Tóm tắt video không giám sát dựa trên deep reinforment learning với phần thưởng là tính đa dạng và khả năng đại diện

Đặng Văn Minh, Đại học Công nghệ thông tin thành phố Hồ Chí Minh

Abstract

Trong bài báo cáo này, tôi sẽ giới thiệu một mô hình DR-DSN(Kaizang Zhou và cộng sự)[1] để giải quyết bài toán tóm tắt video. Tóm tắt video là tạo ra một video khác từ video ban đầu , video được tạo ra đó ngắn hơn vdeo ban đầu nhưng vẫn súc tíc, đa dạng và đại diện được cho video ban đầu hay nói cách khác bản tóm tắt video ngắn gọn hơn nhưng vẫn giữ được những nội quan trọng cần được truyền tải từ video bàn đầu. Bài toán tóm tắt video được mô hình hóa như một quá trình đưa ra một chuỗi quyết định, quyết định khung hình nào của video sẽ được giữ lại. Mô hình DSN sữ dự đoán xác xuất cho mỗi khung hình trong video, xác xuất này sẽ thể hiện khả năng cho một khung hình sẽ được lựa chọn, từ những khung hình được lựa chọn sẽ tọa ra bản tóm tắt cho video ban đầu. Để huấn luyện mô hình DSN, tác giả sử dụng reinforment learning cùng với hàm phần thưởng được thiết kế dựa trên tính đa dạng và khả năng đại điện. Trong quá trình huấn luyện, hàm phần thưởng sẽ đo mức độ đa dạng và tính đại diện của bản tóm tắt được tạo ra, mô hình DSN sẽ cố gắng làm cho hàm phần thưởng lớn hơn. Mô hình được huấn luyện hoàn toàn không cần nhãn và sự tương tác của con người nên mô hình có thể xem như học không có giám sát.

I. Introduction

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của mạng internet đẫn đến số lượng khổng lồ video được tạo ra, vì vậy lĩnh vực nghiên cứu tóm tắt video được sự quan tâm khá lơn. Nhiều phương pháp, mô hình cũng theo đó được phát triển. Gần đây sự phát triển của recurrent neural network(RNN), đặc biệt là long short-term memory(LSTM) đã được khai thác để mô hình chuỗi khung hình cũng như giải quyết vấn để huấn luyện.

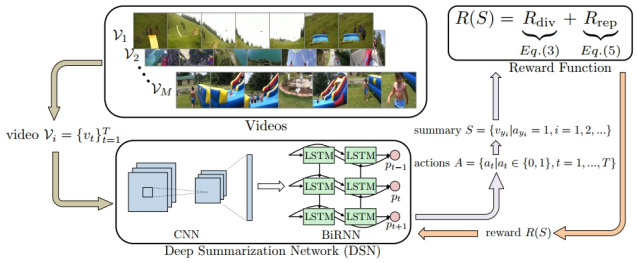
Video là sự kết hợp của nhiều khung hình(frame), vậy nên tóm tắt video là tạo ra video mới ngắn gọn, súc tích bằng cách giữ lại khung hình có thông tin cần thiết, và loại bỏ những khung hình dư thừa không thể hiện nhiều thông tin cần truyền đạt. Đối với mô hình DSN, bài toán tóm tắt video được mô hình hóa như một quá trình đưa ra một chuỗi quyết định. DSN có một kiến trúc mã hóa - giải mã, trong khi mã hóa là convolutional neural network(CNN) có nhiệm vụ trích xuất đặc trưng trong mỗi khung hình video thì giải mã là biderectional LSTM có đầu ra là xác suất, dựa trên xác xuất đó sẽ đưa ra hành động lựa chọn khung hình hay không. Để huấn luyện mô hình DSN tác giả dựa trên nền tảng reinforment learning với hàm phần thưởng được thiết kế dựa trên tính đa dạng và khả năng đại điện của bản tóm tắt đối với video ban đầu.

Dựa trên tiêu chí thông thường của một bản tóm tắt video có chất lượng tốt, hàm phần thưởng được thiết là sự kết hợp của phần thưởng cho tính đa dạng của bản tóm tắt và phần thưởng cho khả năng đại diện của bản tóm tắt. Phần thưởng tính đa dạng sẽ do tính đa dạng giữa những khung hình được chọn với nhau, phần thưởng khả năng đại diện sẽ tính khoảng cách giữa những khung hình với những khung hình được chọn gần nó nhất. Phần thưởng được thiết kế dựa trên sự kết hợp hai tiêu chí là tính đa dạng và khả năng đại diện sẽ làm cho mô hình DSN có khả năng tạo ra bản tóm tắt video đa dạng và đại diện cho video ban đầu.

Tôi sẽ thực nghiệm mô hình trên hai bộ dữ liệu SumMe(Gygli 2014) và TVsum(Song 2015).

II. Methods

Video là sự kết hợp của nhiều khung hình, khi đó tóm tắt video là chọn ra những khung hình quan trọng, và loại bỏ nhữ ng khung hình dư thừa chứa ít thông tin hoặc những thông tin không cần thiết. Bài toán tóm tắt video được mô hình hóa như là một quá trình đưa ra một chuỗi quyết định, tác giả phát triển deep summarization network(DSN) để dự đoán xác suất cho mỗi khung hình(frame) của video và đưa ra quyết định lựa chọn khung hình nào dựa trên phân phối xác suất đó. Để train mô hình DSN tác giả sử dụng reinforment learning. Tác giả thiết kế một hàm phần thưởng để đo chất lượng của bản tóm tắt video, hàm phần thưởng này được thiết kế dự vào hai tiêu chí là sự đa dạng khung hình thể hiện tính đa dạng và tính đại diện cho video ban đầu.



Hình 1: Huấn luyện mô hình DSN dựa trên reinformment learning. DSN sẽ nhận vào một video Vi và đưa ra một chuỗi hành động A, dựa vào đó tạo ra bản tóm tắt S, phần thưởng cho bản tóm tắt sẽ được tính và gửi phản hồi về cho DSN

A. Deep summarization network

DSN là một cấu trúc mã hóa và giải mã, trong đó bộ mã hóa là convolutional neurol network(CNN), có nhiệm vụ trích xuất đặc trưng $\{x_t\}_{t=1}^{T}$ từ những khung hình của video đầu vào $\{v_t\}_{t=1}^{T}$ có độ dài là T. Bộ giải mã là Biderectional recurrent neural network(BiRNN), BiRNN sẽ lấy đầu vào là toàn bộ những đặc trưng $\{x_t\}_{t=1}^{T}$ là đầu ra của bộ mã hóa tương ứng với $\{h_t\}_{t=1}^{T}$, mỗi h_t liên kết với trạng thái ẩn trước h_t^f và trạng thái ẩn sau h_t^b để đóng gói thông tin của những khung hình xung quanh khung hình thứ t, mô hình sẽ kết thúc với hàm sigmoid, hàm này sẽ dự đoán xác suất p_t là khả năng được chọn. Hành động lựa chọn khung hình sẽ có mẫu như sau:

$$p_t = \sigma(Wh_t)$$
$$a_t \approx Benoulli(p_t)$$

Với σ biểu diễn hàm sigmoid, hành động $a_t \in \{0, 1\}$ để diễn tả hành động chọn khung hình đó hay không, một bản tóm tắt video là sự kết hợp của những khung hình được chọn

$$S = \{v_{y_i} | a_{y_i} = 1, i = 1,2,\ldots\}$$

B. Hàm phần thưởng

Trong quá hình huấn luyện, mô hình DSN sẽ nhận phản hồi là một giá trị phần thưởng R(s), phần thưởng này thông báo cho DSN chất lượng của bản tóm tắt video được tạo ra, mục tiêu là huấn luyện mô hình DSN tạo bản tóm tắt video sao cho tối da hóa phần thưởng , hay tạo ra bản tóm tắt video có phần thưởng cao nhất có thể. Hàm phần thưởng đánh giá dựa vào hai tiêu chí là tính đa dạng và khả năng đại diện để từ đó thông tin cần truyền tải của video ban đầu có thể lưu giữ tối đa. Hàm phần thường được kết hợp từ phần thưởng tính đa dạng R_{div} và phần thưởng khả năng đại diện R_{rep}

1) Phần thưởng tính da dạng (diversity reward)

Phần thưởng tính đa dạng sẽ cho thấy mức độ đa dạng của video được tạo ra, nó được tính bằng cách đo sự khác nhau giữa những khung hình được chọn trong không gian đặc tính. Những khung hình được chọn $Y = \{y_i | a_{y_i} = 1, i = 1,2,\ldots |Y|\}$, phần thưởng tính đa dạng R_{div} sẽ được tính như sau:

$$R_{div} = \frac{1}{|Y|(|Y|-1)} \sum_{t \in y} \sum_{\substack{t' \in y \\ t><t'}} d(x_{t'}, x_t)$$

$d(x_{t'}, x_t)$ là hàm được tính bởi:\

$$d(x_{t'}, x_t) = 1 - \frac{x_t^T x_{t'}}{\|x_t\|2\|x_t\|2}$$

Tính đa dạng(sự khác nhau giữa khung hình được chọn) cành cao thì phần thưởng tính đa dạng càng lớn, tuy nhiên sự giống nhau giữa hai khung hình gần nhau nên được bỏ qua để đảm bảo cấu trúc của câu chuyện. Vậy nên $d(x_{t'}, x_t) = 1$ nếu $|t - t'| > \exists$

2) Phần thưởng khả năng đại điện(representativeness reward)

Phần thưởng khả năng đại điện sẽ cho thấy khả năng đại diện của bản tóm tắt cho video ban đầu, phần thưởng khả năng đại diện R_{rep} sẽ được tính như sau:

$$R_{rep} = exp(-\frac{1}{T}\sum_{t=1}^{T} min_{t' \in y}\|x_t - x_{t'}\|_2$$

với hàm phần thưởng được thiết kế như trên, DSN sẽ được khuyến khích chọn những khung khình gần trung tâm trong không gian đặc tính.

3) Huấn luyện mô hình bằng phương pháp policy gradient

Mục tiêu của tác tử là học một hàm policy π_θ bằng cách tối đa hóa phần thưởng mong muốn.

$$J(\theta) = E_{p\theta(a1:T)}[R(S)]$$

Theo thuật toán REINFORCE (Williams 1992) ta có thể tính đạo hàm của hàm $J(\theta)$ như sau:

$$\nabla_\theta J(\theta) = E_{p\theta(a1:T)}\left[R(S)\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|h_t)\right]$$

Với a_t là hành động được đưa ra bởi DSN, và h_t là trạng thái ẩn của biRNN

III. RESULT

Tôi sẽ thực nghiệm mô hình trên hai bộ dữ liệu SumMe(Gygli 2014) và TVsum(Song 2015). Bộ dữ liệu Summe có 25 videos phủ nhiều chủ đề khác nhau, mỗi video có chiều dài từ 1 đến 6 phút, chia 20 video cho huấn luyện, 5 video để kiểm tra. Bộ dữ liệu TVsum có 50 videos, các video có chiều dài từ 2 đến 10 phút, chia 40 video để huấn luyện, 10 video để kiển tra. Mô hình thực nghiệm trên máy tính cá nhân cho ra kết quả như sau:

	Độ chính xác	Thời gian huấn luyện
SumMe	40,9 %	16phút 12 giây
TVsum	53,2 %	56 phút 21 giây

bảng 1: kết quả chạy mô hình DSN trên hai tập dữ liệu SumMe và TVsum

IV. COCLUSION

Trong bài báo cáo đồ án này, tôi đã nêu ra phương pháp DSN(Kaizang Zhou và cộng sự) để giải bài toán tóm tắt video. Mô hinh DSN được huấn luyện dựa trên reinformment learning và hàm điểm thưởng được thiết kế dựa trên hai tiêu chí là tính đa dạng và khả năng đại diện cho vieo ban đầu.

V. REFERENSES

[Kaizang Zhou 2018] Yu Qiao, Tao Xiang, deep reinforment learning fro unsupervised video summarization with diversity-representativeness reward.