ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ РОЗПІЗНАВАННЯ ПОМИЛОК В ДОКУМЕНТАХ

КЕРІВНИК КУРСОВОЇ РОБОТИ

ДОЦ. ІГНАТЕНКО О.П.

ВИКОНАЛА СТУДЕНТКА

ЖИРКОВА А.П.

АКТУАЛЬНІСТЬ ТЕМИ

ПРИ СКЛАДАНІ ДОКУМЕНТІВ ВИНИКАЮТЬ ПОМИЛКИ РІЗНОГО ТИПУ, ЯКІ ВАЖКО ВІДСТЕЖИТИ ТА НАВІТЬ ПРИ ДУЖЕ УВАЖНОМУ ПЕРЕГЛЯДІ МОЖНА ПРОПУСТИТИ.

ОТЖЕ, ВИНИКАЄ ПОТРЕБА ОБРОБЛЯТИ ДОКУМЕНТИ В АВТОМАТИЗОВАНОМУ РЕЖИМІ, ЩОБ ВІДСЛІДКОВУВАТИ ПОМИЛКИ ТА МІНІМІЗУВАТИ ЇХНЮ ПРИСУТНІСТЬ У ДОКУМЕНТАХ.

МЕТА ДОСЛІДЖЕННЯ

Метою курсової роботи є розробка моделі згорткової нейронної мережі для коректної класифікації документів за наявністю або відсутністю печаток на ньому.

ПОСТАНОВКА ЗАДАЧІ

- Прикладами помилок у документах є відсутність печаток, або відсутність підписів, або ж присутність російських слів в україномовному документі.
- **В рамках виконання курсової роботи обрано задачу розпізнавання печаток у документах.**
- Для навчання обраної моделі машинного навчання, на вхід подаються зображення документів, які вона класифікує як 0, якщо на зображенні немає печатки, та як 1, якщо має хоча б одну.

ІСНУЮЧІ МЕТОДИ

01

Двоетапний підхід до вилучення візуальних об’єктів з паперових документів.

02

Підхід до виявлення печаток у документах, який використовує поєднання деяких простих характеристик зображення.

03

Підхід до розпізнавання печаток, який фокусується на розпізнаванні геометричних форм, притаманних їм.

ВЛАСНИЙ МЕТОД

Розв’язувати задачу розпізнавання печаток у документах було вирішено за допомогою згорткових нейронних мереж, які є найефективнішим методом роботи з зображеннями.

ЗАВАНТАЖЕННЯ ДАНИХ

- Усі дані, які підходять для вирішення даної задачі, представлені або як зображення, або як документи у форматі PDF.
- Посилання на сайті Prozorro представлені у форматі

  “https://prozorro.gov.ua/tender/UA-{year}-{mon}-{day}-{uid}-{s}/”
- Усі посилання записуються у файл, з якого потім вони зчитуються та завантажуються на комп’ютер.

ПОПЕРЕДНЯ ОБРОБКА ДОКУМЕНТІВ

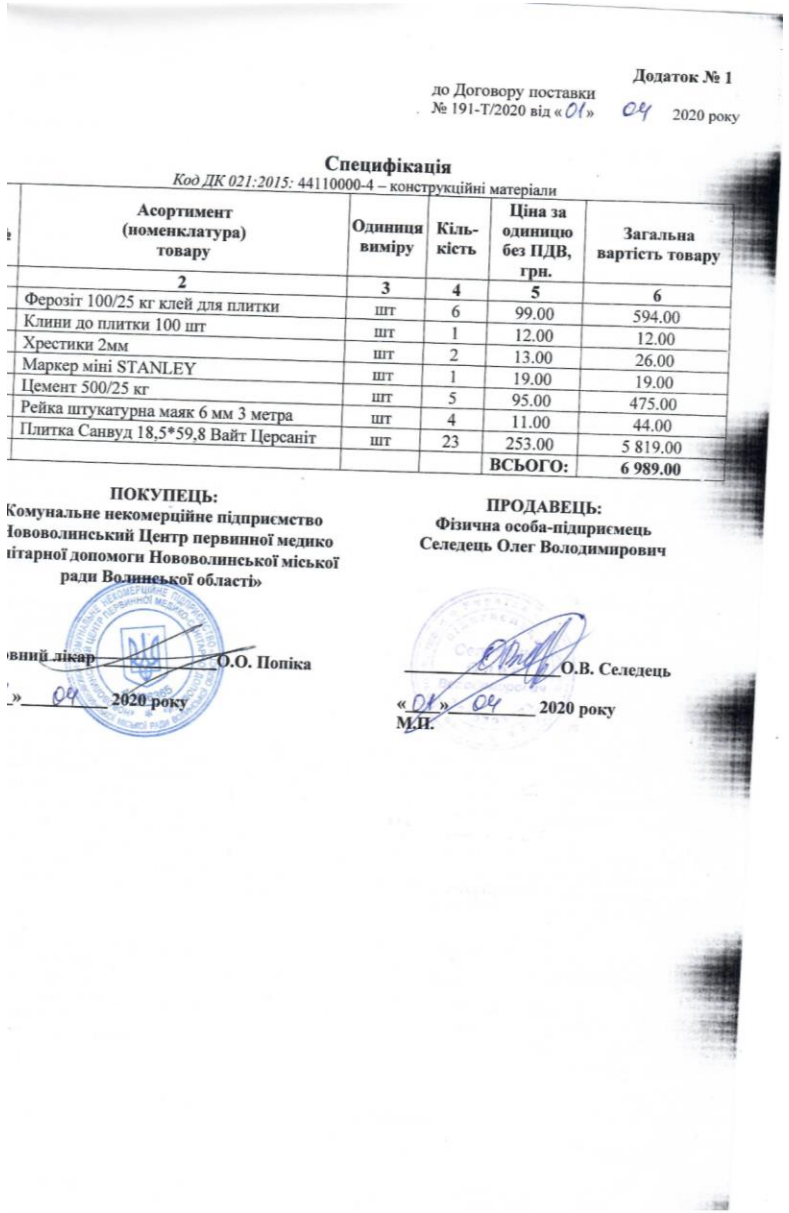
Всі документи, представлені у форматі PDF, перетворюються на зображення.

Формується директорія з зображеннями.

Розмір зображень стає 1350x1900.

Вхідні дані розбиваються на тренувальну та тестову вибірки.

ПРИКЛАДИ ДОКУМЕНТІВ



Додаток № 1

до Договору поставки
№ 191-Т/2020 від «01» 04 2020 року

Специфікація

Код ДК 021:2015: 44110000-4 – конструкційні матеріали

Асортимент (номенклатура) товару	Одиниця виміру	Кіль-кість	Ціна за одиницю без ПДВ, грн.	Загальна вартість товару
2	3	4	5	6
Ферозіт 100/25 кг клей для плитки	шт	6	99.00	594.00
Клини до плитки 100 шт	шт	1	12.00	12.00
Хрестики 2мм	шт	2	13.00	26.00
Маркер міні STANLEY	шт	1	19.00	19.00
Цемент 500/25 кг	шт	5	95.00	475.00
Рейка штукатурна маяк 6 мм 3 метра	шт	4	11.00	44.00
Плитка Санвуд 18,5*59,8 Вайт Церсаніт	шт	23	253.00	5 819.00
			ВСЬОГО:	6 989.00

ПОКУПЕЦЬ:
Комунальне некомерційне підприємство Нововолинський Центр первинної медико санітарної допомоги Нововолинської міської ради Волинської області»

вний лікар **О.О. Попіка**

«» 04 2020 року

ПРОДАВЕЦЬ:
Фізична особа-підприємець Селедець Олег Володимирович

О.В. Селедець

«01» 04 2020 року

М.П.

ЗВІТ

про укладені договори

1. Дата укладення договору – 31.03.2020 р.

2. Номер договору – 752/24/16-20.

3. Найменування замовника – Приватне акціонерне товариство «Акціонерна компанія «Київводоканал».

4. Код згідно з ЄДРПОУ замовника – 03327664.

5. Місцезнаходження замовника – 01015, м. Київ, вул. Лейпцизька,1а.

6. Найменування постачальника товарів, виконавця робіт чи надавача послуг (для юридичної особи) прізвище, ім'я, по батькові (для фізичної особи), з яким укладено договір – Товариство з обмежен відповідальністю «МАКС МАТЕРІАЛИ».

7. Код згідно з ЄДРПОУ/реєстраційний номер облікової картки платника податків постачальника това виконавця робіт чи надавача послуг – 39482858.

8. Місцезнаходження постачальника товарів, виконавця робіт чи надавача послуг (для юридичної осо або місце проживання (для фізичної особи) та номер телефону, телефаксу – 02090, м. К вул. Алма-Атинська, буд. 8, тел. 067-325-25-50.

8.1. Вид предмета закупівлі – закупівля товарів.

9. Конкретна назва предмета закупівлі – поролон.

10. Найменування (номенклатура, асортимент) товарів, робіт чи послуг – згідно зі Специфікаціями.

11. Кількість товарів, робіт чи послуг – кількість вказана в Специфікаціях складених згідно письмо заявок Покупця.

12. Місце поставки товарів, виконання робіт чи надання послуг – 01015, м. Київ, вул. Лейпцизька, 1 02232, м. Київ, вул. Пухівська, 1-Д.

13. Строк поставки товарів, виконання робіт чи надання послуг – 5 робочих днів з дати надання письмо заявки Покупцем.

14. Інформація про технічні та якісні характеристики товарів, робіт чи послуг – відповідно до у Договору.

15. Ціна договору – 50 000 грн. 00 коп. з ПДВ.

16. Строк дії договору – до 31.12.2020 р.

16.1. Джерело фінансування закупівлі – Власні кошти ПрАТ «АК «Київводоканал».

17. Ідентифікатор договору*.

18. Одиниця виміру – пак.

19. Ціна за одиницю**.

Заступник директора з підготовки виробництва – начальник управління матеріально-технічного забезпечення департаменту з підготовки виробництва

О. І. Згонник

КЛАСИФІКАЦІЯ ДОКУМЕНТІВ

ВИКОРИСТОВУЄТЬСЯ GOOGLE COLAB

```
#authenticate on Google Drive
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
```

АУТЕНТИФІКАЦІЯ

```
# Get train and validation data from GD
train = drive.CreateFile({'id': '1ESU6W-8ugAsBPPC0QRANHEbEpOCyQIVT'})
train.GetContentFile('train.zip')

valid = drive.CreateFile({'id': '130pwjSebfaHktA-3ITr-NjWbbiyxpKE8'})
valid.GetContentFile('validation.zip')

!unzip train.zip
!unzip validation.zip
```

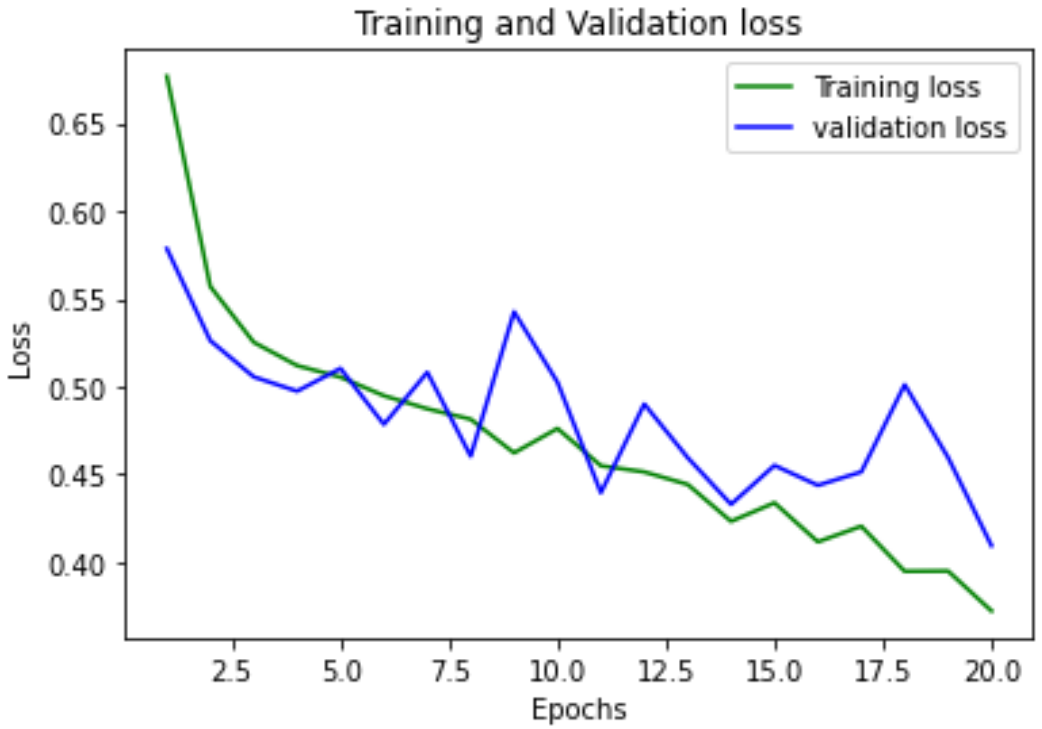
ЗАВАНТАЖЕННЯ ДАНИХ

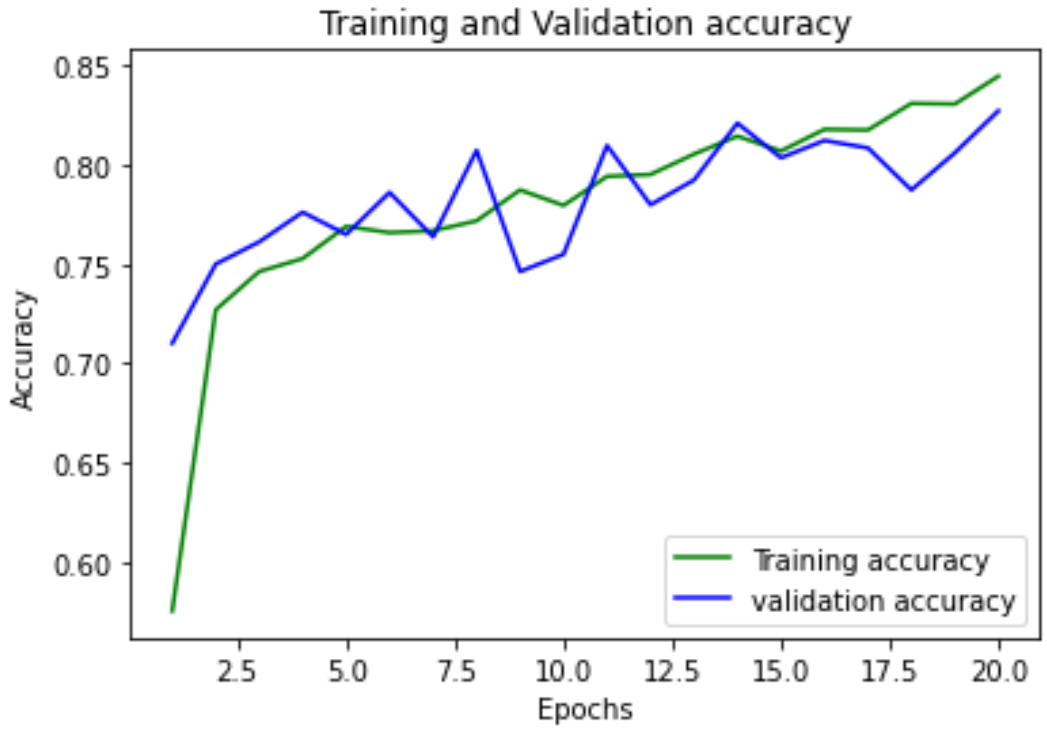
- Завантаження архівів у поточний файловий простір
- Розархівація даних

НЕЙРОННА МЕРЕЖА

```python
# build convolutional neural network and train it
from keras.models import Sequential
from keras.layers import Dense, Dropout, Flatten
from keras.layers import Conv2D, MaxPooling2D

model = Sequential()
model.add(Conv2D(64, kernel_size=3,activation='relu',input_shape=(28,32,3)))
model.add(Conv2D(128, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Conv2D(32, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))
model.add(Flatten())
model.add(Dense(1000, activation='relu'))
model.add(Dropout(0.25))
model.add(Dense(500, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(1, activation='sigmoid'))

model.compile(loss='binary_crossentropy',optimizer='Adam',metrics=['accuracy'])

model.fit(X_train, y_train, epochs=20, validation_data=(X_test, y_test))
```

РЕЗУЛЬТАТИ


Training and Validation loss
Training loss
validation loss
Loss
0.65
0.60
0.55
0.50
0.45
0.40
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epochs


Training and Validation accuracy
Accuracy
0.85
0.80
0.75
0.70
0.65
0.60
Training accuracy
validation accuracy
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epochs

ВПЛИВ РОЗМІРУ ТРЕНУВАЛЬНОГО НАБОРУ ДАНИХ НА РОБОТУ НЕЙРОННОЇ МЕРЕЖІ

Розмір тренувальної вибірки	Розмір валідаційної вибірки	Розмір тестової вибірки	Точність роботи моделі	Час, необхідний на навчання моделі (у хв)	Час роботи мережі на тестових даних (у сек)
435	109	234	**81.19%**	0,66	0,23
1164	291		**80.76%**	1,78	0,22
1896	474		**80.76%**	2,86	0,21
2596	650		**86.32%**	3,86	0,27
3216	804		**88.03%**	4,81	0,37

ВИСНОВКИ

Вивчено різні методи роботи з зображеннями.

Досліджено можливість використання машинного навчання для роботи з зображеннями.

Проведено аналіз існуючих методів вирішення поставленої задачі.

Проведено збір та обробку даних.

Побудовано нейронну мережу для класифікації документів.

Вивчено залежність точності класифікації від кількості даних для навчання моделі.

ДЯКУЮ ЗА УВАГУ!