On Calibration of Scene-Text Recognition Models

Ron Slossberg[1], Oron Anschel[2], Amir Markovitz[2], Ron Litman[2], Aviad Aberdam[2], Shahar Tsiper[2], Shai Mazor[2], Jon Wu[2] and R. Manmatha[2]
{ronslos}@campus.technion.ac.il
{oronans, amirmak, litmanr, aaberdam, tsiper, smazor, jonwu, manmatha}@amazon.com

[1] Technion Institute, Haifa
[2] Amazon AWS AI Labs

Abstract. The topics of confidence and trust in modern scene-text recognition (STR) models have been rarely investigated in spite of their prevalent use within critical user-facing applications. We analyze confidence estimation for STR models and find that they tend towards overconfidence thus leading to overestimation of trust in the predicted outcome by users. To overcome this phenomenon we propose a word-level confidence calibration approach. Initially, we adapt existing single-output T-scaling calibration methodologies to suit the case of sequential decoding. Interestingly, extensive experimentation reveals that character-level calibration underperforms word-level calibration and it may even be harmful when employing conditional decoding. In addition, we propose a novel calibration metric better suited for sequential outputs as well as a variant of T-scaling specifically designed for sequential prediction. Finally, we demonstrate that our calibration approach consistently improves prediction accuracy relative to the non-calibrated baseline when employing a beam-search strategy.

1 Introduction

Scene Text Recognition (STR) – the task of extracting text from a cropped word image, has seen an increase in popularity in recent years. While an active research area for almost three decades, STR performance has recently seen a significant boost due to the utilization of deep-learning models [41, 1, 43, 3, 38, 25]. Some typical applications relying on STR models include assistance to the visually impaired, content moderation in social media and street sign recognition for autonomous vehicles. The above examples, often referred to as user-facing applications entail a high degree of trust. This can be achieved by reliably assessing the prediction confidence, *i.e.*, what the probability for a correct prediction is. Despite the prevalent usage of STR models within critical user-facing applications, confidence estimation for such models has not been thoroughly investigated in the past, thus leading to a misjudgment of the risk introduced by the model into the overall application.

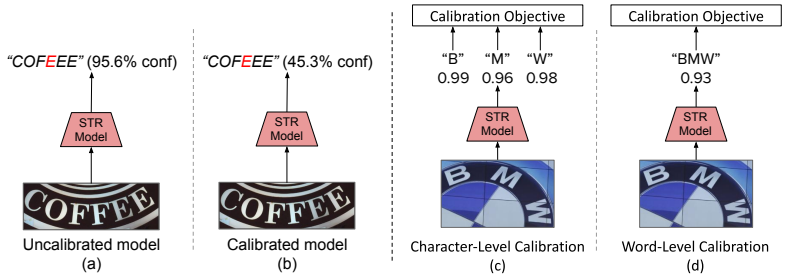

Fig. 1: **Overconfidence displayed by uncalibrated STR models**. Our analysis exposes the miscalibration of many modern STR models as depicted by (a). The calibrated score in (b) signals downstream tasks that the prediction should not be trusted. textbfCharacter vs. Word-Level Calibration. (c) Individual character score calibration. (d) Word-level calibration *i.e.* the confidence score for the entire word is directly optimized. We demonstrate the importance of adopting the word-level approach.

Confidence calibration is the task of tuning a model's confidence scores to match a successful prediction's underlying probability. For example, within the group of samples producing a confidence score of 0.7, we expect to achieve a prediction success rate of exactly 70%. Confidence calibration and model reliability have been active areas of research for many years [5, 7, 33]; however, the task of calibrating sequence-level confidence has received little attention and to the best of our knowledge, in STR, it has yet to be explored. In this work we study the confidence characteristics of modern STR models and expose the overconfidence tendency displayed by them. A similar phenomenon in the context of classification was previously observed by [12, 21, 14, 36] (see Figure 1 (a-b) for illustration). We conduct a comprehensive study encompassing a wide range of recent popular STR methods. Specifically, we examine various encoder choices coupled with conditional as well as non-conditional decoders and propose calibration techniques and practices to improve estimated model confidence scores.

Previous work has examined calibration and confidence intervals for structured prediction in the context of NLP problems [20, 8, 32, 22]. However, these methods focus on the calibration of marginal confidences, analogous to the calibration of each individual decoder output. This methodology resembles single-output model calibration as each classified token is treated individually as depicted in Figure 1 (c).

In this work, we show that calibration of text recognizers at the character-level is sub-optimal and in fact harmful when employing conditional decoders. Contrarily, we demonstrate that word-level calibration of STR models (see Figure 1 (d)) are more suitable to sequential decoding tasks independent of decoder choice. This is demonstrated in Figure 2 where the calibrated sequence length is gradually varied from character-level to word-level calibration. We specifically

note the increase in calibration error when applying character-level calibration to conditional decoders. In contrast we notice that error decreases when applying word-level calibration regardless of the decoder type.

Our main objective is to perform confidence calibration within real-time user-facing STR applications. We therefore limit our scope to methodologies that do not increase run-time complexity. Among this class of calibration methods, Temperature-scaling (T-scaling) was shown to be both simple and effective, often more so than other more complex methods [12, 37]. While previous works have demonstrated successful calibration of single output (classification) models, we adapt the T-scaling method to the sequence prediction task prescribed by STR. In addition, we extend the Expected Calibration Error (ECE) [31] proposed for binary classification problems to the regime of sequence calibration by incorporating a sequence accuracy measure, namely the edit-distance [24] metric. Furthermore, we present a useful application for confidence calibration by combining calibration with a beam-search decoding scheme, achieving consistent accuracy gains. Finally, we propose a sequence oriented extension to Temperature-scaling named Step Dependent T-Scaling, presenting moderate calibration gains for negligible added computational complexity.

Our key contributions are the following:

- We present the first analysis of confidence estimation and calibration for STR methods, discovering that numerous off-the-shelf STR models are badly calibrated.
- We highlight the importance of directly calibrating for word-level confidence scores and demonstrate that performing character-level optimization often has an adverse effect on calibration error.
- We demonstrate consistent accuracy gains by applying beam-search to calibrated STR models.
- We extend a commonly used calibration metric (ECE) to better suite partial error discovery we term "Edit-Distance ECE", and propose a sequence-based extension to T-scaling termed "Step-dependant T-scaling".

2 Related Work

Scene Text Recognition Shi *et al.* [41] proposed an end-to-end image to sequence approach without the need for character-level annotations. The authors used a BiLSTM [11] for modeling contextual dependencies, and Connectionist Temporal Classification (CTC) [10] for decoding. Baek *et al.* [1] proposed a four-stage framework unifying several previous techniques [4, 41, 42, 27]. The framework comprises the following building blocks: image transformation, feature extraction, sequence modeling, and decoding. Numerous subsequent methods also conform to this general structure [25, 43, 38, 3]. Currently, SOTA results are often achieved by methods adopting an attention-based [2] decoder scheme. The attention-based decoders usually consist of an RNN cell taking at each step the previously predicted token and a hidden state as inputs and outputting the next token prediction.

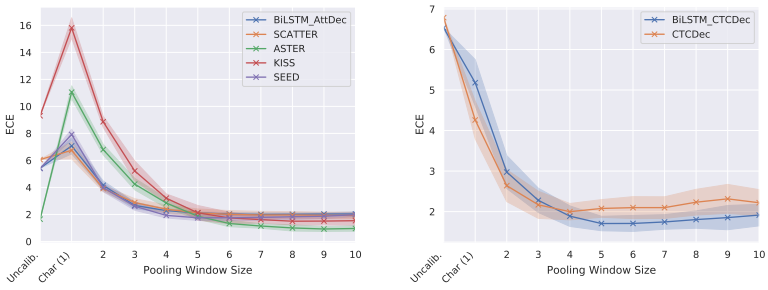

Fig. 2: **Calibrated Word-Level ECE Values vs. Pooling Window Sizes (n-grams)** for different STR methods. We evaluate our results on 10 saved checkpoints for each model, plotting the mean and standard deviation. **Left:** Conditional decoders. **Right:** Non-conditional decoders. All results are demonstrated on a held-out test-set. "Char" refers to character level calibration corresponding to a window size of 1. We observe that attention based models become uncalibrated when individual character calibration is performed and that longer window sizes are preferred over short ones during calibration.

Confidence Calibration Model calibration has been a subject of interest within the data modeling and general scientific communities for many decades [5, 7, 33]. Several recent papers [12, 21, 14, 36, 16, 34] have studied model calibration in the context of modern neural networks and classifier calibration (scalar or multi-class predictions). Empirically, modern neural networks are poorly calibrated and tend towards overconfidence. A common theme among numerous calibration papers is that Temperature-scaling (T-scaling) [12] is often the most effective calibration method even when compared to complex methods such as Monte-Carlo Dropout, Deep-ensemble, and Bayesian methods [35]. Nixon *et al.* [34] conduct a study on several proposed variations of established calibration metrics and suggest good practices for calibration optimization and evaluation. Similarly, we minimize an ECE calibration objective using a gradient framework.

Confidence Calibration for Sequential Models Most of the confidence calibration literature is focused on calibrating a single output classifier. Kuleshov and Liang [19] were the first to propose a calibration framework for structured prediction problems. The framework defines the notion of "Events of Interest" coupled with confidence scores allowing event-level calibration. The practical methods laid out by Kuleshov and Liang [19], however, predate the recent advances in DNNs.

Kumar *et al.* [20] address the problem of miscalibration in neural machine translation (NMT) systems. The authors show that NMT models are poorly calibrated and propose a calibration method based on a T-scaling variant where the temperature is predicted at each decoding step. They were also able to improve translation performance by applying beam-search to calibrated models. Our experiments find this to be beneficial for the task of STR as well.

Desai *et al*. [8] suggest the usage of T-Scaling for calibration of pre-trained transformers models (*e.g*. BERT [9]). The authors differentiate between in and out of domain calibration and propose using T-Scaling, and label-smoothing [36] techniques. We point-out that label smoothing is carried out during the training phase and therefore affects the model accuracy. Here, calibration is also conducted at the individual output level.

In another work [22], a proposed extension to T-scaling calibration for sequences is presented. The authors employ a parametric decaying exponential to model the temperature for each decoding step. Again, similarly to [20] calibration is performed for each decoding step and not for entire sequences.

3 Background

Temperature Scaling T-Scaling [12], a limited version of Platt Scaling [37], is a simple yet effective calibration method. T-scaling utilizes a single parameter $T > 0$. Given a logits vector $\mathbf{z}_i$ the model produces a calibrated score as:

$$\hat{q}_i = \max_k \ \sigma_{\text{SM}} \left(\mathbf{z}_i / T\right)^{(k)}, \tag{1}$$

where $\hat{q}_i$ denotes the estimated confidence of the i^{th} data sample, $\mathbf{z}_i \in \mathbb{R}^K$ is the output logits and K is the number of output classes (number of supported symbols for a STR model). T is a global scaling parameter and σ_{SM} is the softmax function defined as:

$$\sigma_{\text{SM}} \left(\mathbf{z}_i\right)^{(k)} = \frac{\exp\left(\mathbf{z}_i^{(k)}\right)}{\sum_{l=1}^{K} \exp\left(\mathbf{z}_i^{(l)}\right)}. \tag{2}$$

The temperature parameter T scales the logits, either altering the predicted confidence scores as necessary. T-Scaling is a monotonic transform of the confidence values, and therefore *does not affect the classifier model accuracy.*

Reliability Diagrams Figure 3 presents a visual representation of model calibration [7, 33]. Reliability diagrams show the expected accuracy for different confidence bins, where the diagonal represents a perfect calibration. Within each plot, the lower right triangle represents the overconfidence regime, where the estimated sample confidence is higher than its expected accuracy. We observe that the uncalibrated models are overconfident. We note that these plots do not contain the number of bin samples, and therefore, calibration error and accuracy cannot be directly derived from them.

Expected Calibration Error (ECE) Expected Calibration Error (ECE) [31] is perhaps the most commonly used metric for estimating calibration discrepancy. ECE is a discrete empirical approximation of the expected absolute difference between prediction accuracy and confidence estimation. The ECE is formally given by:

$$\operatorname{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m}, \mathbb{1}\left(\hat{y}_i = y_i\right), \tag{3}$$

$$\operatorname{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m}, \hat{q}_i, \tag{4}$$

$$\mathrm{ECE} = \sum_{m=1}^{B} \frac{|B_m|}{N} |\operatorname{acc}(B_m) - \operatorname{conf}(B_m)|. \tag{5}$$

Here, B_m denotes the set of samples belonging to the m^{th} bin, $|B_m|$ is the number of instances residing in bin b, N is the total number of samples, B is the total number of bins and $\mathbb{1}$ is the indicator function.

Since prediction accuracy cannot be estimated for individual samples but rather by taking the mean accuracy over a group of samples, the ECE score employs a binning scheme aggregating close by confidence values together. There is a resolution-accuracy trade-off between choosing more or fewer bins, and bin boundaries should be chosen carefully. During our experimentation, we choose an adaptive binning strategy proposed by [34], where the boundaries are set such that they split the samples into B even groups of N/B samples each.

This scheme adapts the bins to the natural distribution of confidence scores, thus, trading-off resolution between densely and sparsely populated confidence regions while keeping the accuracy estimation error even among the bins. We refer our readers to [12] for details and experimentation with different variations of the ECE metric.

Negative log likelihood The Negative Log Likelihood (NLL) objective is commonly used for classifier confidence calibration. NLL is defined as:

$$\mathcal{L} = -\sum_{i=1}^{n} \log\left(\hat{\pi}\left(y_i \mid \mathbf{x}_i\right)\right), \tag{6}$$

where the estimated probability $\hat{\pi}$ for the ground truth label y_i given the sample x_i is formulated as

$$\hat{\pi}\left(y_i \mid \mathbf{x}_i\right) = \sigma_{\mathrm{SM}}\left(\mathbf{z}_i\right)^{(y_i)}.$$

Brier Score Brier score [5] is a scoring method developed in an effort to predict the reliability of weather forecasts and has been subsequently adapted as a proxy for calibration error. Since the number of possible sequential labels is intractable, we treat the problem as a one vs. all classification, enabling the use of the binary Brier formulation. The Brier score as formalized in Equation 7 is the mean square error between the confidence scores and the binary indicator function over the predicted and ground truth labels.

$$\text{Brier} = \sum_{i=1}^{N} (\mathbf{1}\left(\hat{y}_i = y_i\right) - \hat{q}_i)^2. \tag{7}$$

As analyzed by [30], the Brier score comprises three components: uncertainty, reliability, and resolution. While confidence calibration is tasked to minimize the reliability term, the other terms carry information regarding the data uncertainty and deviation of the conditional probabilities from the mean. Therefore, while the Brier score contains a calibration error term, it is entangled with two other terms leading to sub-optimal calibration.

The main advantage of minimizing the Brier score is that it is parameter independent as it does not depend on data binning as ECE does. In Section 5, we demonstrate that minimization of Brier score leads to reduced ECE on a held-out test-set.

4 Sequence-Level Calibration

We propose to incorporate existing and new calibration methodologies while optimizing for word-level confidence scores, as depicted in Figure 1 (d). Our optimization scheme consists of a calibration model and a calibration objective applied to word-level scalar confidences.

Starting with a pre-trained STR model, we freeze the model weights and apply the T-scaling calibration method by multiplying the logits for each decoding step by the temperature parameter T. Following Kuleshov and Liang [19] we define the "Event of Interest" as the exact word match between predicted and ground-truth words. For each word prediction, we define a scalar confidence score as the product of individual decoding-step confidences. We assess our performance according to the ECE metric with equal-sized bins as suggested by Nixon *et al.* [34].

Calibration of Non-IID Predictions We motivate our choice of word-level calibration from a probabilistic viewpoint. Taking into account inter-sequence dependencies, we assume that the predictions made at each decoding step are non-IID. This is especially evident for RNN-based decoders *e.g.* Attention-based decoders, where each decoded character is provided as an input for the following prediction. In this case, the following inequality holds:

$$\mathbb{P}(\hat{Y} = Y|x) \neq \prod_{i} \mathbb{P}(\hat{y}_i = y_i|x). \tag{8}$$

Here, $\mathbb{P}(\hat{Y} = Y|x)$ denotes the correct prediction probability for the predicted sequence $\hat{Y}$ for input x, $\mathbb{P}(\hat{y}_i = y_i|x)$ are the marginal probabilities at each decoding step and $\hat{Y} = (\hat{y_1}, ..., \hat{y_L})$ are the predicted sequence tokens.

The calibration process attempts to affect the predicted scores so that they tend towards the prediction probabilities. Therefore, Equation 8 implies that

calibration of the marginals $\mathbb{P}(\hat{y}_i = y_i|x)$, corresponding to character-level calibration (Figure 1 (c)), will not lead to a calibrated word-level confidence (Figure 1 (d)). This insight leads us to advocate for direct optimization of the left hand side of Equation 8 *i.e.* the word-level scalar confidence scores.

Objective Function In previous work, several calibration objective functions have been proposed. Three of the commonly used functions are ECE, Brier, and NLL. Typically T-scaling is optimized via the NLL objective. Since the proposal of ECE by Naeini *et al.* [31], it has been widely adopted as a standard calibration measure. Adopting the findings of Nixon *et al.* [34], we utilize ECE as the calibration optimization objective. In our experiments, we also examine the Brier and NLL objectives. We find that while Brier can reduce ECE, it does not converge to the same minima. As for the NLL score, we find that it is unsuitable for sequence-level optimization as directly applying it to multiplied character scores achieves the same minima as character-level optimization. This is undesirable due to the inequality from Equation 8. (see supplementary for more details).

Edit Distance Expected Calibration Error (ED-ECE) A single classification is either correct or incorrect in its prediction. Sequential predictors, however, present a more nuanced sense of correct prediction *e.g.* correctly predicting 4 out of 5 characters is not as bad as predicting only 3. When calibrating in order to minimize the accuracy-confidence gap, the estimated accuracy is obtained by Equation 3. The indicator function implies a binary classification task, or in the multi-label setting, a one versus all classification. Sequence prediction, however, is more fine-grained and allows for partial errors. We propose to incorporate the error rate into a new calibration metric. By doing so we allow the end user to assess not only the probability for absolute error but also the amount of incorrect predictions within the sequence. To this aim, we propose to manipulate the ECE metric from by replacing Equation 3 with the following:

$$\mathrm{acc_n}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\left(\mathrm{ED}(\hat{y}_i, y_i) \leq n\right). \tag{9}$$

Here, ED refers to the Edit Distance function, also know as the Levenshtein Distance [24]. The Edit Distance produces an integer enumerating the minimal number of insertions, deletions, and substitutions performed on one string to produce the other.

We term our modified ECE metric – Edit Distance Expected Calibration Error (ED-ECE). Essentially, ED-ECE is a relaxation of the ECE metric where the error term is relaxed to better suite the sequential nature of textual prediction. ED-ECE provides a fine-grained per-character-centric metric, whereas ECE is a coarse-grained word-centric metric proposed in the context of single-prediction classification tasks. ED-ECE can be minimized for any desired string distance n and produces a confidence scores signifying the likelihood of an erroneous prediction up to an Edit Distance of n. This information is helpful for downstream applications such as dictionary lookup, beam-search or human correction, as

each example may be sent down a different correction pathway according to some decision scheme.

Step Dependent T-Scaling (STS) We extend T-Scaling to better suit the case of sequence prediction. As T-Scaling applies a single, global parameter to all model outputs, it does not leverage existing inter-sequence dependencies. This is especially true for context-dependent models such as Attention and Transformer based decoders. Therefore, we propose extending the scalar T-scaling to *Step Dependent T-scaling (STS)* by setting an individual temperature parameter for each character position within the predicted sequence. We replace the scalar temperature T with a vector $\mathbf{T} \in \mathbb{R}^{\tau+1}$, $\mathbf{T} = \{T_0, ..., T_\tau\}$, where τ is a truncation length. This may be formulated as:

$$\hat{q}_{i,j} = \max_k \sigma_{\mathrm{SM}} \left(\mathbf{z}_{i,j}\mathbf{T}_j\right)^{(k)} . \tag{10}$$

Here, $\mathbf{z}_{i,j}$ is the logits vector for the j^{th} character of the i^{th} sample, and $\mathbf{T}_j$ is the temperature value applied to the j^{th} character for all sequences.

Applying this method directly, however, results in sub-optimal results. This is due to the increase in trainable parameters for the same size calibration set. Furthermore, longer words are scarce and present high variability; thus, it may skew the temperature values of time steps above a certain index. We propose a meta-parameter τ that applies to all time steps over a certain value, such that $\mathbf{T}_{j \geq \tau} = \mathbf{T}_\tau$. We establish a value for τ on a held-out subset of the calibration dataset.

5 Experiments

In the following section, we carry out an extensive evaluation and analysis of our proposed optimization framework. We begin by detailing our experimental setup, including datasets, evaluated models, optimization methodology, and implementation details in Section 5.1.

In Section 5.2 we provide a deeper analysis, including sequence calibration for various aggregation window sizes (n-grams lengths). We further provide detailed results and analysis for the aspects described thus far: ED-ECE metric, calibration by decoding T-scaling step (STS), and the gains obtained through calibration and beam-search based decoding.

5.1 Experimental Setup

Datasets Following the evaluation protocol set by [44] who also focus on non-accuracy related aspects of STR models, all STR models are retrained on the SynthText [13] dataset. Models are then evaluated using both regular and irregular text datasets. Regular text datasets containing text with nearly horizontally aligned characters include: IIIT5K [29], SVT [45], ICDAR2003 [28], and ICDAR2013 [18]. Irregular text datasets are comprised of arbitrarily shaped text

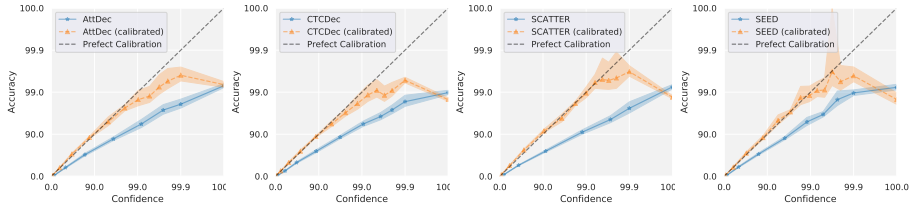

Fig. 3: **Reliability Diagrams [7]**: (i) AttDec – a variant of [1] with an attention decoder, (ii) CTCDec – a variant of [1] with a CTC decoder, (iii) SCATTER [25] and (iv) SEED [38]. We calibrate using T-scaling coupled with an equal-bin-size ECE objective applied to the word-level scalar confidence scores. The accuracy here is measured w.r.t exact word match. The figure shows accuracy vs. confidence plotted for equally-sized confidence bins, before and after calibration. Over-confidence can be observed for STR models, where the confidence of the model is higher than the expected accuracy.

(*e.g.* curved text), and include include: ICDAR2015 [17], SVTP [39], and CUTE 80 [40]. Our calibration set is comprised of the training portions of the aforementioned datasets where available (ICDAR2013, ICDAR2015, IIIT5K and SVT), while testing was performed on the testing portion of all seven regular and irregular datasets. By following this protocol we keep a strict separation between training, calibration and testing sets.

Text Recognition Models Our experiments focus on several recent STR models. Baek *et al*. [1] proposed a framework text recognition model comprising four stages: transformation, feature extraction, sequence modeling, and decoding. We consider architectures using four of their proposed variants, including or omitting BiLSTM combined with either a CTC [10] or an attention [23] decoder. In ASTER [43], oriented or curved text is rectified using an STN [15], and a BiLSTM is used for encoding. A GRU [6] with an attention mechanism is used for decoding. SCATTER [25] uses a stacked block architecture that combines both visual and contextual features. They also used a two-step attention decoding mechanism for providing predictions. Bartz *et al*. [3] proposed KISS, combining region of interest prediction and transformer-based recognition. SEED [38] is an attention-based architecture supplemented with a pre-trained language model.

In our work, we evaluate and analyze each of these models' calibration-related behavior, highlighting differences between various decoder types.

Optimization The task of calibrating confidence scores boils down to minimizing the model parameters w.r.t a given loss function. T-scaling based calibration methods take the predicted logits $\mathbf{z}_i$ as input and apply a modified SoftMax operation to arrive at the calibrated confidence score.

Backpropagation is only conducted through the optimized calibration parameter, while the STR model remains unchanged. The model formulations for T-scaling and STS are provided in Equation 1 and Equation 10 respectively.

Method	Uncalib.	ECE	Brier	NLL
CTC [1]	6.9	2.2	5.9	6.8
BiLSTM CTC [1]	6.7	2.0	6.0	6.6
Atten. [1]	5.9	1.8	5.2	5.9
BiLSTM Atten. [1]	5.4	2.0	4.8	5.3
ASTER [43]	1.8	0.8	1.8	1.7
SCATTER [25]	5.8	1.8	4.5	5.7
KISS [3]	9.6	1.4	5.3	9.4
SEED [38]	5.7	2.0	5.7	5.6
Average	5.98	**1.75**	4.9	5.88

Table 1: **Calibrated ECE Scores for Different Objective Functions.** Unsurprisingly, ECE values are best optimized w.r.t. the ECE objective. Brier loss is also suitable for reducing calibration error but is less effective. Finally, we observe that NLL is unsuitable for sequence level calibration as detailed in Section 4 and in the supplementary.

We use the L-BFGS optimizer [26] coupled with several calibration objective functions to demonstrate our calibration methodology. Our tested loss functions include ECE, ED-ECE, Brier, and NLL. Table 1 presents ECE achieved by calibrating for NLL, Brier, and the ECE objective functions. We observe that ECE obtains the best calibration error as expected while Brier succeeds to a lower degree. We also demonstrate that, as mentioned, NLL is not suitable for word-level calibration.

5.2 Results and Analysis

Aggregation Window In order to gain a deeper understanding of the relation between aggregation and calibration performance, we experimented with calibration via partial sequences. To this end, we break up our calibration datasets into all possible sub-sequences of length $\leq n$. We note that when $n = 1$ the calibration is carried out at character-level as depicted by Figure 1 (c), and for $n = \max \text{length}(w_i)$ we are calibrating on the full sequence (Figure 1 (d))

Calibration is performed by the T-scaling method coupled with the ECE objective function. All reported results are measured on a held-out test-set. In an attempt to reduce noise, we test the calibration process on 10 training checkpoints of each model and plot the mean and variance measurements in Figure 2.

We find that for attention decoders (Figure 2 (Left) $n = 1$ provides worse calibration than the uncalibrated baseline. CTC decoders (Figure 2 (Right), on the other hand, also exhibit worse ECE scores on per-character calibration; however, the error is still reduced relative to the uncalibrated models. We postulate that this phenomenon relates to the difference between IID and non-IID decoding discussed in Section 4. This key observation emphasizes the importance of score aggregation during the calibration process as opposed to individual character calibration.

Beam-Search Although calibration methods based on T-scaling do not alter prediction accuracy, it is still possible to indirectly affect a model's accuracy rate. This can be achieved through a beam-search methodology, where the space of possible predicted sequences is explored within a tree of possible outcomes. At each leaf, the total score is calculated as the product of all nodes leading up to the leaf.

We note that each predicted character's confidence score does not change ordering with relation to other scores due to the monotonic nature of T-scaling. In contrast, aggregated word-level confidence scores do change ordering in some cases. Score reordering can take place if individual character scores have a non-monotonic dependence on other parameters than the calibrated temperature. This is the case for state-dependant decoders where the scores depend on the internal decoder state, therefore allowing for a reordering of the word-level confidence scores.

We present our calibrated beam-search results in Table 2, showing a consistent gain for each calibrated mode1 relative to the non-calibrated baseline. We also show that this holds for all tested beam widths between two and five. In the supplementary material, we further break down the results according to the individual test datasets. We note that while the overall absolute improvement is below 1%, this is significant when considering the baseline of improvement offered by the uncalibrated beam-search method, in some cases more than doubling beam-search effectiveness. While beam-search on its own is quite intensive, requiring several inference steps for each word, our calibration is performed offline and adds virtually no further computational burden while increasing effectiveness by as much as two fold.

	bw=1	bw=2		bw=3		bw=4		bw=5	
Method\Calibration	X	X	✓	X	✓	X	✓	X	✓
Atten. [1]	86.21	0.18	0.39	0.23	0.43	0.26	**0.47**	0.25	0.45
BiLSTM Atten. [1]	86.2	0.11	0.2	0.18	0.33	0.2	**0.35**	0.21	**0.35**
ASTER [43]	86.03	0.52	0.57	0.63	0.77	0.64	**0.82**	0.64	0.81
SCATTER [25]	87.36	0.16	0.27	0.2	**0.3**	0.19	0.28	0.2	0.29
SEED [38]	81.18	0.2	0.29	0.23	0.35	0.32	**0.42**	0.28	0.4
Average	84.53	0.23	0.34	0.29	0.44	0.32	**0.47**	0.32	0.46

Table 2: **Beam-search accuracy gains** achieved for calibrated (✓) vs. uncalibrated (X) models. We apply beam widths (bw) between 1 and 5. Displayed results are averaged across test datasets and are reported relative to the baseline of bw = 1, which is equivalent to not using beam-search. CTC based methods are omitted as beam-search requires decoding dependence in order to be effective. We demonstrate consistent improvement across all models and all datasets (see supplementary for breakdown) over the uncalibrated baseline.

**Method **	$E_d = 0$		$E_d \leq 1$		$E_d \leq 2$	
Calibration	X	✓	X	✓	X	✓
CTC [1]	6.9	2.2	4.1	1.8	8.2	1.7
BiLSTM CTC [1]	6.7	2.0	4.2	1.5	7.9	1.9
Atten. [1]	5.9	1.8	2.0	0.9	4.5	0.7
BiLSTM Atten. [1]	5.4	2.0	2.1	1.0	5.0	0.7
ASTER [43]	1.8	0.8	5.3	2.1	9.7	1.6
SCATTER [25]	5.8	1.8	1.6	1.2	3.9	0.9
KISS [3]	9.6	1.4	1.0	0.8	5.2	1.2
SEED [38]	5.7	2.0	2.1	1.2	4.9	0.7
Average	5.98	1.75	2.8	1.31	6.16	1.18

Table 3: **ED-ECE Values** for uncalibrated (X) and calibrated (✓) models. Calibration was performed by the T-scaling method and ED-ECE objective for $E_d = 0$ (equivalent to ECE) and $E_d \leq 1, 2$. We observe that the optimization process reduces ED-ECE values on the held-out test-set. The calibrated ED-ECE scores may be used to estimate the number of incorrect predictions within the sequence for down-stream applications.

Edit Distance Expected Calibration Error In Section 4 we present a new calibration metric termed Edit Distance Expected Calibration Error (ED-ECE). We calibrate for ED-ECE with $n = [1, 2]$ and present the results in Table 3. As expected, the ED-ECE is reduced significantly due to the optimization. It is worth noting that ED-ECE is often lower than the original ECE score, leading to a more accurate confidence estimation. Once calibrated, three scores corresponding to ECE (ED-ECE for $n = 0$) and ED-ECE for $n = [1, 2]$ are produced for each data sample. This allows us to submit the data to further review according to thresholds on the output scores.

For example given a predicted label "COFEEE" for a ground-truth label of "COFFEE", the absolute and $E_d = 1$ predicted confidences are 75% and 98% respectively. We might recognize such an example and perform a focused search by testing the confidence scores of words that differ by an Edit-Distance of 1 and selecting the most confident prediction within the search space.

Step Dependent T-Scaling (STS) In Section 4 we propose *Step Dependent T-Scaling (STS)*. STS extends the previously presented T-scaling by assigning a temperature for each character position in the sequence. Table 4 lists the calibrated ECE values achieved by T-scaling as well as STS calibration schemes coupled with an ECE calibration objective function. We find that a value of $\tau = 5$ is optimal for the held-out validation set and therefore select $\tau = 5$ temperature values while a 6^{th} value is used to calibrate the subsequent sequence positions. Our experimentation demonstrates that the Time-Stamp scaling is beneficial or on par with T-scaling for all but one of the models. Overall, when averaging on all tested models, STS shows a slight benefit over T-scaling. Although relatively small, this benefit is achieved for very low additional complexity to the offline

Method	Uncalib.	TS	STS
CTC [1]	6.9	**2.2**	**2.2**
BiLSTM CTC [1]	6.7	2.0	**1.8**
Atten. [1]	5.9	1.8	**1.7**
BiLSTM Atten. [1]	5.4	2.0	**1.7**
ASTER [43]	1.8	**0.8**	1.0
SCATTER [25]	5.8	1.8	**1.6**
KISS [3]	9.6	**1.4**	**1.4**
SEED [38]	5.7	**2.0**	**2.0**
Average	5.98	1.75	**1.67**

Table 4: **ECE Values Comparing T-scaling and STS** for uncalibrated (Uncalib.) and calibrated confidence scores obtained on a held-out test-set. We optimize according to our proposed method utilizing the T-scaling (TS) and the proposed STS calibration methods. We demonstrate that STS is slightly advantageous over global TS.

calibration process. We hypothesize that STS is able to improve calibration error due to its finer-grained calibration and exploitation of inter-sequence relations.

6 Conclusion

In this work, we analyze the calibration characteristics of off the shelf STR models finding that they are commonly over confidant in their estimation. We further demonstrate that word-level and, in general, sequence-level calibration should be optimized directly on the per-word scalar confidence outputs. This is motivated by probabilistic reasoning and demonstrated empirically for various STR methods.

To the best of our knowledge, we are the first to conduct an in-depth analysis of the current state of calibration in scene-text recognition models. We perform extensive experimentation with STR model calibration and propose ED-ECE, a text-oriented metric and loss function, extending ECE to calibrate for sequence-specific accuracy measures (*e.g.* Edit-Distance).

Furthermore, we demonstrate that the calibration of STR models boosts beam-search performance, consistently improving model accuracy for all beam-widths and datasets with relative accuracy improvement of up to double relative to the non-calibrated baseline. Finally we propose to extend the T-scaling calibration method to a sequence-level variant we termed Step dependent T-Scaling (STS), showing moderate gains for very little effort.

Bibliography

[1] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4715–4723 (2019)

[2] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 (2015)

[3] Bartz, C., Bethge, J., Yang, H., Meinel, C.: Kiss: Keeping it simple for scene text recognition (2019)

[4] Borisyuk, F., Gordo, A., Sivakumar, V.: Rosetta: Large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 71–79 (2018)

[5] Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly weather review **78**(1), 1–3 (1950)

[6] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1179, https://www.aclweb.org/anthology/D14-1179

[7] DeGroot, M.H., Fienberg, S.E.: The comparison and evaluation of forecasters. Journal of the Royal Statistical Society: Series D (The Statistician) **32**(1-2), 12–22 (1983)

[8] Desai, S., Durrett, G.: Calibration of Pre-trained Transformers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020)

[9] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423

[10] Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376. ACM (2006)

[11] Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 6645–6649. IEEE (2013)

[12] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 1321–1330. ICML'17, JMLR.org (2017)
[13] Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2315–2324 (2016)
[14] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks (2017)
[15] Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
[16] Ji, B., Jung, H., Yoon, J., Kim, K., et al.: Bin-wise temperature scaling (bts): Improvement in confidence calibration performance through simple scaling techniques. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 4190–4196. IEEE (2019)
[17] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: Icdar 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1156–1160. IEEE (2015)
[18] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: Icdar 2013 robust reading competition. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1484–1493. IEEE (2013)
[19] Kuleshov, V., Liang, P.S.: Calibrated structured prediction. In: Advances in Neural Information Processing Systems. pp. 3474–3482 (2015)
[20] Kumar, A., Sarawagi, S.: Calibration of encoder decoder models for neural machine translation. arXiv preprint arXiv:1903.00802 (2019)
[21] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems. pp. 6402–6413 (2017)
[22] Leathart, T., Polaczuk, M.: Temporal probability calibration. arXiv preprint arXiv:2002.02644 (2020)
[23] Lee, C.Y., Osindero, S.: Recursive recurrent nets with attention modeling for ocr in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2231–2239 (2016)
[24] Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710 (1966)
[25] Litman, R., Anschel, O., Tsiper, S., Litman, R., Mazor, S., Manmatha, R.: Scatter: Selective context attentional scene text recognizer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
[26] Liu, D.C., Nocedal, J.: On the limited memory bfgs method for large scale optimization. Mathematical programming **45**(1-3), 503–528 (1989)
[27] Liu, W., Chen, C., Wong, K., Su, Z., Han, J.: Star-net: A spatial attention residue network for scene text recognition. In: BMVC (2016)

[28] Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R.: Icdar 2003 robust reading competitions. In: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. pp. 682–687. Citeseer (2003)

[29] Mishra, A., Alahari, K., Jawahar, C.: Scene text recognition using higher order language priors (2012)

[30] Murphy, A.H.: A new vector partition of the probability score. Journal of applied Meteorology **12**(4), 595–600 (1973)

[31] Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence. vol. 2015, p. 2901. NIH Public Access (2015)

[32] Nguyen, K., O'Connor, B.: Posterior calibration and exploratory analysis for natural language processing models. In: EMNLP (2015)

[33] Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning. pp. 625–632 (2005)

[34] Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)

[35] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems. pp. 13991–14002 (2019)

[36] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:1701.06548 (2017)

[37] Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)

[38] Qiao, Z., Zhou, Y., Yang, D., Zhou, Y., Wang, W.: Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13528–13537 (2020)

[39] Quy Phan, T., Shivakumara, P., Tian, S., Lim Tan, C.: Recognizing text with perspective distortion in natural scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 569–576 (2013)

[40] Risnumawan, A., Shivakumara, P., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. Expert Systems with Applications **41**(18), 8027–8048 (2014)

[41] Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2298–2304 (2016)

[42] Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4168–4176 (2016)
[43] Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE transactions on pattern analysis and machine intelligence (2018)
[44] Wan, Z., Zhang, J., Zhang, L., Luo, J., Yao, C.: On vocabulary reliance in scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11425–11434 (2020)
[45] Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: 2011 International Conference on Computer Vision. pp. 1457–1464. IEEE (2011)