2021「中技社科技獎學金」
2021 CTCI Foundation Science and Technology Scholarship
境外生生活助學金
Living Grant for International Graduate Students

Fire Detection using Transformer Network

3rd year PhD student– Mohammad Shahid Advisor-Prof Kai-Lung Hua
Dep't of computer science and information engineering, NTUST Taiwan.

Abstract

Every year, fire is becoming one of the severe natural hazards that cause threats to the ecology, economy, human life and even more. This work shows that Vision Transformer (ViT) is a viable tool for automated fire detection by aggregating features from the whole spatial context. The proposed method is tested on benchmark fire datasets to reveal the framework's strength and effectiveness.

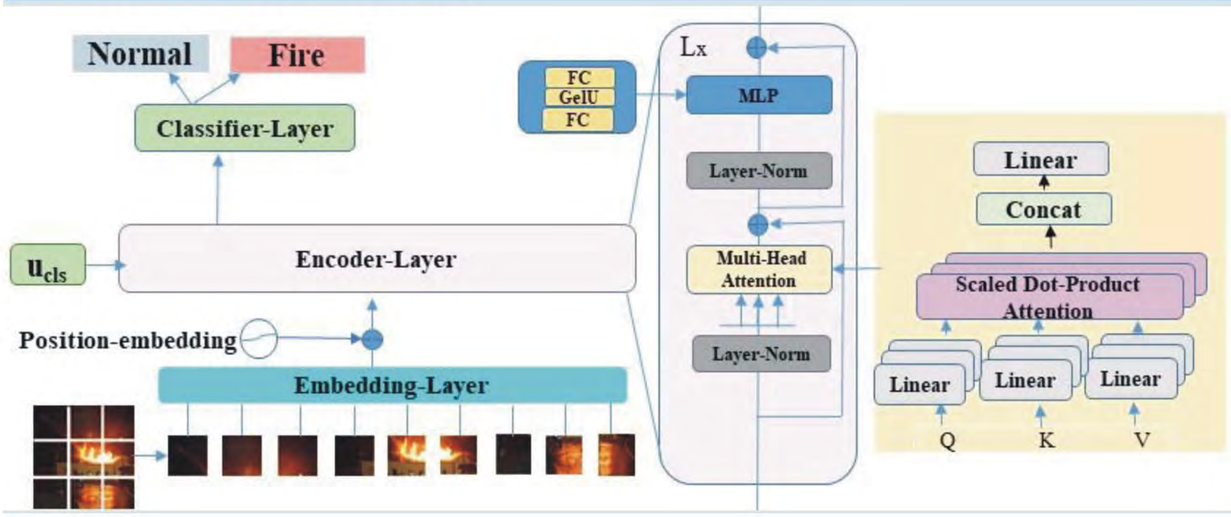
Motivation

- Computer vision system has fast reaction speed, operate in open environments, and shown a higher success probability than sensor-based
- It only requires adjusting in CCTV camera to decrease costs on new systems installation and procurement

Method

- Divided an image into fixed sizes patches
  For an input image $(x) \in R^{H \times W \times C}$ *split* into N patches $(x)p \in R^{N \times (P2C)\backslash}$ where c is no of channels and $(H \times W)$ is the resolution of the image x, and $(P \times P)$ is the resolution of x_p.
- Image patches are reshaped.
- from these reshape patches, Created lower-dimensional embeddings.
- Combined with positional embedding.
- Fed these sequence to transformer encoder

Network architecture



Dataset

- Adopted an Imagenet-21k-pretrained model that is fine-tuned on Imagenet1k.
- And further fine-tuned it on Fire Datasets;
- Evaluated in terms of False Positive (FP), False Negative (FN), Recall, Precision, accuracy and F1-score.

All experiments are conducted on the machine (Intel(R) Core(TM) i7-7700K) with a RAM of 32 GIGabytes memory capacity and NVidia GTX 1080Ti graphics processing unit (GPU) of eleven GIGabytes. As for the software, all codes are implemented using the Pytorch deep learning framework on the Ubuntu system

Result

Performance comparison on

Methods	FP(%)	FN(%)	Acurracy(%)
Squeezenet [1]	11.18	89.90	90.95
InceptionV1 [2]	18.17	3.83	89.81
ViT-B/32	2.15	1.02	94.03
ViT-L/32	2.80	1.03	94.01
ViT-B/16	3.75	1.06	93.30
ViT-L/16	2.95	1.04	93.70

a) Foggia Dataset

Methods	Recall	Precision	F1-score
Squeezenet [1]	0.9498	0.9401	0.9449
InceptionV1 [2]	0.8289	0.8919	0.8592
ViT-B/32	0.9709	0.9910	0.9808
ViT-L/32	0.9703	0.9835	0.9768
ViT-B/16	0.9690	0.9800	0.9744
ViT-L/16	0.9699	0.9801	0.9749

b) Bowfire Dataset

Conclusion

Unlike CNN's, the ViT model employs images as a group of patches to gain a long-range relationship. and the preliminary results showed the significance of new structures to enhance detection accuracy and match state-of-the-art technique.

References

1. Muhammad et al. "Efficient deep CNN-based fire detection and localization in video surveillance applications." IEEE Transactions on Systems, Man, and Cybernetics: Systems 49.7 (2018)
2. Andrew et al "Experimentally defined convolutional neural network architecture variants for non-temporal real-time fire detection." 2018 25th IEEE (ICIP)