TransNet: A deep network for fast detection of common shot transitions

https://github.com/soCzech/TransNet

Tomáš Souček et al: 2019

Ngày 31 tháng 10 năm 2022

MỤC ĐÍCH/OUTPUT

- Đề xuất TransNet, một kiến trúc có thể mở rộng với nhiều hoạt động dilated 3D convolutional trên mỗi lớp (thay vì chỉ một như bình thường) dẫn đến trường xem lớn hơn với ít tham số có thể đào tạo hơn. Mặc dù kiến trúc được đào tạo chỉ dựa trên hai kiểu chuyển đổi phổ biến (hard cuts và dissolves), nhưng nó vẫn đạt được kết quả hiện đại trên tập dữ liệu RAI. Mạng sử dụng các dilated convolution và chỉ hoạt động trên các frame được thay đổi kích thước nhỏ.
- Quá trình đào tạo sử dụng các chuyển đổi được tạo ngẫu nhiên bằng cách sử dụng các ảnh được chọn từ bộ dữ liệu TRECVID IACC.3

Dataset

Dataset: RAI dataset và TRECVID IACC.3 dataset

- **TRECVID IACC.3 dataset:** được cung cấp với một tập hợp các phân đoạn thời gian được xác định trước. Xem xét các phân đoạn của 3000 video IACC.3 được chọn ngẫu nhiên. Hơn nữa, các phân đoạn có ít hơn 5 khung hình đã bị loại trừ và từ tập hợp còn lại chỉ mỗi phân đoạn khác được chọn $\rightarrow$ 54884 phân đoạn được chọn.
- Các ví dụ đào tạo được tạo ra theo yêu cầu trong quá trình đào tạo bằng cách lấy mẫu ngẫu nhiên hai bức ảnh và kết hợp chúng với một kiểu chuyển đổi ngẫu nhiên. Chỉ hard cuts và dissolves được xem xét để train. Vị trí của quá trình chuyển đổi được tạo ngẫu nhiên.
- Đối với dissolves, chiều dài của nó cũng được tạo ngẫu nhiên từ khoảng thời gian [5, 30]. Độ dài N của mỗi chuỗi huấn luyện được chọn là 100 khung hình. Kích thước của khung đầu vào được đặt thành 48 $\times$ 27 pixel.
- Validation: thêm 100 video IACC.3 (tức là khác với bộ đào tạo) đã được gắn nhãn thủ công -> có 3800 shots.
- Test: Sử dụng RAI dataset.

CÔNG TRÌNH LIÊN QUAN

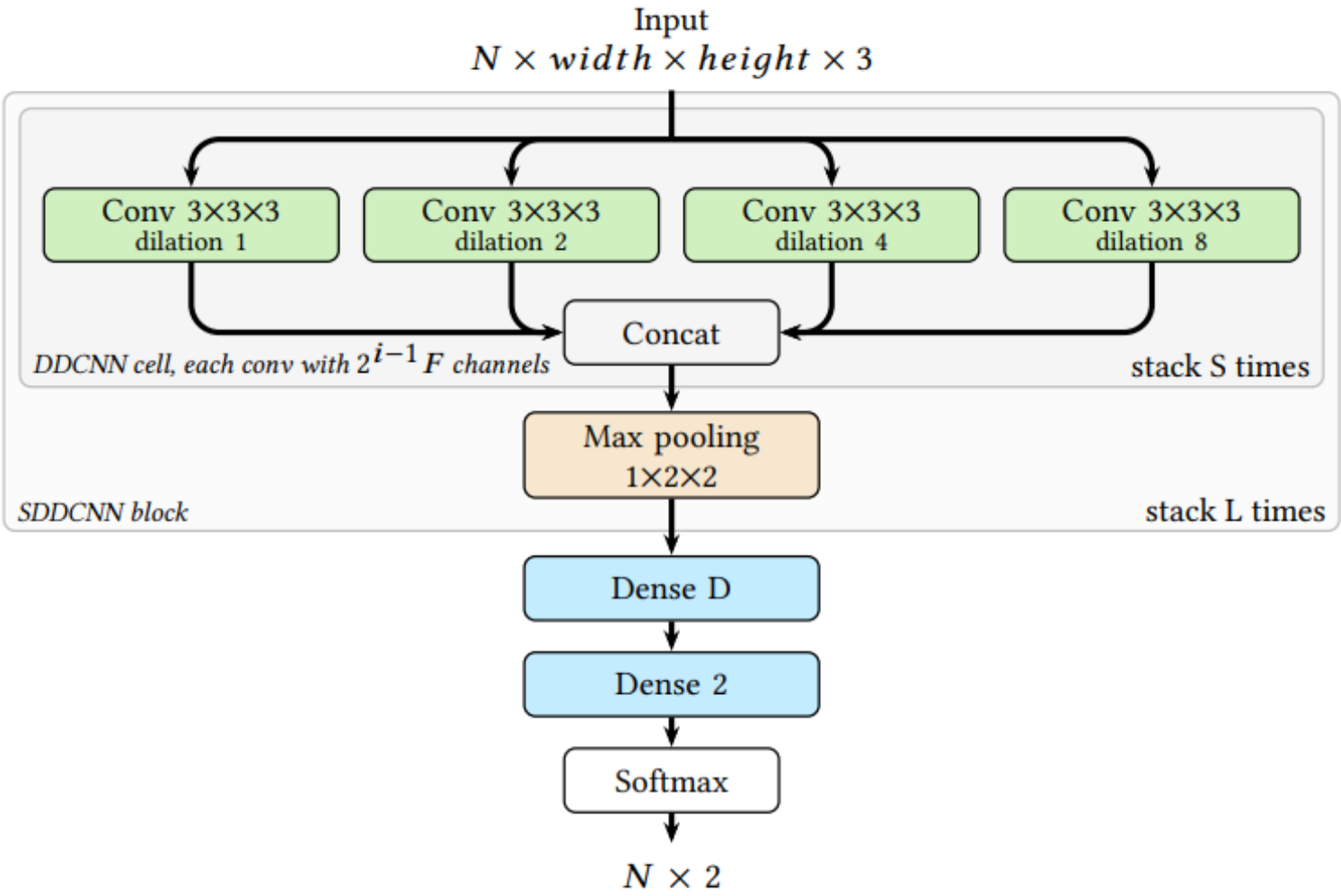
- **Shao et al[2015]:** sử dụng HSV và biểu đồ gradient để phát hiện shot boundary.
- **Apostolidis et al[2014]:** không chỉ sử dụng biểu đồ mà còn sử dụng một tập hợp các bộ mô tả SURF để phát hiện sự khác biệt giữa một cặp frame.
- **Baraldi et al[2015]:** sử dụng phân cụm quang phổ đưa ra một tập hợp các tính năng cho mọi khung hình được trích xuất bởi deep siamese network.
- **Gygli[2018]:** đã sử dụng một mạng nơ-ron tương đối nông với các chập 3D với chiều thứ ba theo thời gian. Mặc dù các chập 3D làm tăng đáng kể độ phức tạp tính toán và yêu cầu bộ nhớ so với các chập 2D tiêu chuẩn do kích thước được bổ sung, nhưng lại tăng độ chính xác và tốc độ.
- **Hassanien et al[2017]:** cũng sử dụng 3D CNN tuy nhiên đầu ra được cung cấp thông qua bộ phân loại SVM và quá trình xử lý hậu kỳ tiếp theo được thực hiện để giảm cảnh báo sai về quá trình chuyển đổi dần dần thông qua sự khác biệt thời gian theo biểu đồ.

MODEL ARCHITECTURE

Kiến trúc TransNet được đề xuất (Hình 1) tuân theo công trình của Gygli[2018] và các kiến trúc tích chập tiêu chuẩn khác.

- Input: mạng lấy một chuỗi N khung hình video liên tiếp và áp dụng một loạt các phức hợp 3D trả về dự đoán cho mọi khung hình trong đầu vào. Mỗi dự đoán thể hiện khả năng một khung hình nhất định là 1 shot boundary.
- Xây dựng block chính của mô hình (Dilated DCNN cell) được thiết kế dưới dạng 4 3D $3\times3\times3$ convolutional operations.
- Các chập sử dụng các tốc độ giãn nở khác nhau cho thứ nguyên thời gian và đầu ra của chúng được nối với thứ nguyên kênh. Cách tiếp cận này làm giảm đáng kể số lượng các tham số có thể đào tạo so với các biến thể 3D tiêu chuẩn với cùng một trường xem.
- Nhiều ô DDCNN chồng lên nhau, theo sau là tổng hợp tối đa theo không gian tạo thành một khối DDCNN xếp chồng lên nhau.

MODEL ARCHITECTURE



Hình 1. Kiến trúc mạng TransNet shot boundary detection cho S = 1 và L = 1. Lưu ý rằng N đại diện cho độ dài của chuỗi video, không phải batch size. Case N = 100.

MODEL ARCHITECTURE

- TransNet bao gồm nhiều khối SDDCNN, mỗi khối tiếp theo hoạt động trên độ phân giải không gian nhỏ hơn nhưng kích thước channel lớn hơn, tăng thêm sức mạnh biểu đạt và trường tiếp nhận của mạng.
- Hai lớp fully connected sẽ tinh chỉnh các tính năng được trích xuất bởi các lớp convolutional và dự đoán shot boundary có thể có cho mọi biểu diễn khung hình một cách độc lập (trọng lượng của các lớp được chia sẻ).
- ReLU activation function được sử dụng trong tất cả các lớp với ngoại lệ duy nhất của lớp fully connected cuối cùng có đầu ra softmax.

CHI TIẾT TRAINING

Kiến trúc được đề xuất cung cấp các meta-parameters sau khi đã được dò tìm bằng grid search:

- **S:** số lượng ô DDCNN trong lớp SDDCNN.
- **L:** số lớp SDDCNN.
- **F:** số bộ lọc trong tập hợp các lớp DDCNN đầu tiên (nhân đôi trong mỗi lớp SDDCNN tiếp theo).
- **D:** số lượng neurons trong lớp dense.

- Training: batch size $= 20$ được sử dụng cho tất cả các mạng dò tìm. Để ngăn việc bị overffit, chỉ có 30 epochs, mỗi epoch là 300 batch. Sử dụng Adam optimizer với lr $= 0{,}001$ và cross-entropy loss function. Theo đánh giá sơ bộ thì dropout không cải thiện kết quả.
- Tùy thuộc vào kiến trúc, toàn bộ khóa đào tạo mất khoảng hai đến bốn giờ để hoàn thành trên một GPU Tesla V100.
- Việc tăng trọng số của các chuyển đổi trong hàm mất mát không tạo ra kết quả tốt hơn so với việc hạ thấp ngưỡng chấp nhận θ dưới $0{,}5 \rightarrow$ sử dụng giảm ngưỡng.

ĐÁNH GIÁ

Trong quá trình validation và test, danh sách các ảnh được xây dựng theo cách sau:

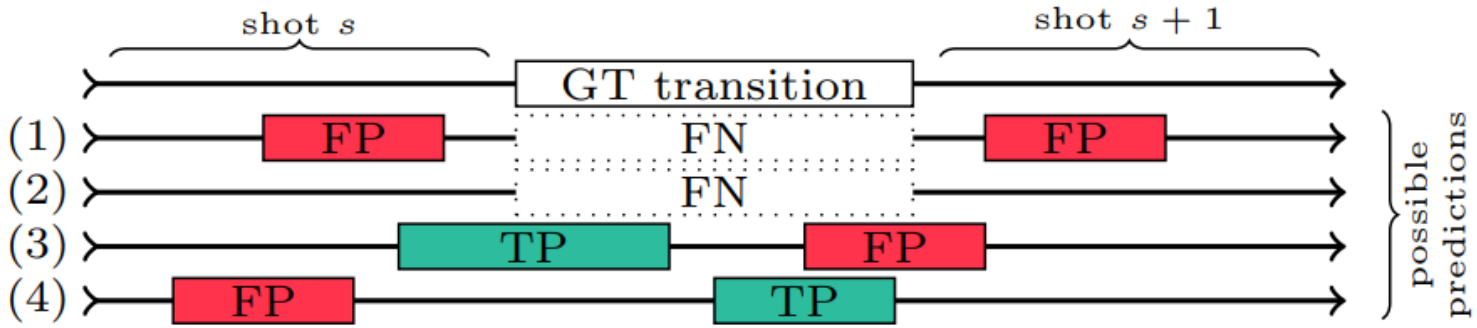
- Ảnh bắt đầu ở frame đầu tiên khi dự đoán giảm xuống dưới ngưỡng θ và kết thúc ở frame đầu tiên khi dự đoán vượt quá θ.
- Lưu ý rằng chỉ các dự đoán cho frame 25-75 được sử dụng do thông tin thời gian không đầy đủ cho các frame đầu tiên / cuối cùng. Do đó, khi xử lý video, cửa sổ đầu vào được dịch chuyển 50 frame giữa các lần chuyển tiếp riêng lẻ qua mạng.

Evaluation metric

- F1 score được sử dụng làm metric đánh giá, được tính là điểm trung bình của từng điểm F1 cho mỗi video. Hình 2 cho thấy các trường hợp khi các bức ảnh được phát hiện xem là True positive, false positive, hoặc false negative.

1: Evaluation metric

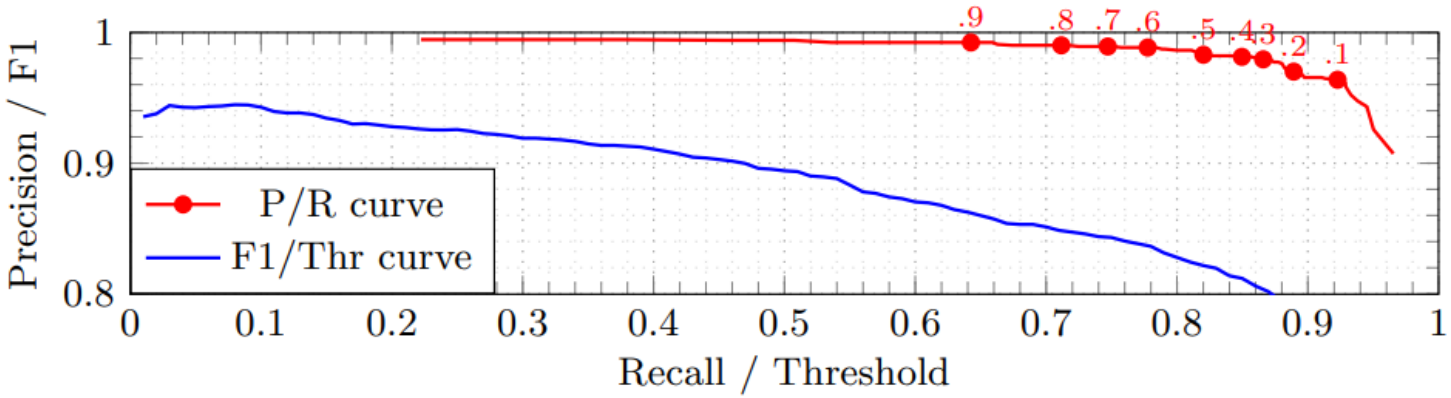
- true positive chỉ được phát hiện nếu chuyển cảnh quay được phát hiện chồng lên ground truth (3, 4 màu xanh lá cây).
- False positive được phát hiện nếu quá trình chuyển đổi được dự đoán không trùng lặp với ground truth (1,4 màu đỏ) hoặc quá trình chuyển đổi được phát hiện lần thứ hai (3 màu đỏ).
- False negative được phát hiện nếu không có quá trình chuyển đổi trùng lặp với ground truth (1, 2 dấu chấm) - ground truth bị bỏ qua.



Hình 2: Hình dung về cách tiếp cận đánh giá. Các chuyển đổi dự đoán được hiển thị bằng các hình chữ nhật có dấu chấm và không rõ ràng.

2: Results

Lưu ý rằng các top-performing weights cho từng cấu hình mô hình đã được chọn dựa trên kết quả trên tập dữ liệu xác thực sau mỗi epoch. Confidence threshold θ cho biết quá trình chuyển đổi được đặt thành $\theta = 0{,}1$ vì nó hoạt động khá tốt đối với hầu hết các mô hình. Ảnh hưởng của θ đến precision, recall và F1 socre được mô tả trong Hình 4.



Hình 2: Precision/Recall cho mô hình hoạt động tốt nhất với các ngưỡng tương ứng θ bên cạnh điểm (màu đỏ) và điểm F1 phụ thuộc vào ngưỡng (màu xanh lam). Được đo trên tập dữ liệu RAI.

2: Results

- Mô hình hoạt động tốt nhất là mô hình có 16 filters ở lớp đầu tiên, hai ô DDCNN xếp chồng lên nhau ở mỗi một trong ba blocks SDDCNN và với 256 neurons trong lớp dense (F = 16, L = 3 S = 2, D = 256).
- Mô hình không cần post-processing.
- Mô hình TransNet hoạt động tốt nhất cũng phải đối mặt với các vấn đề với việc phát hiện một số transitions, chẳng hạn như false positives trong dynamic shots và false negatives trong gradual transitions.