). For a baseline, we grow the training set by adding examples in a random order. If the valuations signify utility of the examples for the model, the resulting performance should indicate when the model is learning from high- or low-valued examples, especially in comparison to the baseline where training examples are added randomly.

4.2. Valuation of Utterances

The preceding method allows for computing Shapley values of acoustic frames. Identifying high quality *frames* may prove useful for understanding the quality of data. However, we are ultimately interested in identifying subsets of high quality *utterances* because speech data for an ASR are almost always annotated at the utterance level. One approach is to aggregate the Shapley values of the constituent frames, for example summing the Shapley values for the frames in an utterance, and normalizing the sum by the duration of the utterance (or number of frames).

A more principled approach is inspired by the notion of a data vendor offering a batch of data in a market [5]. Treating an utterance as a collection of frames that are to be "sold" or valued together, or not at all, we can use similar methods for valuing the utterance. Unfortunately, we cannot apply the KNN-Shapley algorithm which values individual examples rather than batches of examples. Instead, we devise a *hybrid* approach by using the proxy KNN in conjunction with the MC algorithm. The general MC scheme is to uniformly sample the size of the (random) subset, then randomly select examples to fill that subset. Measuring the model performance of the training subset with and without an example provides the marginal contribution of the example (for one random subset). This is repeated several times until the values converge. In practice, we use an equivalent and computationally efficient sampling method, by sampling a random permutation of the training data, then scanning the permutation. For each example being scanned, we add it to the examples preceding it, then measure the change in performance with and without the example, to find its marginal contribution. Once the permutation has been scanned, we have a marginal contribution statistic for each example in the training set. After several rounds, we average the marginal contribution of an example to derive its estimated Shapley value. The TMC algorithm [8] uses a heuristic for early stopping: as a training subset becomes larger, the changes in performance will diminish allowing us to stop scanning the permutation once the change falls below a tolerance level.

For our hybrid approach of valuing an utterance, we modify this procedure. We assign an index to each utterance, then sample a permutation of these utterance indices. When an utterance is included in a set, all of its constituent frames are included in the training set. Removing an index from the permutation list corresponds to removing the constituent frames in that utterance from any training

set. Thus, we compute the marginal contribution of each utterance as the marginal contribution of all the constituent frames included together. The MC procedure requires re-training the model for each computation of the marginal contribution. The computational burden is alleviated through the use of the proxy KNN acoustic model. Our hybrid approach assumes that the effect of dropping or adding an utterance from the ASR training set can be approximated by the effect of adding or dropping the constituent frames from the proxy KNN model. We can evaluate the Shapley values on the original ASR model to see how well the assumption holds.

5. EXPERIMENTS AND RESULTS

5.1. Data and ASR Model

For computational reasons, our initial experiments focus on a small ASR task with a limited vocabulary. The CMU AN4 dataset [19] consists of 1,078 utterances, about 50 minutes total, from 84 male and female speakers. The utterances have an average duration 3 seconds with speakers describing personal information and control words. The data is split into 948, 100, and 130 utterances for the training, validation and test sets. The validation set has 8 speakers, all common with the training set. The test set has 10 speakers, independent of the training set. Our inputs are 83-dimensional vectors per 25 ms frame, using 80-dimensional log-Mel filterbank coefficients concatenated with a 3-dimensional pitch vector. We employ an end-to-end ASR system using a hybrid CTC-Attention model [20] using the standard recipe from the open-source ESPnet ASR framework [21]. We train the ASR model using the AN4 recipe [22] for 20 epochs. Using the final model to decode the validation and test sets results in word error rates (WER) of 16.8 and 9.8, respectively.

5.2. Frame Valuation

We use the ASR encoder to produce 320-dimensional neural features from the acoustic frames. Our input data for the proxy acoustic model consists of 56,854, 6,559, and 8,894 frames in the training, validation, and test sets, respectively. We obtain ground truth phonetic labels with 78 classes for the frames through forced alignment using the Kaldi speech recognition framework [23]. We learn a KNN model to map neural features to phonetic labels. Tuning on a held out set, we determine an optimal value of $k = 8$ for the KNN model. The proxy model achieves an accuracy of 77.7% on a test set with unseen speakers. The relatively high performance validates our approach of learning a proxy KNN for the ASR's acoustic model.

We estimate Shapley values of acoustic frames using the KNN-Shapley algorithm [17]. We evaluate the utility of the Shapley values for the proxy KNN acoustic model using a standard approach in the data valuation literature. We rank the examples in ascending order of Shapley value (i.e., from lowest-valued, or "worst", examples to highest-valued, or "best", examples). Starting with the complete training set of acoustic frames, we train a proxy KNN model and measure its performance on a held out set. We iterate over the ranked data, in fixed size batches, to incrementally drop the lowest-valued examples, from the training set. Each time we drop a batch of examples, we use the remaining (higher-valued) examples to train a proxy KNN model and measure the resulting performance. This produces a "drop worst-first" curve. We repeat the procedure by reversing the ranking, so that we gradually drop the highest-valued frames first. This produces a "drop best-first curve". Finally, we repeat the procedure with training data in a random order to produce a baseline. If the Shapley values are ordering the data points in a meaningful way, dropping the "best" examples should result in a large drop in overall performance and dropping the "worst" examples should result in a minimal drop in performance or even a performance increase (if those examples are mislabeled, for example).

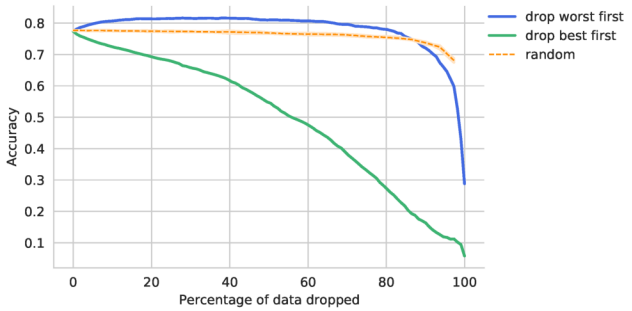

Fig. 1. Evaluation of Shapley values of acoustic frames. We measure proxy test accuracy as we drop frames used for training the KNN model. In drop-worst-first, we drop data with lowest Shapley values first, while in drop-best-first, we drop data with highest Shapley values first. The random curve shows results from 5 runs with the mean performance depicted as a dashed line.

Evaluating the Shapley values. The Shapley evaluation curves are shown in Figure 1. From the "drop worst-first" curve, we see that that dropping about 12% of lowest-valued frames yields an improvement in classification accuracy from 77.4% to 81.0%. Dropping about 32% of the lowest-valued frames provides the best model performance with 81.6% accuracy. This suggests that Shapley values are able to identify examples that may be misleading or difficult to learn, and thus not of utility to the model. We also see that it is possible to drop up to 82% before we see a decline in model performance below that from using the entire training data. Thus a model trained with one-fifth of the data, when selected appropriately, can achieve better performance than a model trained with all of the data. As the highest-valued frames are being dropped, the model performance begins to decline rapidly. This suggests that many frames have redundant information and convey similar patterns to the model. This is further supported by the "random" curve, which is relatively flat, suggesting that randomly dropping frames does not produce a significant drop in performance. This concurs with observations made for neural net learning [24]. About 30% of the available data is sufficient to produce the best performing acoustic model and can be identified by the (highest) 30% of Shapley values. Thus, we find that Shapley valuation of data can be used to curate data effectively.

We investigate why some frames received particularly low values and notice that low-valued frames tend to have annotation errors in the ground truth data. For example, in the utterance "ftmj-an213-b", which has a reference transcription of "RUBOUT P N A M X SEVENTY TWO", one frame with phone label "IY" has a particularly low Shapley value. We also note that the frame appears mislabeled in the ground truth annotation from the Kaldi model. Studying the example in Praat, we note that when the speaker utters the letters "P N" (with phonemes "P IY EH N"), the forced alignment is confused toward the end of the "IY" phone when the utterance transitions to the "EH" phone. The low Shapley value occurs because of this error in the annotation. We decode this training utterance using the ESPnet ASR model and receive a hypothesized transcription of "RUBOUT T M A M X SEVENTY T O". We note that the ASR model is also confused near this region, transcribing "P N" as "T M". Thus the

Shapley values are identifying annotation quality issues at the frame level and potential for confusion for the ASR model.

5.3. Utterance Valuation

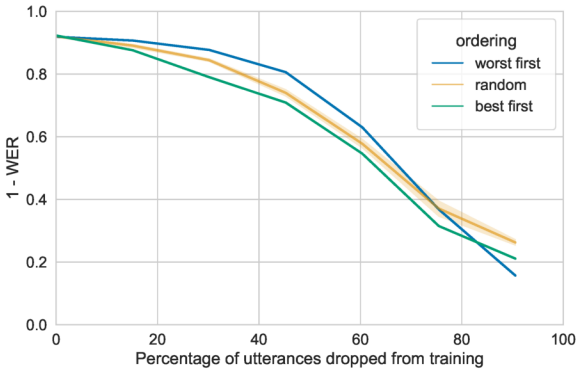

Fig. 2. Evaluation of Shapley values estimated for AN4 utterances. Performance scores are computed as word accuracy rate, $1 - \text{WER}$. Utterances are dropped from the training set in an order determined by their Shapley values. The random curve is the mean of 3 runs, with a 95% bootstrapped confidence interval.

We derive utterance values by aggregating the values of constituent frames and normalizing for duration. The resulting values do not provide a measure of utility of the training examples. This is because of a high variance in frame values within any given utterance. We proceed by valuing an utterance as a collection of frames that must be added or dropped from the training set. We use our hybrid approach with the Truncated Monte-Carlo algorithm and proxy KNN model for estimating the value of utterances. Figure 3 compares the distributions of the Shapley values of utterances to the Shapley values of the constituent frames. There is no clear trend between the two values, as shown by the regression line. For any given utterance, the Shapley values of frames have a high variance. This confirms our earlier findings that low quality frames are widely dispersed among utterances and it is hard to value utterances by aggregating frame values. Further, only about 3% of the utterances, having negative values, are hurting model performance. Thus we expect that most utterances in this dataset help model performance.

We evaluate the Shapley values by training the ASR with all of the data, then incrementally drop batches of utterances in an order determined by their Shapley value. We also perform 3 repetitions of randomly ordered training data. Figure 2 shows the results from our ASR evaluation. We see that especially in the middle of the chart, the best-first and worst-first curves are different from each other and the random curves. This suggests that the valuations have some utility. We notice within the first 10% of data, on the left of the chart, dropping the worst- or best-valued utterances both degrade performance in a similar way. This is different from our previous observations, where we see performance improve when dropping the lowest quality data. It may be that the AN4 data is relatively uniform in quality, and few to none of the examples mislead the model in any significant way. From the frame-level valuations, we observed a large number of redundant frames across utterances, with the lowest valued frames occurring in isolation across many utterances. Thus, there may be too few utterances that stand out as high or low quality.

We also see that the curves are well separated in the middle

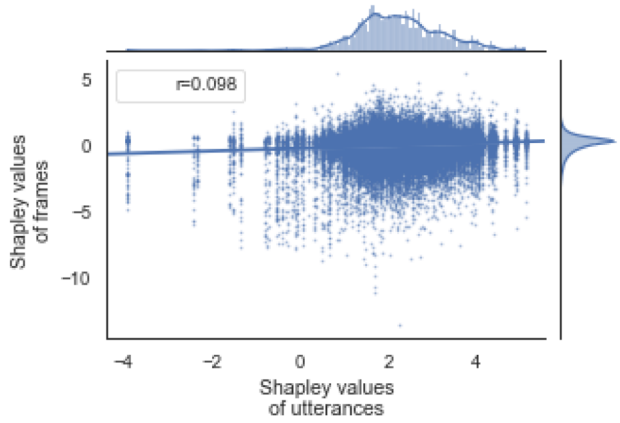

Fig. 3. Comparing standardized Shapley values of frames and utterances. We plot values of utterances against the corresponding values of the frames comprising those utterances. There is a wide variance of frame values within any utterance. The regression line shows the lack of trend between the two values.

region of the curve. Our method is better at identifying, for example, the best or worst 30% of training utterances rather than the best or worst 5% to 10%. It is possible that our Monte Carlo estimation scheme is better at producing valuations at a coarse or macroscopic level rather than producing fine-grained rankings of the data. Since our method treats an utterance as a "bag of frames", it loses any information about the sequential structure of the utterances, and does not account for it in the valuation. Another reason is that the neural network also learns representation from the data while performing the structured prediction task, which may require a significant amount of data. Finally, our valuation is ultimately based on the (proxy) acoustic model and may not account for the relative importance of data for the entire end-to-end model.

6. CONCLUSIONS

Data valuation methods have the potential to aid in collection and curation of high quality data for faster development of ASR systems. The methods proposed in data valuation have been limited to classification model and we present a method for applying these to sequential models performing a structured prediction task. We demonstrate how a KNN model can serve as a good proxy for the acoustic model component of an end-to-end ASR. We also show the acoustic model can be used to value acoustic frames and identify high and low quality subsets. We present a method for valuing utterances as collections of frames and are able to identify high and low quality batches of utterances for the end-to-end ASR model. Our method requires a Monte Carlo method for estimation and works well for relatively small datasets. In ongoing work, we are working on scaling the method for valuation of larger datasets. In addition to scalability issues, another disadvantage of our method is that it ignores the sequential structure of utterances and does not take into account the value of data for an ASR model beyond the acoustic model component. This remains an open problem for future work.

7. REFERENCES

[1] Hasan Ferdowsi, Sarangapani Jagannathan, and Maciej Zawodniok, "An online outlier identification and removal scheme for improving fault detection performance," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 908–919, 2013.

[2] Benoit Frenay and Michel Verleysen, "Classification in the presence of label noise: a survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2013.

[3] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey Gordon, "An empirical study of example forgetting during deep neural network learning," *arXiv:1812.05159*, 2018.

[4] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite, "Deep learning on a data diet: Finding important examples early in training," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20596–20607, 2021.

[5] Ramesh Raskar, Praneeth Vepakomma, Tristan Swedish, and Aalekh Sharan, "Data markets to support ai for all: Pricing, valuation and governance," *arXiv:1905.06462*, 2019.

[6] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, et al., "Dataperf: Benchmarks for data-centric ai development," *arXiv preprint arXiv:2207.10062*, 2022.

[7] Lloyd Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[8] Amirata Ghorbani and James Zou, "Data shapley: Equitable valuation of data for machine learning," in *International Conference on Machine Learning*, 2019, pp. 2242–2251.

[9] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gurel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos, "Towards efficient data valuation based on the shapley value," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 1167–1176.

[10] Shuyue Wei, Yongxin Tong, Zimu Zhou, and Tianshu Song, "Efficient and fair data valuation for horizontal federated learning," *Federated Learning: Privacy and Incentive*, pp. 139–152, 2020.

[11] Siyi Tang, Amirata Ghorbani, Rikiya Yamashita, Sameer Rehman, Jared A Dunnmon, James Zou, and Daniel L Rubin, "Data valuation for medical imaging using shapley value and application to a large-scale chest x-ray dataset," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[12] Weixin Liang, Kai-Hui Liang, and Zhou Yu, "Herald: an annotation efficient method to detect user disengagement in social conversations," *arXiv preprint arXiv:2106.00162*, 2021.

[13] Enis Berk Çoban, Ali Raza Syed, Dara Pir, and Michael I Mandel, "Towards large scale ecoacoustic monitoring with small amounts of labeled data," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 181–185.

[14] Qizhe Xie, *Towards Data-Efficient Machine Learning*, Ph.D. thesis, Carnegie Mellon University, 2020.

[15] Pang Wei Koh and Percy Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*. PMLR, 2017, pp. 1885–1894.

[16] Jon Kleinberg, Christos Papadimitriou, and Prabhakar Raghavan, "On the value of private information," in *Theoretical Aspects Of Rationality And Knowledge*, 2001, vol. 8.

[17] Ruoxi Jia, Xuehui Sun, Jiacen Xu, Ce Zhang, Bo Li, and Dawn Song, "An empirical and comparative analysis of data valuation with scalable algorithms," *arXiv:1911.07128*, 2019.

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

[19] Alex Acero, *Acoustical and environmental robustness in automatic speech recognition*, vol. 201, Springer Science & Business Media, 1992.

[20] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John Hershey, and Tomoki Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, 2017.

[21] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson-Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al., "Espnet: End-to-end speech processing toolkit," *Proceedings of Interspeech*, 2018.

[22] "ESPnet: AN4 recipe," Available at `https://github.com/espnet/espnet/tree/master/egs/an4/asr1`. Accessed: 2023-03-13.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[24] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al., "A closer look at memorization in deep networks," in *ICML*, 2017.