Towards the Unseen: Iterative Text Recognition by Distilling from Errors

Ayan Kumar Bhunia1 Pinaki Nath Chowdhury1,2 Aneeshan Sain1,2 Yi-Zhe Song1,2
1SketchX, CVSSP, University of Surrey, United Kingdom.
2iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{a.bhunia, p.chowdhury, a.sain, y.song}@surrey.ac.uk.

Abstract

Visual text recognition is undoubtedly one of the most extensively researched topics in computer vision. Great progress have been made to date, with the latest models starting to focus on the more practical “in-the-wild” setting. However, a salient problem still hinders practical deployment – prior state-of-arts mostly struggle with recognising unseen (or rarely seen) character sequences. In this paper, we put forward a novel framework to specifically tackle this “unseen” problem. Our framework is iterative in nature, in that it utilises predicted knowledge of character sequences from a previous iteration, to augment the main network in improving the next prediction. Key to our success is a unique cross-modal variational autoencoder to act as a feedback module, which is trained with the presence of textual error distribution data. This module important translates a discrete predicted character space, to a continuous affine transformation parameter space used to condition the visual feature map at next iteration. Experiments on common datasets have shown competitive performance over state-of-the-arts under the conventional setting. Most importantly, under the new disjoint setup where train-test labels are mutually exclusive, ours offers the best performance thus showcasing the capability of generalising onto unseen words (Figure 1 offers a summary).

1. Introduction

Text recognition being a longstanding problem in computer vision plays a pivotal role in a diverse set of applications, ranging from OCR systems [4, 42, 48, 54], navigation and guiding board recognition [10], to more recent ones such as visual question answering [5]. With the advance of deep learning [50, 62, 10, 43], recognition accuracy have notably increased over traditional methods [36]. Research focus has thus shifted to the more practical “in-the-wild” setting in an attempt towards ubiquity. Of these, irregular scene text recognition [50, 10, 58, 60] has gained considerable attention, yet the focus is placed on irregular-image rectification process [62, 58] other than the core recognition problem itself.

In this paper, we continue this push towards practicality, albeit with a different perspective – we importantly focus on the understudied problem of unseen (or rarely seen) word recognition, where no (or limited) word image of a particular character sequence is present during training. Our motivation is straightforward – humans can recognise a word image, even when it falls beyond the scope of known vocabulary. In fact, robustness of a text recognition framework largely depends on its performance on rarely or unseen words [52]. Note that unlike the conventional zero-shot [59] setting where the transfer happens on class-level, here the combination of characters is “unseen” although the characters individually have come up in training. The fact that the sequences not being encountered during training is what makes this task challenging. Our solution for this “unseen” problem is intuitive: (i) we leverage an iterative framework with a feedback mechanism to guide the model to revisit its false predictions, and (ii) we explicitly ask such feedback to encapsulate useful information that would help the model to correct itself at the following iteration.

Our first contribution is therefore an iterative framework, where characters predicted in the previous iteration provide clues through a feedback loop [17] to enhance performance in the subsequent iterations. This is fundamentally different to current state-of-the-arts [50, 10, 29], most of which adapt
a feed-forward framework consisting of three-components (feature-extracting backbone, bidirectional-RNN encoder, and attention-based recurrent decoder). Despite the attention mechanism, its single-pass nature still dictates wrong predictions, thereby leaving no chance for the model to recover. To this end, our iterative design enables the revision of incorrect intermediate predictions in its subsequent steps, via a novel cross-modal (i.e., text prediction to image feature-maps) feedback mechanism. The key to our success lies with how feedback is progressed at each iteration. A naive solution might be to apply an independent spelling correction network [13, 56] chained serially to a basic text recognition model. Apart from not being end-to-end trainable, this also ignores the intermediate visual features from the recognition network, ultimately bisecting the feedback loop. We on the other hand advocate that earlier word predictions (text labels) should be fed back cross-modal to the main text recognition network and directly modulate the visual feature maps at the next iteration. That is, the feedback module triggers a mapping from the discrete predicted label space, to a continuous space of affine transformation parameters (akin to [41]) which are consequently used to condition visual features (hence closing the feedback loop).

Simply knowing how feedback works is not enough – we still need to devise what information should be fed back to give a model its best shot at rectifying itself. For this, we resort to distilling knowledge from textual error distributions collected from state-of-the-art text recognition models – this is akin to humans who use prior experiences to help them to make corrections. For example, ‘hello’ might be wrongly predicted as ‘nello’ or ‘bello’ due to partial structural similarity of ‘h’ with ‘n’ or ‘b’. By distilling such error distributions into the feedback module (during training only), the model will gain knowledge of correct character associations. Our second contribution is therefore designing the feedback module via a conditional variational auto-encoder (CVAE) [51] that learns from such error distributions. More specifically, we augment the vanilla CVAE with an auxiliary decoder that tries to directly reconstruct the correct word, given any incorrect prediction at each iteration. Note that deterministic alternatives such as typical feedback networks [47, 17] or spelling correction (prediction refining) networks, [13, 56] would not work well since they do not model the uncertainty among multiple erroneous alternatives, dictating a variational formulation like ours.

Our contributions are: [a] We for the first time propose an iterative approach to specifically tackle the “unseen” text recognition problem. [b] We design a conditional variational autoencoder to act as a feedback module, which works cross-modal to propagate predicted text labels from an earlier iteration to condition the visual features from the main network. [c] Our novel cross-modal feedback module is trained by distilling knowledge learned from textual error distributions that model multiple erroneous character sequences to a given candidate word.

Experiments confirm our framework to be capable of adopting unseen words better than state-of-the-art frameworks on various public scene-text recognition and handwriting recognition datasets. Further ablative studies demonstrate the superiority of our iterative framework over naive spell checking [56, 13] and language model alternatives [22], and that the proposed feedback module can be plug-and-play with more than one base network.

2. Related Works

Text Recognition: With the rising applicability of deep learning methods, Jaderberg et al. [20] employed convolutional neural network for text recognition, but such methods were constrained to dictionary words. Although this limitation was eliminated by Jaderberg et al. [18], it still needed huge resources for character level localisation. Sequence discriminative training using connectionist temporal classification (CTC) [11] coupled with recurrent neural networks, dealt with the need for character level localisation. This lead to an end-to-end trainable convolutional recurrent neural network for reading texts [48]. This was further enhanced by incorporating the idea of attention [3] for text recognition [26, 49]. Usually a two-fold process is implemented [58, 49, 50, 62], where an irregular image is first passed through a rectification network, and then followed by a text recognition network. Ideas like 2D attention mechanism, focusing attention network (FAN) [9] has been explored recently along-with the possibility of enhancing text reading accuracy using synthetically generated large datasets [63]. Despite such extensive research, reading irregular curved texts had not been explored much in details until recently in [10], which describes how arbitrarily oriented texts can be read by extracting four directional features from the 2D input image. Baek et al. [2] has conducted a comparative study where different popular text recognition architectures were trained in a similar setting. Hence it can be observed that although recent works in text reading have emphasised on designing a better rectification network [62, 58], all such research [31, 57] essentially used an off-the-shelf [48] recognition module. On the other side, handwriting recognition poses a tougher challenge owing to a free flow nature [4] of writing. Poznanski et al. [42] employed a ConvNet to estimate an n-gram frequency profile along with a large dictionary having true frequency profiles for recognition.

Feedback Mechanism: Carreira et al. [7] augments input space based on a corrective signal output manifold, improving on human pose estimation and generalising to instance segmentation tasks [28]. While Wei et al. [55] used a ConvNet followed by a similar module having a larger receptive field, Newell et al. [37] developed an hour-glass network
design, stacked together for merging information across all input scales. Zamir et al. [61] proposed a novel network architecture aligned with a feedback notion, functionally akin to ResNet architecture. Others include instance segmentation [28], few-shot learning [64], object detection [8], super resolution [35] and image generation [47,17]. However, here we introduce cross-modal iterative feedback for text recognition with a conditional variational autoencoder that enables modelling of a prior knowledge of linguistically correct character-sequences. 

**Error Correction:** Earlier efforts to utilise this idea of refining character-sequence prediction can be found in automatic speech recognition (ASR) community. In works such as Rozovskaya et al. [46] and Hans et al. [15] preposition errors were detected using classifiers. Grammatical error correction was approached using statistical machine translation methods by Ng et al. [38]. Recently, the idea of using recurrent encoder-decoder architecture in conjunction with attention mechanism was presented by Xie et al. [56] to address complex orthographic errors. Later this idea was adopted by Guo et al. [13] for tasks related to ASR.

### 3. Methodology

Instead of generating results in a single feed forward pass [50,30,58,65], we propose an iterative approach towards text recognition. As our work solely contributes to the text recognition part, we have refrained from detailing on the initial rectification network. We have used an off-the-shelf rectification network based on Spatial Transformer Network [21] and Thin Plate Splines [2] from Shi et al. [50].

#### 3.1. Text Recognition Module

The rectified image received from the rectification network [50] is fed into this text recognition network \( T \) aiming to produce a character sequence \( Y = \{y_1, y_2, ..., y_K\} \), where \( K \) denotes the variable length of text. Given an image \( I \in \mathbb{R}^{H \times W \times C} \), the convolutional feature extractor tries to learn rich visual information and produce a feature map of size \( \mathbb{R}^{H' \times W' \times D} \), where \( H', W', \) and \( D \) are the height, width and number of channels in the output feature map respectively. That output is reshaped into a sequence of feature vectors \( B = [b_1, b_2, ..., b_L] \), where \( L = H' \times W' \), and \( b_i \in \mathbb{R}^{D} \). Every \( D \) dimensional feature vector \( b_i \) encodes a particular local image region based on its receptive fields. Thereafter a bidirectional-LSTM is employed to capture the long range dependency in both the directions, thus alleviating the constraints of limited receptive fields. It outputs an updated feature sequence of same length, denoted by \( H \in [h_1, h_2, ..., h_L] \). Following this \( y_k \) is decoded based on three factors viz; bidirectional-encoder output \( H \), the previous internal state \( s_{k-1} \), and the character \( y_{k-1} \) predicted in the last step. At time step \( k \), a recurrent decoder network generates an output vector \( o_k \) and a new state vector \( s_k \), defined as: \( (o_k, s_k) = \text{RNN}(s_{k-1}, [s_k, E(y_{k-1})]) \) where \( g_k \) is the glimpse vector [50] that encodes the information from specific relevant parts of the encoded feature \( H \) to predict \( y_k \); \( E \) is an embedding layer, and \([\cdot]\) signifies a concatenation operation. Here, \( g_k \) is computed as:

\[
g_k = \sum_{i=0}^{L} \left( \frac{\exp(a_{k,i})}{\sum_{j=0}^{L} \exp(a_{k,j})} \right) h_i \quad \text{and,}
\]

attention score \( a_{k,i} = v^T \tanh(W_s s_{k-1} + W_h h_i + b_o) \), where \( v, W_s, W_h, b_o \) are the learnable parameters. Finally, the current step character \( y_k \) is predicted by:

\[
p(y_k) = \text{softmax}(W_o o_k + b_o) \quad \text{where} \quad W_o \text{ and } b_o \text{ are trainable parameters.}
\]

#### 3.2. Cross-Modal Variational Feedback

**Overview:** Unlike previous attempts of devising mostly deterministic feedback modules [17,47,61], we propose a Cross-Modal Variational Feedback network instilling benefits of variational autoencoder (VAE) [25], a powerful class of probabilistic models. Let us consider a text recognition network \( T \) that has been dissected into two parts, namely \( T_A \) and \( T_B \) where the particular position of dissection is obtained empirically as described in section 4.3. The prior knowledge from the resultant discrete character space is obtained empirically as described in section 4.3. The prior knowledge from the resultant discrete character space is defined as:

\[
\hat{\Psi} = \Psi(\hat{\Phi} ; \Psi')
\]

Learning \( \hat{\Phi} \) and \( \Psi' \) can be done in two ways: by back-propagation or by using a Generative Adversarial Network (GAN) [41].

**Feedback Conditioning Layer:** The primary objective of this layer is the propagation of prior knowledge from previous feed-forward pass prediction to \( T \). This action couples the rich visual information extracted by \( T_A \), with the prior feedback signal coming from earlier prediction. A noteworthy mention in this context would be the work by Perez et al. [41], where a general purpose conditioning layer FiLM has been designed based on simple feature-wise affine transformation operation for visual reasoning. Works involving visual-question-answering [41], image style transfer [16] and semantic image synthesis [39] have been seen to endorse a similar idea as well. Based on some prior or conditional information, the intermediate activation map \( \Psi \) could be modulated as: \( \Psi = \Psi \circ \gamma + \omega \), where \( \gamma, \omega \) are global affine transformation parameters. They both have dimensions \( \mathbb{R}^T \), and are usually predicted by a network receiving the conditional information as input. In lieu of transforming each channel globally, we allow local transforma-
Inference in the model consists of predicting $Y'$ using any arbitrary number of correction steps.

(iii) Inspired by the auxiliary task approach towards improving primary task objective [66], we aim to decode the ground truth character sequence $\hat{Y}$ directly from $z$ by using an auxiliary character sequence decoder. Firstly, this approach tackles the problem of vanishing latent variable [6] and provides better gradient to regularize the learning of feedback module. Secondly, we discover a choice of training our feedback module (refer Figure 2) via relativistic information of prediction:ground-truth pairs (say $G$) generated from other state-of-the-arts, instead of depending on $T$ predicted $Y$ alone. This helps the module in learning an associative relation between candidate correct words and closely related erroneous instances. Loosely speaking ‘hello’ might be predicted as ‘nello’ or ‘bello’ due to a partial structural similarity of ‘h’ with ‘n’ or ‘b’. Learning this sense from error distribution imparts the model a semantic knowledge of required character association to form a valid word. Additionally, using text-only data $G$ alleviates the issue of limited availability of image-paired datasets during training.

Hence, we adapt Eqn. 1 for our feedback network to generate a prior knowledge $P^*$, instead of transformation parameters only. This prior knowledge has two components. One encapsulates relationship between $Y$ and $\hat{Y}$, while the other generates $P$ and injects $Y$ into $T$ for next prediction. Therefore adopting the aforementioned variational lower bound expression (Eqn. 1), and assuming conditional independence of those two knowledge components (given $z$ and $Y$) denoted by $p(P^*|z,Y) = p(P|z,Y)p(\hat{Y}|z,Y)$, we get a modified lower bound as:

$$
\tilde{L}'(Y, P^*; \theta, \psi) = -KL(q_\psi(z|\hat{Y},Y)||p_\theta(z|Y)) + \\
\mathbb{E}_{q_\psi}[\log p(P|z,Y)] + \mathbb{E}_{q_\psi}[\log p(\hat{Y}|z,Y)]
$$  

(ii) We assume that posterior of $z$ depends on the actual ground-truth label $Y$ instead of $P$, so $q_\psi(z|P,Y) \approx q_\psi(z|\hat{Y},Y)$ which effectively makes the latent space $z$ aware of exact ground-truth character sequence $Y$. 

Informally spoken, our goal is to model a conditional distribution over the activations at each layer. Compared to global tuning [41], such local harmonisation of features ensure a better fine-grained control over the activations at each layer.

Feedback Network: Our goal is to model a conditional distribution $p(P|Y)$ from a discrete predicted character space $Y$ to the transformation parameter space $P$ of feedback conditioning layer via a feedforward network based on CVAE [51]. Sohn et al. [51] has shown that the variational lower bound of this conditional distribution can be written as:

$$
\tilde{L}(Y,P; \theta, \psi) = -KL(q_\psi(z|P,Y)||p_\theta(z|Y)) + \\
\mathbb{E}_{q_\psi}[\log p(P|z,Y)] \leq \mathbb{E}_{q_\psi}[\log p(P|Y)]
$$  

(i) Ideally speaking, $P$ is supposed to adjust the activation map $\Psi$ in a way that predicts $\hat{Y}$ when passed through a fixed $T_B$. In other words, for time-step $t$, given $P^t$ predicted by the feedback network with a conditioning on $Y^{t-1}$, the modulated feature map $\hat{\Psi}^t$ is fed into $T_B$ obtains the output of next iteration $Y'$. Therefore, $Y' = T_B(\hat{\Psi}^t)$, where $\hat{\Psi}^t = \Phi(\Psi; P^t)$. Thus, minimizing the cross-entropy loss between $Y'$ and $\hat{Y}$ is equivalent to maximizing the likelihood of $P$. 

Figure 2. Training of the network consists of 2 steps: (i) Predicting $Y$ (green arrow); (ii) Refining $Y$ to $Y'$ using feedback mechanism. Feedback module excluding $F_P$ can be trained using additional text-only data $G$ (red). However, when trained using prediction $Y'$, the entire path is followed (red+blue).
Network Components: Ignoring the time-step notation, let us consider any word image \( I \) having ground-truth \( \tilde{Y} \), that undergoes the first forward-pass of \( T \) predicting \( Y \) and \( Y' \) in current and successive post-feedback iterations respectively. During training, we obtain an embedding representation of both \( \tilde{Y} \) and \( Y \) through a shared encoder network \( \mathbf{F}_{\text{enc}}: Y_{\text{enc}} = \mathbf{F}_{\text{enc}}(Y) \) and \( \tilde{Y}_{\text{enc}} = \mathbf{F}_{\text{enc}}(\tilde{Y}) \). Moving on we have two independent branches: A posterior network \( \mathbf{F}_{\text{post}} \) to estimate parameters of posterior distribution, and a prior network \( \mathbf{F}_{\text{prior}} \) to do the same for prior distribution. Following this, we have: \( \mu_{\text{post}}, \sigma_{\text{post}} = \mathbf{F}_{\text{post}}([Y_{\text{enc}}, \tilde{Y}_{\text{enc}}]) \) and \( \mu_{\text{prior}}, \sigma_{\text{prior}} = \mathbf{F}_{\text{prior}}(Y_{\text{enc}}) \). A latent variable \( z \) is sampled from the posterior (or prior during testing) distribution and merged with \( Y_{\text{enc}} \) before feeding it to the auxiliary ground-truth character sequence decoder \( \mathbf{F}_{\text{aux}} \) and transformation parameter prediction sub-module \( \mathbf{F}_{\mathcal{P}} \). Therefore we have, \( \mathcal{P} = \mathbf{F}_{\mathcal{P}}([Y_{\text{enc}}, z]) \) and \( \hat{Y} = \mathbf{F}_{\text{aux}}([Y_{\text{enc}}, z]) \) where \( \mathcal{P} \) encapsulates both \( \Gamma \) and \( \Omega \), and \( \hat{Y} \) is output from the auxiliary decoder \( \mathbf{F}_{\text{aux}} \). Thereafter, we evaluate the prediction of next iteration as: \( \hat{Y}' = \mathbf{T}_B(\hat{Y}) \), where \( \Psi = \Phi(\Psi; \mathcal{P}) \). Please refer to Figure 2 for clarity.

Learning Objectives: Our baseline text recognition network \( T \) (parameters \( \theta_T \)) is trained using Cross-Entropy (ce) loss summed over the ground-truth output sequence \( \hat{Y} \)

\[
L^\theta_T = L_{ce}(Y, \hat{Y}) = - \sum_{k=1}^{K} \hat{y}_k \log p(y_k | \hat{y}_k^{-1}) \quad (3)
\]

Feedback module \( F \) is to be trained from two input sources: (a) by using the prediction \( Y \) obtained from \( T \). In cases where the forward-pass prediction \( Y \) is accurate the very first time, a diverged prediction value in the next iteration is highly undesirable. Hence, along with the lower bound (Eqn. 2), we impose a monotonically decreasing constraint \( L_c \). This enforces the loss value (Eqn. 3) related to the current iteration \( \{Y', \hat{Y}' \} \) to be less than its previous \( \{Y, \hat{Y} \} \), thus converging predictions to a higher precision. Hence, we optimize all the parameters \( \theta_F \) of feedback module using:

\[
L^\theta_F = \lambda_1 L_{ce}(Y', \hat{Y}') + \lambda_2 L_{ce}(Y, \hat{Y}) + \lambda_3 L_{KL} + \lambda_4 L_c
\]

where, \( L_c = \max(0, L_{ce}(Y', \hat{Y}') - L_{ce}(Y, \hat{Y})) \) \quad (4)

(b) By using pre-collected text-only data \( G \) generated from existing text-recognition methods \([30, 10, 2, 50]\). Doing so develops a semantic sense of associative relation between candidate correct words and erroneous alternatives. Keeping \( F_P \) fixed, rest of the feedback module (parameters denoted by \( \theta_F \)) is optimised solely using auxiliary decoder reconstruction loss and KL divergence loss as:

\[
L^\theta_F = \lambda_2 L_{ce}(\hat{Y}, \tilde{Y}) + \lambda_3 L_{KL}
\]

During testing we exclude both \( \mathbf{F}_{\text{aux}} \) and \( \mathbf{F}_{\text{post}} \), where iterative prediction is done using the transformation parameter predicted by \( \mathbf{F}_P \) (see Figure 2).

Comparative Discussions: While state-of-the-art text recognition frameworks model a conditional distribution \( p(Y|I) \), our modelling objective is \( p(Y'|Y^{-1}, I) \). This could be decoupled into two marginal distributions related via a transformation parameter space \( P^t \) (more strictly prior knowledge \( P^t \)). Assuming the obvious conditional independence, we reformulate it as: \( p(Y'|Y^{-1}, I) = p(Y'|I, \mathcal{P}') \times p(\mathcal{P}'|Y^{-1}) \), where \( p(Y'|I, \mathcal{P}') \) is our modified feed forward text recognition model and \( p(\mathcal{P}'|Y^{-1}) \) resembles our feedback network.

Algorithm 1 Training algorithm of the proposed framework

1: Input: Image and ground-truth pairs \( D \); prediction and ground-truth pairs \( G \).
2: Initialise hyper params: \( \alpha_1, \alpha_2 \) learning rate for \( T, F \) respectively.
3: Initialise model params: \( \theta_T, \theta_F \) (\( \theta_F \subset \theta_T \) excluding \( F_P \)).
4: while not done training do
5: Sample a mini-batch \( D_i \) from \( D \)
6: Update \( \theta_T := \theta_T - \alpha_1 \nabla_{\theta_T} (L^\theta_T) \quad \triangleright \text{Eqn. 3} \)
7: Update \( \theta_F := \theta_F - \alpha_2 \nabla_{\theta_F} (L^\theta_F) \quad \triangleright \text{Eqn. 4} \)
8: Sample a mini-batch \( G_i \) from \( G \)
9: Update \( \theta_{F'} := \theta_{F'} - \alpha_2 \nabla_{\theta_{F'}} (L^\theta_{F'}) \quad \triangleright \text{Eqn. 5} \)
10: end while
11: Output: \( \theta_T, \theta_F \).

4. Experiments

Datasets: We train our method using an approach similar to \([62, 58, 2, 10, 50, 30]\), on synthetic datasets such as Synth90k \([19]\) and SynthText \([14]\) that hold 8 and 6 million images respectively. Evaluation is done on: IIIT5K-Words, Street View Text (SVT), SVT-Perspective (SVT-P), ICDAR 2013 (IC13), ICDAR 2015 (IC15), and CUTE80. IIIT5K-Words \([34]\) presents randomly picked 3000 cropped word images. Street View Text \([53]\) contains 647 images, mostly being blurred, noisy or having low resolution. SVT-Perspective \([44]\) offers 645 samples from side-view angle snapshots having perspective distortion. ICDAR2013 \([24]\) presents 1015 words while ICDAR2015 \([23]\) has 2077 images, 200 of which are irregular. CUTE80 \([45]\) distinguishes itself by presenting 288 cropped high quality curved text images. Handwritten Text Recognition (HTR) and Scene-Text Recognition (STR) both share a common objective in terms of recognition, which is usually handled by similar network architecture. Hence, we validate our results on two different public HTR datasets. The evaluation setup described in \([4]\) is employed on two large standard datasets viz, IAM \([33]\) containing 1,15,320 words and RIMES having 66,982 words. For IAM we use the same partition for training, validation and testing as provided. For RIMES, we follow the partition released by ICDAR 2011 competition.
Table 1. Comparison of unconstrained WRA for novel words not encountered during training. $t=0$ signifies no feedback.

<table>
<thead>
<tr>
<th>Methods</th>
<th>BICSK</th>
<th>SVP</th>
<th>TIC3</th>
<th>TIC5</th>
<th>SVP</th>
<th>LAM</th>
<th>TIMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shi et al. [50] (t=0) No-Feedback</td>
<td>84.1</td>
<td>84.2</td>
<td>82.6</td>
<td>65.7</td>
<td>74.4</td>
<td>61.6</td>
<td>54.1</td>
</tr>
<tr>
<td>Baseline Seq-SCM</td>
<td>85.6</td>
<td>84.1</td>
<td>83.7</td>
<td>65.5</td>
<td>75.8</td>
<td>63.4</td>
<td>57.6</td>
</tr>
<tr>
<td>Baseline Deterministic-Feedback</td>
<td>87.9</td>
<td>86.8</td>
<td>85.9</td>
<td>70.4</td>
<td>78.6</td>
<td>64.7</td>
<td>59.9</td>
</tr>
<tr>
<td>Shi et al. [50] + CVAE-Feedback</td>
<td>90.6</td>
<td>88.7</td>
<td>89.3</td>
<td>72.2</td>
<td>79.6</td>
<td>65.1</td>
<td>64.5</td>
</tr>
<tr>
<td>Shi et al. [50] + CVAE-Feedback</td>
<td>90.8</td>
<td>88.9</td>
<td>89.4</td>
<td>72.6</td>
<td>79.6</td>
<td>65.1</td>
<td>64.5</td>
</tr>
<tr>
<td>Shi et al. [50] + CVAE-Feedback</td>
<td>90.8</td>
<td>88.9</td>
<td>89.4</td>
<td>72.6</td>
<td>79.6</td>
<td>65.1</td>
<td>64.5</td>
</tr>
<tr>
<td>Shi et al. [50] + CVAE-Feedback</td>
<td>90.7</td>
<td>88.8</td>
<td>89.4</td>
<td>72.5</td>
<td>79.6</td>
<td>65.8</td>
<td>64.6</td>
</tr>
</tbody>
</table>

Relative Gain (t=0 vs t=2)

<table>
<thead>
<tr>
<th>Methods</th>
<th>6.5↑</th>
<th>4.7↑</th>
<th>6.8↑</th>
<th>6.9↑</th>
<th>5.2↑</th>
<th>4.5↑</th>
<th>10.5↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Show, Attend and Read [27] (t=0) No-Feedback</td>
<td>85.8</td>
<td>86.5</td>
<td>84.7</td>
<td>68.4</td>
<td>82.2</td>
<td>71.8</td>
<td>57.9</td>
</tr>
<tr>
<td>Show, Attend and Read [27] + CVAE-Feedback (t=2)</td>
<td>91.5</td>
<td>90.5</td>
<td>91.2</td>
<td>74.8</td>
<td>87.1</td>
<td>75.0</td>
<td>68.0</td>
</tr>
</tbody>
</table>

Relative Gain (t=0 vs t=2)

<table>
<thead>
<tr>
<th>Methods</th>
<th>5.7↑</th>
<th>4.0↑</th>
<th>6.5↑</th>
<th>6.4↑</th>
<th>4.9↑</th>
<th>3.2↑</th>
<th>10.1↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCATTER [29] (t=0) No-Feedback</td>
<td>84.7</td>
<td>86.9</td>
<td>84.3</td>
<td>71.8</td>
<td>82.6</td>
<td>69.3</td>
<td>59.0</td>
</tr>
<tr>
<td>SCATTER [29] + CVAE-Feedback (t=2)</td>
<td>91.1</td>
<td>90.9</td>
<td>90.0</td>
<td>77.7</td>
<td>87.3</td>
<td>72.7</td>
<td>68.7</td>
</tr>
</tbody>
</table>

Relative Gain (t=0 vs t=2)

<table>
<thead>
<tr>
<th>Methods</th>
<th>6.4↑</th>
<th>4.0↑</th>
<th>6.6↑</th>
<th>5.9↑</th>
<th>4.7↑</th>
<th>3.4↑</th>
<th>10.2↑</th>
</tr>
</thead>
</table>

4.1. Implementation Details

**Network Design:** Keeping text rectification and recognition networks similar to Shi et al. [50], we implement our framework using open sourced codes [2, 30] in PyTorch [40]. A bi-directional LSTM of hidden size 256 is used in designing $F_{enc}$ of our feedback module which accepts a one layer MLP embedded 128-dimensional representation of discrete character sequences. For both posterior $F_{post}$ and prior $F_{prior}$ networks, we use 2 layer MLPs with tanh non-linearity. The latent variable $z$ has a size of 256. The auxiliary decoder $F_{aux}$ is a one layer LSTM decoder whose initial hidden state is initialized by applying a FC-layer on the concatenated representation of $Y_{enc}$ and sampled $z$. The parameter prediction network $F_p$ is a convolutional decoder network inspired from [67]. The first layer is a fully-connected layer implemented through a 1x1 convolution that maps to a tensor of size equal to the last CNN layer of $T$ via a reshaping operation. Thereafter we introduce a sequence of residual decoder blocks which upsamples feature maps to higher spatial dimensions in the reversed order of down-sampling followed by $T$. In other words, if a ResNet encoder block in $T$ halves the height of feature map, the corresponding ResNet decoder block [67] in $F_p$ will double it. This strategy essentially formulates intermediate feature maps of a decoder ResNet decoder block ($F_p$), whose spatial size is similar to the corresponding layer in the encoder ResNet block ($T$), thus predicting required affine transformation parameters of similar dimensions as well. Since we need to predict both $\Gamma$ and $\Omega$, we double the number of convolutional filters in the last layer of $F_p$ and split the output channel-wise to obtain $\Gamma$ and $\Omega$ individually.

**Training Details:** It has been observed in practice, that warming up individual components in the initial phase, followed by a joint training operation provides better stability than training the entire framework at one go. Text recognition network ($T$), along with rectification network, is trained using ADADELTA optimizer having a learning rate of 1.0 and a batch size of 64. Meanwhile Feedback module $F$ is trained via Adam optimizer with a learning rate 0.001, and gradient clipping at 5. In the warm-up phase, firstly, text recognition and rectification networks are trained from a union of MJSynth and SynthText datasets, for 600K iterations. The rectification network is frozen thereafter. Then, to capture a linguistic prior from text-data, the feedback module $F$ is trained independently from $G$ for 300K iterations, thereby ignoring the $F_p$ part (Eqn. 5). Finally, keeping $T$ fixed, we train the complete feedback module using $Y$ (Eqn. 4), with the same training specifications for 300K iterations. Now, for the joint training (see algorithm 1), after updating $T$, $F$ is updated with the prediction $Y$ from $T$, as input. Thereafter, keeping $F_p$ fixed, $F$ is updated using $G$. During this training, we reduce the learning rate of $T$ to 0.01 which continues for 600K iterations. Owing to corresponding data sizes in HTR, iterations for warm-up phases are 100K, 50K and 50K respectively, while for joint training it is 100K. For both STR and HTR, we resize the image to 32x100 and train our model in a 11 GB Nvidia RTX 2080-Ti GPU. $\lambda_1, \lambda_2, \lambda_3$ and $\lambda_4$ has been assigned the values of 1, 1, 50 and 0.5 respectively for the same purpose. In order to generate the text-only data $G$, we follow an approach similar to [68]. The entire training dataset is divided into train-validation split in a cross validation setting, where top ten beam search decoded hypotheses are collected from different state-of-the-art (SOTA) models [30,10,2,50], to harness (using open-source code) the error information from closest possible erroneous alternatives. Such a collection may consist of correctly or incorrectly predicted words. This requires the feedback module to learn a language prior for fixing a wrongly predicted word, as well as for a one-to-one character-sequence mapping for correctly predicted words.

4.2. Performance Analysis

**Baselines:** Aligned to our iterative approach we explore two alternatives as baselines. Seq-SCN: Inspired from Automatic Speech Recognition community [56], a naive baseline could be designed where we train an independent Spelling Correction Module (SCM) based on sequence-to-sequence architecture from text-only training data [13] consisting of paired model-hypothesis and corresponding ground-truth. Deterministic-Feedback: Here, we simply replace the CVAE based feedback module by a deterministic encoder(bi-LSTM)-decoder (parameter prediction network) architecture along with Shi et al. [50].
New Evaluation Setup: Here we design a new disjoint train-test split (DS) in addition to conventional train-test split (CS). While training we remove all words from MJSynth and Synth90K whose ground-truth pairs appear in any of the mentioned STR testing datasets. Additionally we ensure, \( G \) does not contain any information from the testing set. Due to size constraint in HTR datasets, we split it such that all word-image pairs corresponding to a particular ground-truth character sequence must collectively fall in either one of training or testing set, thus ensuring disjointedness. This evaluation protocol grades our model’s recognition performance on word images whose ground-truth character sequences never appeared in the training dataset. We use it to verify generalising capability of our model for unseen or rarely seen words. Furthermore, superiority of our model in this scenario confirms a fair result on datasets having less unique words. As large datasets are rarely available for text recognition (specifically handwriting) apart from English, it is ideal to use a dataset created by collecting multiple instances of a small set of unique words which thus need to be annotated just once per unique word. To do so any model needs to learn character-specific fine-grained details from a small available set of unique words, and generalise onto other unseen character sequences. Please note that we re-train SOTA models [30,10,2,50], used to collect the error distribution, by ensuring no words from evaluation set appear in their training (ensuring no “leak” of information), while collecting error distribution for DS setup.

Result Analysis: Along with the designed baselines, we incorporate our iterative design on top of three popular state-of-the-art (SOTA) feed-forward text recognition frameworks – a) Shi et al. [50], b) Show, Attend and Read [27] c) SCATTER [29]. Show, Attend and Read [27] extends [50] by including 2D attention and SCATTER [29] couples multiple BLSTM layers for richer context modelling. We follow similar training protocol in [50,27,29] respectively. Nevertheless, ours is a meta-framework and could be added on top of most SOTA frameworks. Table 1 (highest scores are in red) depicts unconstrained word recognition accuracy (WRA) on unseen words (DS setup).

Comparison with Baselines: Seq-SCN performs inferior to our method as the rich visual feature is not harnessed while refining the prediction. At times it fails to copy an already accurate prediction, leading to a lower accuracy in certain datasets. Deterministic-Feedback being an iterative framework, performs better than other baselines at \( t = 2 \), however, it lags behind our design since any uncertainty handling potential is absent. [ii] Significant improvement under DS setup: From Table 1, improvement due to our iterative pipeline is quite evident in the DS scenario over three SOTA baselines [50,27,29]. Similarly, in handwritten dataset (Table 1), the performance drop in DS setting is way more severe than its STR counterpart. This signifies that HTR poses a greater challenge due to its free-flow nature of writing. Improvement against Shi et al. [50] reaches to 10.5% and 10.8%, without lexicon information (unconstrained) in HTR dataset of IAM and RIMES respectively, while for STR dataset of IC15 it secures a 6.9% rise in DS setup. [iii] Additional Observations: Improvement at \( t=2 \) w.r.t \( t=0 \) is shown as relative gain where our method outperforms fairly in CS setup, and largely in DS setup (Table 2). In fact optimal WRA is seen at iteration \( t = 2 \), which then diminishes. Contrary to a possible impression, that our feedback module might remember the erroneous pairings from \( G \), the improvement on “unseen” words empirically validates against it. Figure 3 shows qualitative results.

4.3. Further Analysis and Insights

Ablation Study: We have done a thorough ablative study to justify the contribution of every design choice on both
IAM (HTR) and IC15 (STR) datasets using Shi et al. [50] as baseline. [i] Significance of auxiliary decoder: On removing this part along with its respective loss function from Eqn. 4 we see a performance drop of 2.19%(1.98%) and 5.01%(2.94%) in the CS and DS setting for IAM (IC15) dataset respectively, thus confirming its contribution. [ii] Significance of monotonically decreasing constraint: Removing it, destabilizes by 5.47% (3.87%) in DS setup, for IAM (IC15) datasets respectively, thus confirming its importance. [iii] Significance of using error distribution: Discarding data inclusion from $G$ (error distribution), leads to a performance drop of 2.17% (1.87%) in conventional train-test split, and a further drop of 4.9% (2.48%) in disjoint setup, for IAM (IC15) datasets. Please refer to Table 4 for more analysis. [iv] Finding optimal block for feedback: We evaluated the performance by providing a feedback signal into every ResNet block of backbone feature extractor of T at a time. Table 3 shows a complete analysis on both IAM and IC15 datasets in CS and DS setups which shows the optimum result to be obtained by supplying feedback signal to Block-3 of ResNet convolutional architecture. Moreover, local transformation is seen to outperform the global one. [v] Performance with varying text length It is often observed that any text recognition framework struggles to recognize lengthy words. Owing to its iterative refining approach, along with a modelled linguistic prior, our method shows a considerably higher performance in comparison to no-feedback baseline on increasing character sequence length, as shown in Figure 3. [vi] Computational cost: Finally, we want to notify that benefit of any iterative pipeline [7, 8, 28] does incur extra computational expenses, be it text rectification [62] or in our case text-recognition. A thorough study on complexity and speed analysis (Intel(R) Xeon(R) W-2123 CPU @ 3.60GHz) in Table 5 reveals that $T_{f}$ takes most time due to its sequential decoding operation, whereas $F$ adds minimal burden. Please refer to supplementary as well.

Table 3. WRA of using feedback after a specific ResNet Block (abbreviated as ‘Blk’) such as Block 1 (16 × 50 × 64), Block 2 (8 × 25 × 128), Block 3 (4 × 25 × 256), Block 4 (2 × 25 × 256), Block 5 (1 × 25 × 256) as described in ASTER [50].

<table>
<thead>
<tr>
<th>Methods</th>
<th>Convolutional Setup (CS) &amp; Disjoint Setup (DS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CS</td>
</tr>
<tr>
<td>Block 1</td>
<td>84.7</td>
</tr>
<tr>
<td>Block 2</td>
<td>84.9</td>
</tr>
<tr>
<td>Block 3</td>
<td>78.5</td>
</tr>
<tr>
<td>Block 4</td>
<td>78.5</td>
</tr>
<tr>
<td>Block 5</td>
<td>82.3</td>
</tr>
</tbody>
</table>

Feedback module vs Language Model: We could have substituted our iterative approach towards refining text prediction with a Language model(LM). For fairness, we use a state-of-the-art RNN-LM [13] trained from text corpus (librispeech) at character level [1] that aims to predict the next likely character. This could be fused with the text recognition decoder using two state-of-the-art methods introduced in [12] via Shallow Fusion (weighted sum of predicted scores) and Deep Fusion (fusing their hidden states). Our method performs better in both CS and DS setups in comparison to these LM integrations (Table 6). The limited performance of off-the-shelf LM [22] can be attributed to: (i) LM is mostly used in speech recognition tasks, where data is present at sentence-level which provides enough context. For distract word recognition (our focus) however, LM cannot harness such extent of context information. (ii) The LM corpus is significantly different from that used for training word-image recognition system. This leads to a biased incorrectness [13]. (iii) LM being an independent post-processing step, not only ignores rich visual features from the input image, but is also unaware of the error distribution of the model. On the contrary, our model revisits the rich visual features iteratively after every prediction, considering the error distribution while training. Furthermore to align with the evaluation standards for unconstrained word recognition we cite all results in our work using greedy decoding only – no LM based post-processing.

Table 4. Relative contribution (WRA) of three design choices behind training Feedback Network for iterative prediction: (a) Regularisation constraint via $L_C$ (Eqn. 4), (b) Modified variational lower bound for feedback network accommodating $F_{aux}$, (c) Using $G$ to capture the prior-knowledge from error distribution.

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Auxiliary Decoder $F_{aux}$</th>
<th>Use Error Dist. $G$</th>
<th>CS Setup</th>
<th>DS Setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_C$</td>
<td>✓</td>
<td>✓</td>
<td>76.8</td>
<td>85.3</td>
</tr>
<tr>
<td>$L_C$</td>
<td>✓</td>
<td>✓</td>
<td>76.6</td>
<td>84.5</td>
</tr>
<tr>
<td>$L_C$</td>
<td>✓</td>
<td>✓</td>
<td>76.9</td>
<td>85.3</td>
</tr>
<tr>
<td>$L_C$</td>
<td>✓</td>
<td>✓</td>
<td>78.8</td>
<td>87.5</td>
</tr>
</tbody>
</table>

Table 5. Complexity and speed analysis against no. of parameters and flops (Multiply-Add), for both individual component (left) and varying no. of iterations (right), using CPU time during inference.

Table 6. Comparison with different LM integration methods.

5. Conclusion

Here we have proposed a novel iterative approach towards text recognition. Using a conditional variational autoencoder (CVAE) as a feedback module, the knowledge of predicted character sequences is passed from the previous iteration, into the main recognition network, improving subsequent predictions. Our feedback network learns to use the error distribution among multiple character sequences that are closely related to a candidate word. Experiments on various STR and HTR datasets show our network to outperform others on the conventional setting, and more significantly on the more practical disjoint (unseen) setting.
References


[8] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In CVPR, 2018. 3, 8


[27] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In AAAI, 2019. 6, 7


