Машинне навчання для виявлення фальсифікацій у відеоконтенті

Курсова робота студента 3-го року навчання
Освітня програма: Комп'ютерні науки, 122
Автор: Максим Шетеля
Керівник: Салата К.В.

Актуальність дослідження

Deepfake-технології створюють серйозні виклики для безпеки та достовірності цифрового контенту, що вимагає розробки ефективних методів для їх автоматичного виявлення.

Основні цілі

- Дослідити сучасні технології генерації deepfake
- Дослідити методи виявлення фальсифікацій
- Розробити модель для виявлення фальсифікованого контенту.

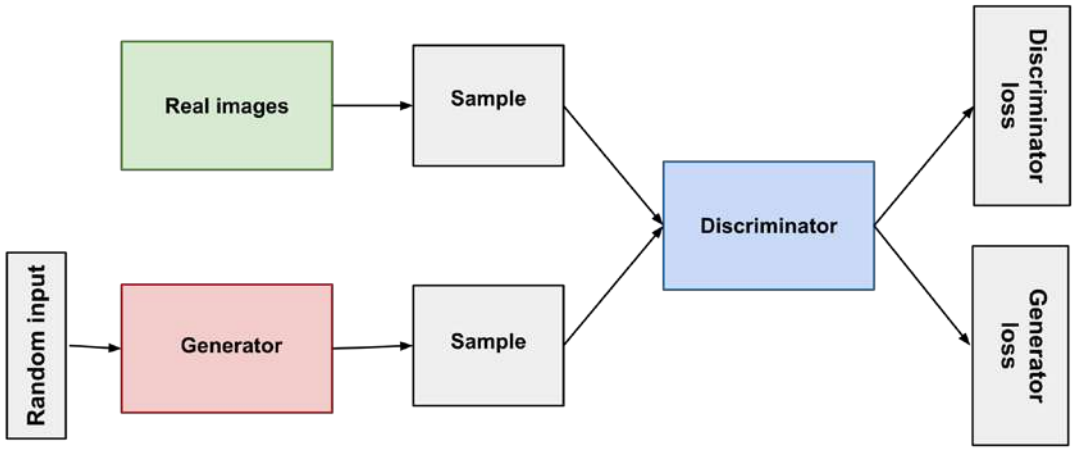
Теоретичні основи deepfake

- Deepfake — технологія створення реалістичних підробок відео та зображень

- У 2017 році користувач Reddit вперше застосував автоенкодери для автоматичної заміни обличчя у відео

Теоретичні основи deepfake

- Розвиток deepfake завдяки генеративним змагальним мережам Generative adversarial network (GAN) [1]



Теоретичні основи deepfake

- Сучасні deepfake-моделі:

  StyleGAN

  First Order Motion Model

  Wav2Lip

  дифузійні моделі

Ознаки фальсифікації

Основні ознаки [2, 3]:

- Візуальні артефакти
- Аудіо-розсинхронізація
- Поведінкові аномалії
- Технічні недоліки

Методи детекції

- Згорткові нейронні мережі (CNN)
- Рекурентні нейронні мережі (RNN, LSTM)
- Attention-механізми і трансформери
- Комбіновані підходи

Практична реалізація

- Вибір датасету: DeepFakeFace, FaceForensics++.
- Побудова моделі CNN + BiLSTM + Attention.
- Тренування та оцінка точності (Loss, Accuracy, Precision, Recall).

Результати

Таблиця – Результати тренування та тестування CNN+BiLSTM+Attention

Epoch	Train Loss	Train Acc (%)	Train Precision	Train Recall	Test Loss	Test Acc (%)	Test Precision	Test Recall
1	0.6975	49.69	0.4966	0.4625	0.6849	56.25	0.5410	0.8250
2	0.6907	54.37	0.5261	0.8812	0.6879	50.00	0.0000	0.0000
3	0.6971	54.69	0.6056	0.2687	0.6812	58.75	0.5522	0.9250
4	0.6878	56.87	0.5679	0.5750	0.6765	61.25	0.5672	0.9500
5	0.6755	58.44	0.5818	0.6000	0.6664	60.00	0.6176	0.5250
6	0.6734	60.31	0.6025	0.6062	0.6539	60.00	0.6176	0.5250
7	0.6681	60.62	0.6037	0.6188	0.6479	63.75	0.6410	0.6250
8	0.6631	61.25	0.6111	0.6188	0.6447	70.00	0.6379	0.9250
9	0.6661	61.56	0.6370	0.5375	0.6469	66.25	0.6140	0.8750
10	0.6563	59.38	0.5882	0.6250	0.6350	68.75	0.6744	0.7250
11	0.6445	64.38	0.6386	0.6625	0.6424	58.75	0.6061	0.5000
12	0.6305	64.69	0.6460	0.6500	0.6479	62.50	0.5862	0.8500
13	0.6041	66.25	0.6461	0.7188	0.6496	63.75	0.6571	0.5750
14	0.6125	65.31	0.6503	0.6625	0.6476	61.25	0.5818	0.8000
15	0.6029	67.50	0.6609	0.7188	0.6345	66.25	0.6066	0.9250

Висновки

Джерела

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. **Generative adversarial nets.** *Advances in Neural Information Processing Systems.* 2014. Vol. 27. P. 2672–2680 .
2. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J. **Deepfakes and beyond: A survey of face manipulation and fake detection.** *Information Fusion.* 2020. Vol. 64. P. 131–148 .
3. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M. **FaceForensics++: Learning to Detect Manipulated Facial Images.** *Proc. IEEE International Conference on Computer Vision (ICCV).* 2019. P. 1–11 .

Дякую за увагу