Supplementary materials for
”Inkorrect: Digital Ink Spelling Correction”

Contents

1 Contents of the archive

The set of supplementary materials for the submission includes this file and other files in the *supplementary.zip* archive. We list their contents below.

- *deepwriting_data_prepare_external.py*: run to obtain a set of data needed to run the spelling correction models.
- *deepwriting_model_run_external.py*: run to generate the results of **DW** approach.
- *deepwriting_data_lib_external.py* and *deepwriting_model_lib_external.py* are the libraries needed to run the two scripts above.
- *deepwriting_valid_corrected_labels.txt*: the spell-corrected labels for the samples in the Deepwriting validation dataset.
- *user_study.zip*: Archive with images from the user study, and the sample responses.

All of the .py files can be run as:

```
python3 filename.py <args>
```

To list the required arguments, run:

```
python3 filename.py --help
```

Running the DW model requires having access to the Deepwriting dataset and pre-saved model[1]. The *deepwriting_data_lib_external.py* and *deepwriting_model_lib_external.py* are a version of the original code of the DW approach (available at the same address), modified by the authors of Inkorrect.

Due to the size limitation of the supplementary materials (50mb), we are unable to include all of the user study images covering the **DeepWriting** validation dataset, therefore we include the first 250. We include all responses from all users.

2 CDE

As mentioned in the main paper, for two sets of points P and Q representing the inks, sorted by their *x-axis*:

$$
\begin{aligned}
&CD(P,Q) = \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}||p-q||_2 + \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}||p-q||_2 \\
&CDE(P,Q,K) = \min_{(P_1,\ldots,P_k),(Q_1,\ldots,Q_k)}\sum_{i=1}^{K} CD(P_i,Q_i) \\
&CDE(P,Q,K) = F^K_{|P|,|Q|} \\
&F^k_{i,j} = \min_{l<i,m<j} F^{k-1}_{l,m} + CD(\{p_l,\ldots,p_i\},\{q_m,\ldots,q_j\})
\end{aligned}
$$

The boundary conditions for F are:

1. $F^0_{0,0} = 0$
2. $F^0_{i,j} = \infty\ \forall i > 0$
3. $F^0_{i,j} = \infty\ \forall j > 0$
4. $F^k_{0,0} = \infty\ \forall k > 0$

2.1 Opimizing CDE

Listing below contains a sample implementation of CDE.

```python
import numpy as np

def CD(P: np.npdarray, Q: np.ndarray) -> float:
  """Sample CD implementation for [Nx2] and [Mx2] numpy arrays."""
  pass

def CDE(P: np.pdarray, Q: np.ndarray, K: int) -> float:
  # Ensure the correct order of points.
  P = np.sort(P, axis=0)
  Q = np.sort(Q, axis=0)

```

[1] https://ait.ethz.ch/projects/2018/deepwriting/

```
# Initialize F.
N, M = len(P), len(Q)
F = np.zeros((N + 1, M + 1, K + 1)) + np.inf
F[0, 0, 0] = 0

# Optimization.
for k in range(K):
  for i in range(N):
    for j in range(M):
      if F[i, j, k] == np.inf:
        continue
      # Transition step: consider all consecutive subsegments.
      for next_i in range(i+1, N+1):
        for next_j in range(j+1, M+1):
          F[next_i, next_j, k+1] = min(
              F[next_i, next_j, k+1], F[i, j, k] + CDO(P[i:next_i, j:
  next_j]))

# Allow splitting into less than K parts in case there are less then
  K points.
return np.min(F[N, M, :K+1]
```

Listing 1: Sample CDE implementation

2.2 Selection of K

As we mention in the main paper, we select K to be the number of words in the original label plus the edit distance between the original and spell-corrected label, because the two main sources of the ink misalignment are shifts between the words and shifts within the words due to spelling corrections. In fact, a better estimate for K could be obtained by replacing edit distance by the number of places within the words where inks have different letters (so, for example, if two consecutive letters were replaced, that would incur the edit distance of 2, but in fact K=1 would suffice). In practice, we found virtually no difference between these two approaches, probably because most samples had an edit distance of 1. Since edit distance between strings is a well-known entity, and computing number of places where inks don't align requires additional implementation, we decided to use edit distance.

3 User study

Figure 1 shows the setup of the user study.

The open-ended feedback from the participants on the criteria they used is listed below:

1. 1. Recognizability. 2. Size. 3. Angle. 4. Connectivity. 5. Shape of the individual letters.

2. Many times my decision was based on which synthesis actually spelled all the words correctly and had all letters readable. Only if both satisfied this then I looked at the style.

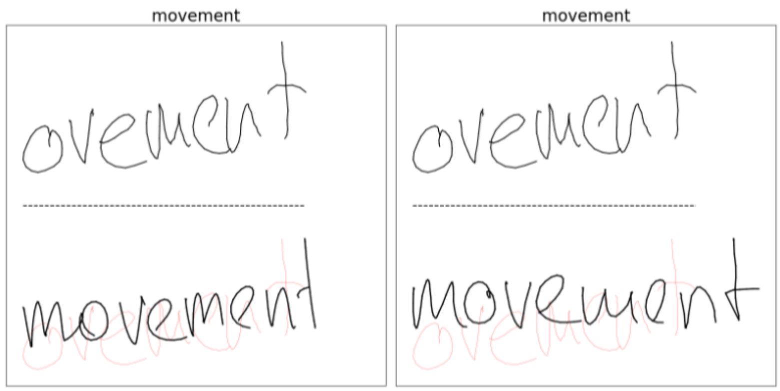

Figure 1: Example of the user study. Users were shown the original ink twice (on top), and spell-corrected versions from each model (order randomized). The spell-corrected versions were superimposed on the original ink (in thin red) to help the user make a better judgement about which spelling-corrected version is best. The labels on top are the spell-corrected labels, to help understand whether the spell-corrected ink matches the spell-corrected label. Users selected the left or right option by clicking.

3. Main criteria for me were similarity to style of original ink and more reasonable positioning relative to the original ink. In addition to size and angle, also the overall shape of the letters / connectedness / etc. And for position mainly the unchanged words, but also reasonable positioning for the spell-corrected ones.

4. 1. Is the ink readable, does it match the target label? (that was enough to decide on more than half of the samples). 2. Closeness to original style. 3. Closeness to original size.

5. 1. Correctness - matching the labels. 2. Delayed strokes present if also present in original. 3. Matching the original styles. 4. Readability

6. 1. Recognizability. 2. Ease of reading (prefer the one I can parse faster, even if I can parse both). 3. Similarity of angle, size, and cursiveness.

7. 1. Recognizability. 2. Overall closeness.

8. 1. How easy it was to read. 2. How close it was to the original style. 3. How pretty it is.

9. 1. How easy is it to read and parse; 2. If both can be easily parsed, prefer one with no discrepancy in height.

10. 1. Ink could be parsed as the spell-corrected label; 2. Connectedness similar to the original ink; 3. Size and angle similar to the original ink.

4 Additional results

Figures 3 and 2 contain some additional results for the two datasets used.

Figure 2: Additional results on **HANDS-VNOnDB** dataset. Label on the left if the original label, on the right is the spell-corrected label.

Figure 3: Additional results on **DeepWriting** dataset. Label on the left is the original label, on the right is the spell-corrected label.