

Adapting OCR with Limited Supervision

Bc. Pavel Raur, Bc. Marián Zimmermann, Bc. Jiří Žilka

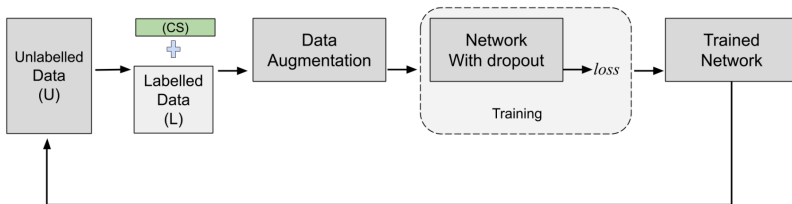
Autoři článku: doc. Ing. Deepayan Das and C V Jawahar , Ph.D.



11. dubna 2024

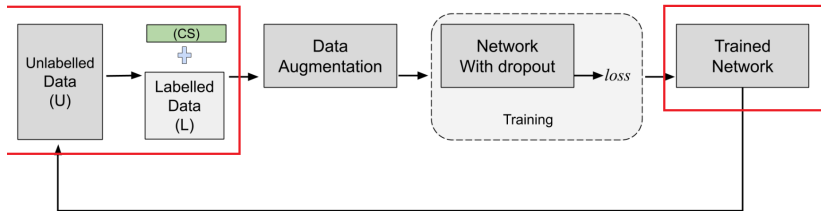


- OCR - optical character recognition - optické rozpoznání znaků v obraze
- Přepis dokumentů z naskenovaného dokumentu do textové podoby.
- Self-supervised učení – učení na neanotovaných datech.
- Semi-supervised učení – učení na kombinaci anotovaných a neanotovaných dat.

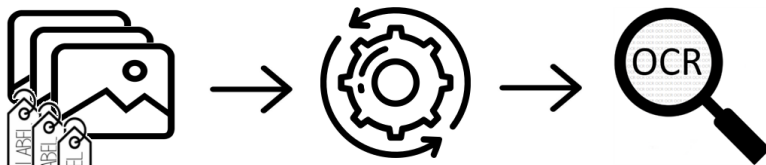


Semi-supervised učení

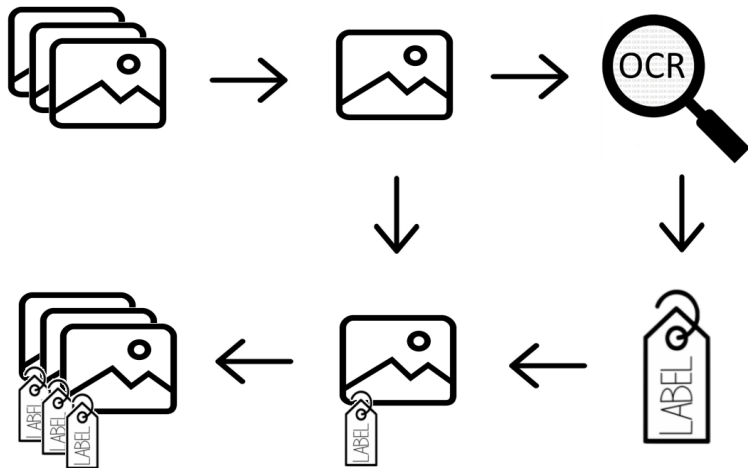
- Selekcce dat k trénování
- Augmentace dat – úprava dat pro lepší generalizaci
- Úprava sítě pro lepší generalizaci



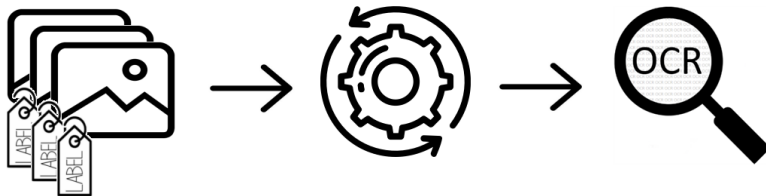
- Řeší problém s nedostatkem trénovacích dat
- Selekce dat pro trénování
- Strojové generování labelů



- Trénování na části datasetu s labely.
- Málo trénovacích dat s labely.



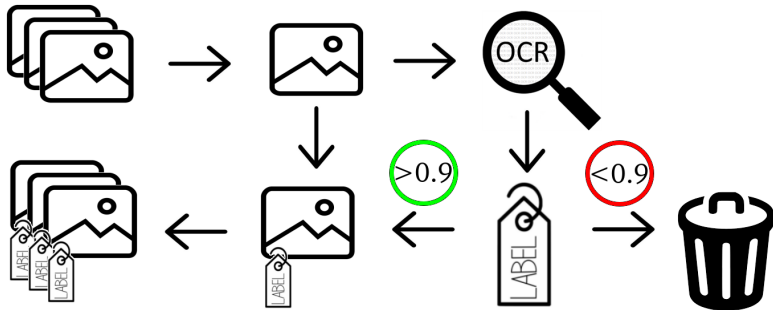
- Pseudo-labely



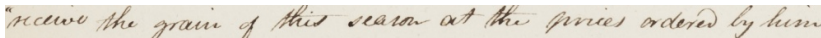
- Zahrnutí nových dat do trénovací sady



- **Confirmation bias** – přeučení sítě na svých vlastních chybách
- Nemusí souviset s generalizací – síť skvěle rozpozná znaky na vstupu, jen se je naučila špatně přepisovat



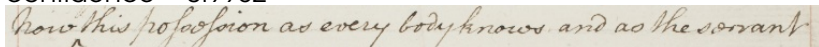
- Kvalitní přepisy – základ
- Výběr labelů – u OCR např. použití slovníku pro rozpoznání slov, které alespoň existují
- Nemusí existovat doménová znalost, pak výběr dle confidence přepisu



GT: "receive the grain of this season at the prices ordered by him" vs.

P: "because the grain of this weaver at the previous ordered by hisin"

confidence = 0.9982



GT: Now this possession as everybody knows and as the servant vs.

P: how this rejection as every body knows and as the servant

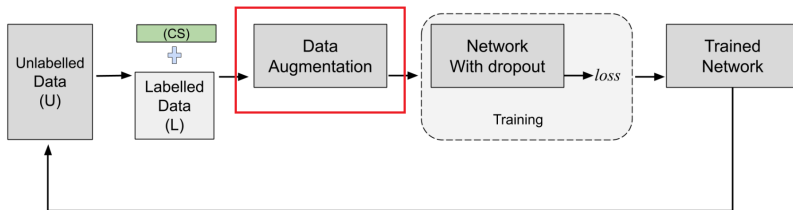
confidence = 0.4727

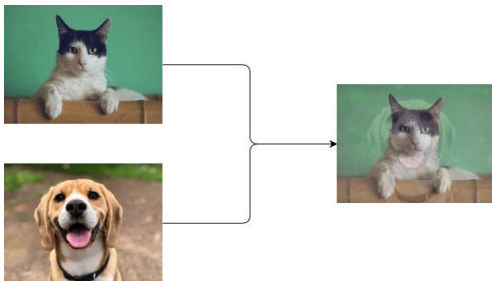
Formule pro výpočet score:

$$\text{score} = -\alpha \sum_{i=1}^t \log(p_i) + (1 - \alpha) \frac{1}{m} \log P(w_1, \dots, w_m)$$

Vážený součet predikcí jazykového modelu a OCR.

- 20% vzorků s největším score přidáno do datasetu.



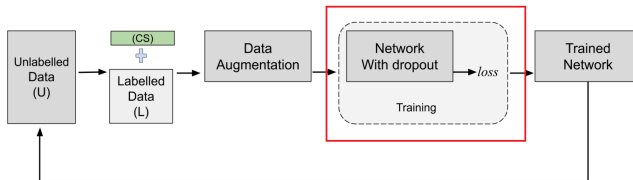


$datapoints(x_1, y_1), (x_2, y_2)$

$$\lambda \in [0, 1]$$

$$\hat{x} = \lambda x_i + (1 - \lambda) x_j$$

$$\hat{y} = \lambda y_i + (1 - \lambda) y_j$$



- konvoluční vrstvy (VGG, ResNet18)
- rekurentní vrstvy - Bi-LSTM
- metody: Full, Last, Chain Thaw, Unfreeze
- Optimalizátor Adam
- Slanted Triangular Learning Rate (STLR)
- early stopping criterion - proti overfitting
- Dropout, Drop-connect

- Network regularization against confirmation bias
- Weight Dropped LSTMs - does not affect long-term dependency as much
- 50% probability on FC layers after Bi-LSTM

Heuristics	English		Hindi	
	CRR	WRR	CRR	WRR
ST	94.22	85.12	92.13	85.41
+ STL	94.72	86.88	92.17	85.45
+ noise	95.61	91.87	92.26	85.57
+ dropout	95.99	92.57	92.26	85.54
+ mixup	96.48	93.57	92.57	86.23

- Self-training < Fine-tuning < ST + FT