Supplementary material for
Text is Text, No Matter What: Unifying Text Recognition using Knowledge Distillation

Ayan Kumar Bhunia[1] Aneeshan Sain[1,2] Pinaki Nath Chowdhury[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{a.bhunia, a.sain, p.chowdhury, y.song}@surrey.ac.uk.

A. Motivation of Unified HTR-STR model

While modality-specific training has its own benefits, there are a few scenarios which require simultaneous recognition of both modalities – recognising hand-written road signs, posters, graffiti, or some scene-text images pasted on handwritten documents. Our work is motivated at a high level by the philosophy of *general AI* where the goal is to develop a single model handling multiple purpose, such as solving multiple tasks [4, 5, 10] via multitask learning, working over multiple domains [2, 7], and employing universal adversarial attack [6]. Furthermore, this work paves the way towards a unified text recognition paradigm, which has the potential of significantly reducing the extended effort of training models separately.

B. Generalizability across different architectures

We explore the generalisation potential of our method across different architectures within the attention decoder based paradigm. **A**ttentional **D**ecoder **(AD)** is the de-facto state-of-the-art choice over CTC loss [3] based alternatives due to their better overall accuracy via modeling an implicit language model [9]. Handling non-identical student-teacher networks may be challenging due to: **(a)** *feature dimension mismatch* for character localized hint loss – solved by a learnable linear embedding layer [8], or **(b)** *spatial size mimatch* for attention map – that requires a differentiable bilinear-interpolation or learnable up-sampling/down-sampling layer. Nevertheless, for any auto-regressive decoder based architecture, the affinity matrix $\mathcal{A}_{i,j}$ could be calculated using the glimpse vectors $\{g_1, g_2, \dots, g_K\}$ as shown in Eq. 8.

We further scrutinize the generalisation potential for our method across different architectures by employing the same teacher network like ours but using the popular ASTER [9] model with 1D attention as student. While the specialised teacher results in 82.3% (HTR) and 74.8% (STR), our unified student performs 82.4% (HTR) and 74.8% (STR).

C. More details on experimental setup and analysis:

(i) End-to-end training with binary classifier would give an overly complicated baseline that needs thorough hyper-parameter tuning to optimise as it involves a *non-differentiable* operation like selection (e.g., via Gumble-softmax). Hence, in this study use the binary classifier to only select the specific model between HTR and STR for text recognition.

(ii) We ensured that hyper-parameters of all the baselines are optimized against the same validation set. For HTR we use the standard validation split within the dataset, while for STR we use the protocol adopted by [1].

(iii) For Binary-Classifier based two-stage alternative, instead of restricting ourselves to standard classifiers readily available in PyTorch – ResNet18 being the simplest there, we further compare with alternative simpler binary-classifier. A 3 layer CNN with hidden size 128, having 0.3M parameters, and 0.03 GFlops for $84\times84\times3$ input – typically used for few-shot classification in mini-imagenet dataset. We find similar acccuracy of 74.1% on STR-IC15 and 82.8% on HTR-IAM - which is slightly lower than ours.

References

[1] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *ICCV*, 2019. 1

[2] Hakan Bilen and Andrea Vedaldi. Universal representations: The missing link between faces, text, planktons, and cat breeds. *arXiv preprint arXiv:1701.07275*, 2017. 1

[3] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006. 1

[4] Lukasz Kaiser, Adian N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017. 1

[5] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017. 1

[6] Hong Liu, Rongrong Ji, Jie Li, Baochang Zhang, Yue Gao, Yongjian Wu, and Feiyue Huang. Universal adversarial perturbation via prior driven uncertainty approximation. In *ICCV*, 2019. 1

[7] Sylvester-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *NeurIPS*, 2017. 1

[8] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antonie Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 1

[9] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai. Aster: An attentional scene text recognizer with flexible rectification. *TPAMI*, 2018. 1

[10] Amir R Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 1