KUZUSHIJI RECOGNITION: 2ND PLACE SOLUTION

Konstantin Lopukhin, Scrapinghub

November 11th, 2019

ABOUT ME

- Kaggle Grandmaster since 2019, competing since 2016
- Work on web information extraction at Scrapinghub
- Member of Open Data Science Community [ods.ai]
- Enjoy working on unusual (for me) challenges
- Like if high solution quality can be achieved

MY MOTIVATION

- I never worked on character recognition
- Great organisation of the competition
- Dataset is clean and looks beautiful
- I love Japan!

SOLUTION: TWO STAGES

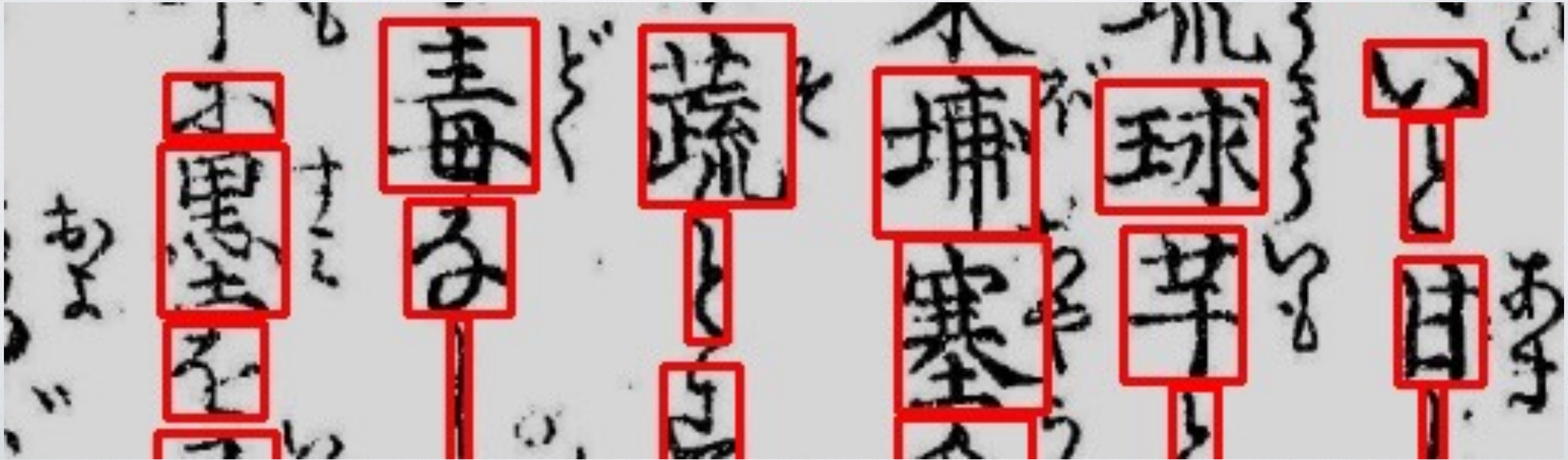
- Stage 1: detect characters
- Stage 2: classify characters

STAGE 1: DETECT CHARACTERS

- **Input: image**
- Output: boxes around all characters

STAGE 1: DETECT CHARACTERS

- Input: image
- **Output: boxes around all characters**

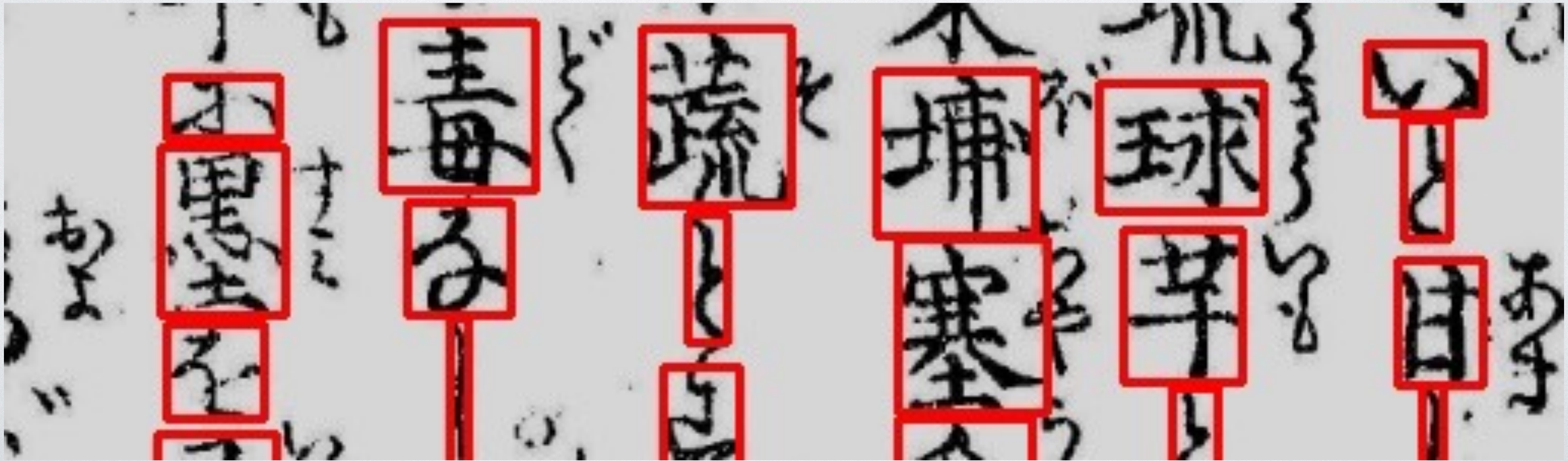


DETECTION CHALLENGES

- False detection: annotations, characters in images, unknown characters.
- Characters consisting of multiple parts: sometimes incorrectly broken into multiple boxes.
- Still, detection quality is high: 0.992 assuming perfect classification.

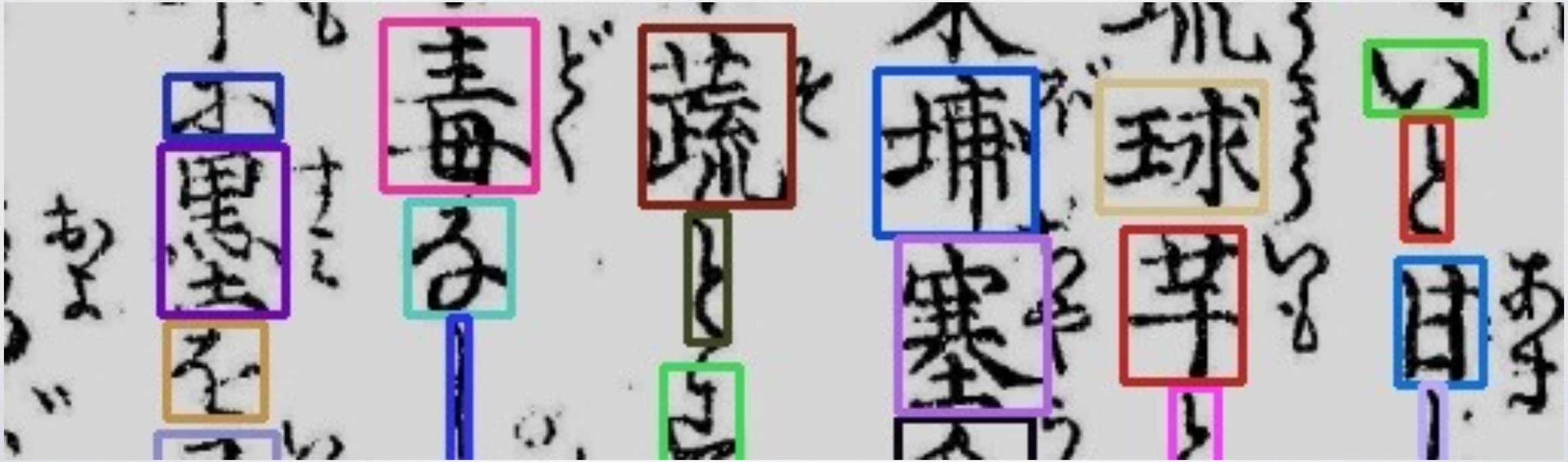
STAGE 2: CLASSIFY CHARACTERS

- **Input: image and boxes from stage 1**
- Output: classes for all boxes



STAGE 2: CLASSIFY CHARACTERS

- Input: image and boxes from stage 1
- **Output: character classes for all boxes**



CLASSIFICATION CHALLENGES AND SOLUTIONS

- Context is important: process all characters at once.
- Large number of classes (more than 4k): use large models, regularisation, data augmentation, careful tuning.
- Detection errors: allow the model to reject false detections.
- High variability between different books: use test adaptation.

RESULTS

- Simple baseline: **0.838**
- Tuning baseline: **0.922**
- Using bigger model: **0.934**
- Adapting to test set: **0.945**
- Mixing models from 5 folds: **0.950**

LESSONS LEARNED

- Dataset is very clean and rich
- Classification is harder than detection
- Modern computer vision methods work very well here
- Hard to analyse errors without domain knowledge
- Two stages were not required (see 1st place)

THANK YOU!

- Thank you dear organising team, it was a great success
- Thank you Kaggle and all who participated
- Thank you CODH for inviting me
- Thank you everyone for attention

Code and models: https://github.com/lopuhin/kaggle-kuzushiji-2019