Data Incubation — Synthesizing Missing Data for Handwriting Recognition

Rick Chang, Martin Bresler, Youssouf Chherawala, Adrien Delaye, Thomas Deselaers, Ryan Dixon, Oncel Tuzel
ICASSP 2022 · Apple Inc.

#0000 INSERT POSTER ID

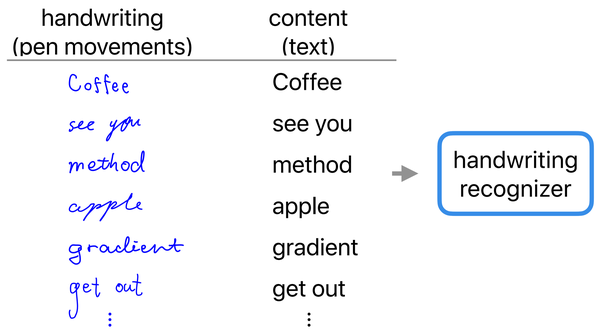
Abstract

Can we generate more training data from existing ones?

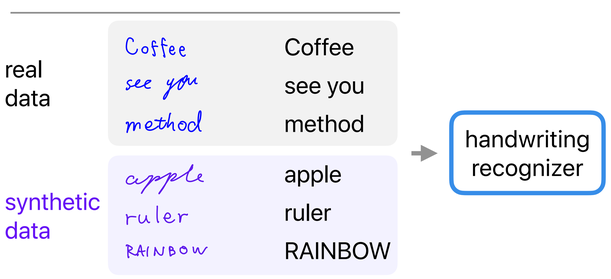
- We use an observation that handwriting data distribution can be factorized into **content** and **style**
- By learning a controllable generative model, we can fill in the missing content and style

ERM vs. Data incubation

Empirical Risk Minimization (ERM)
directly train handwriting recognizers



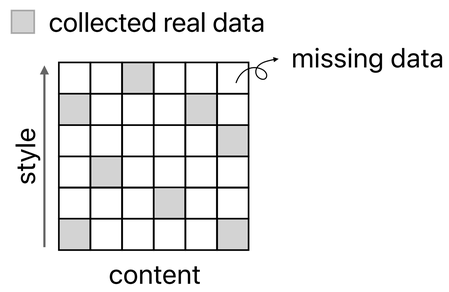
Data incubation
first learn a generative model on the training data



Observations

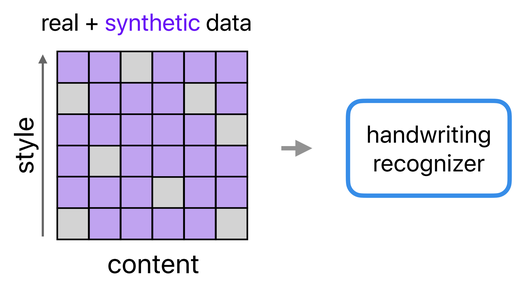
Missing content and missing style in collected data

- corpus cannot contain all the word combinations
- fixed number of users in the data collection

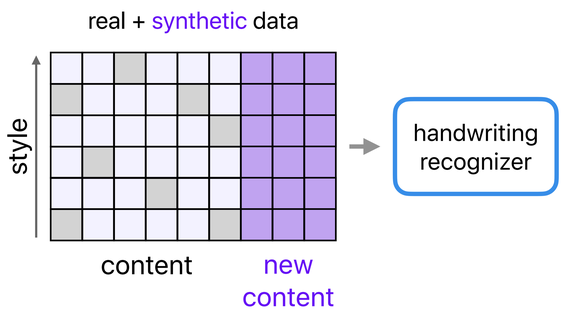


What we propose (data incubation)

Fill in missing style



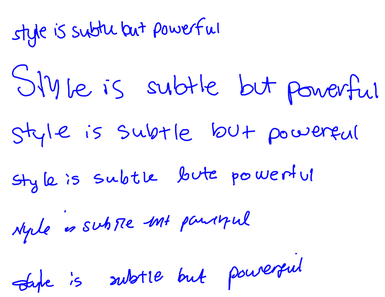
Fill in missing style + content



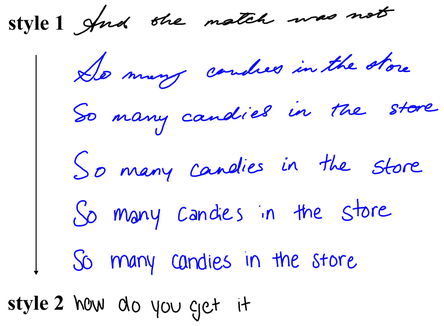
Results

Examples of synthesized handwriting

style sampled from prior distribution



interpolate between two reference styles



more results & details

Does filling in missing style help?
English handwriting dataset (0.3M real data)
Synthesize using the same corpus as the training data (i.e., same content)

	Character Error Rate (↓)	
train on collected data (0.3M)	4.9%	
collected + same style as collected (1M)	4.7% (4% ↓)	← same content and same style
collected + new style (1M)	4.0% (18% ↓)	← filling in style
collected + new style (**4M**)	3.6% (28% ↓)	← filling in a lot more style

Does filling in missing content help?
Multilingual handwriting dataset (0.6M real data)

	CER on similar corpus (↓)	CER on underrepresented content (↓)
train on collected data (0.8M)	8.2%	18.5%
collected + new style (3M)	4.5% (46% ↓)	16.9% (9% ↓)
collected + new style + **new content** (8.1M)	2.8 % (66% ↓)	6.3 % (66% ↓)