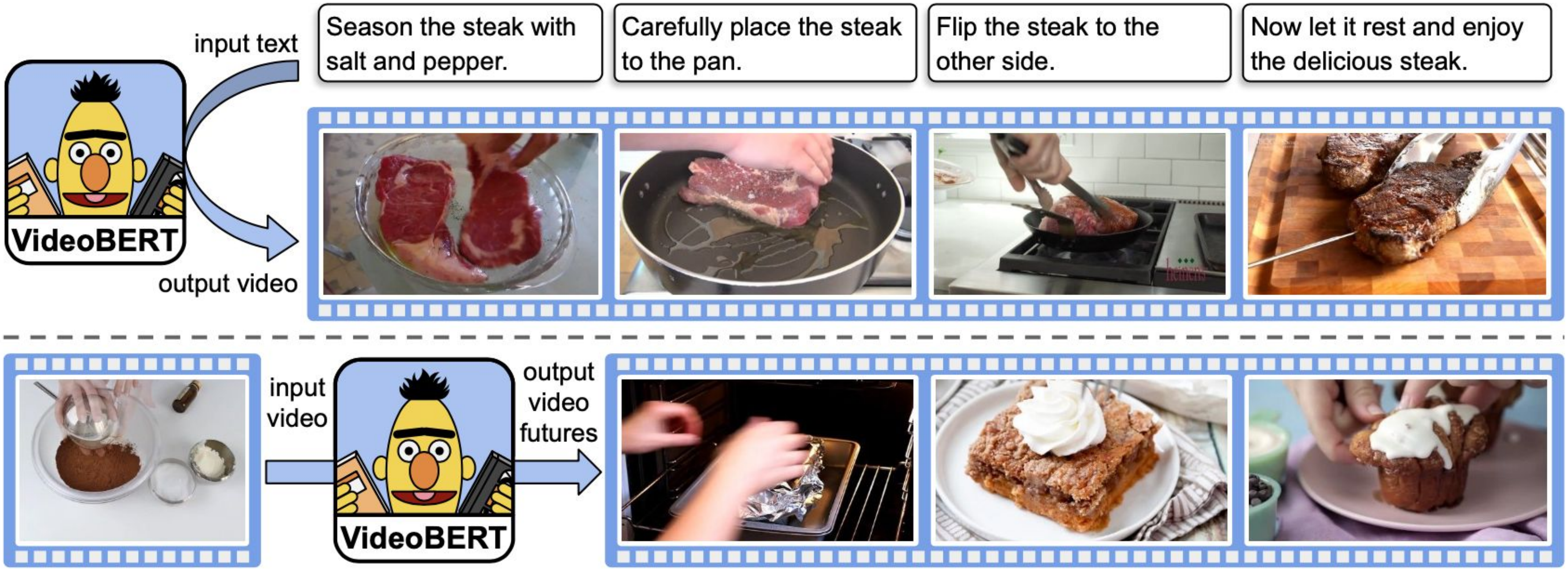
VideoBERT:
A Joint Model for Video and Language Representation Learning

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid

Google Research

발표자 김연수

Contents



- Text-to-Video Generation
- Future Forecasting

Abstract

- joint visual-linguistic model **to learn high-level features** without any explicit supervision
- BERT model to learn bidirectional joint distributions over sequences of visual and linguistic tokens
- 다양한 task에도 적용해봄 (e.g., action classification, video captioning)
- Large training data & cross-modal information are critical to performance
- SOTA on video captioning (YouCook 2 Dataset)

Introduction

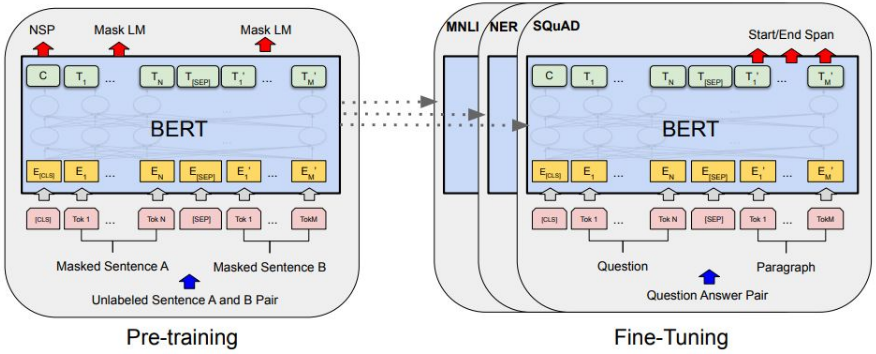
- interested in discovering high-level semantic features
- In this paper, we exploit the key insight that human language has evolved words to describe high-level objects and events, and thus provides a natural source of **“self” supervision.**
- model the relationship between the visual domain & linguistic domain
    - combine 1) automatic speech recognition(ASR) ⇒ speech to text
      2) vector quantization(VQ)
      3) BERT

Introduction

- GOAL
  - apply BERT to learn a model of the form p(x,y)
    - x : sequence of visual words
    - y : sequence of spoken words
- Summary
  - a simple way to learn high level video representation (semantically meaningful & temporally long-range)

Related Work : Self-supervised learning

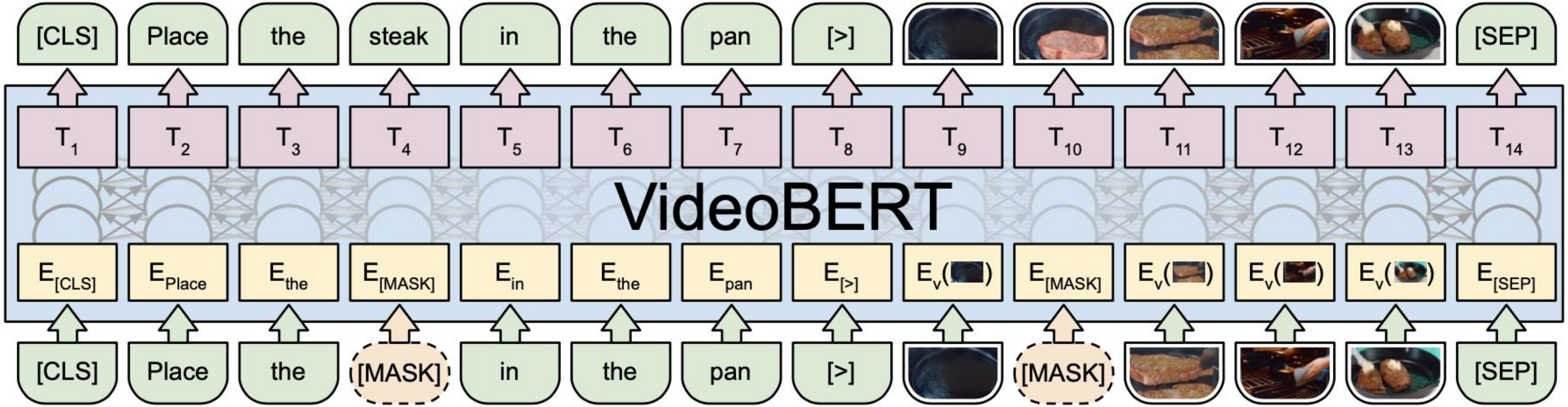
- learn conditional models of the form $p(x_{t+1:T}|x_{1:t})$.
  - partitioning (e.g., gray scale & color, previous frame & next frame)
  - try to predict one from the other
- Our approach uses quantized visual words (instead of pixels)
- BERT



Related Work : Self-supervised learning

- Cross-modal learning
  - most videos contain synchronized **audio and visual signals**, the **two modalities can supervise each other to learn strong self-supervised video representations**

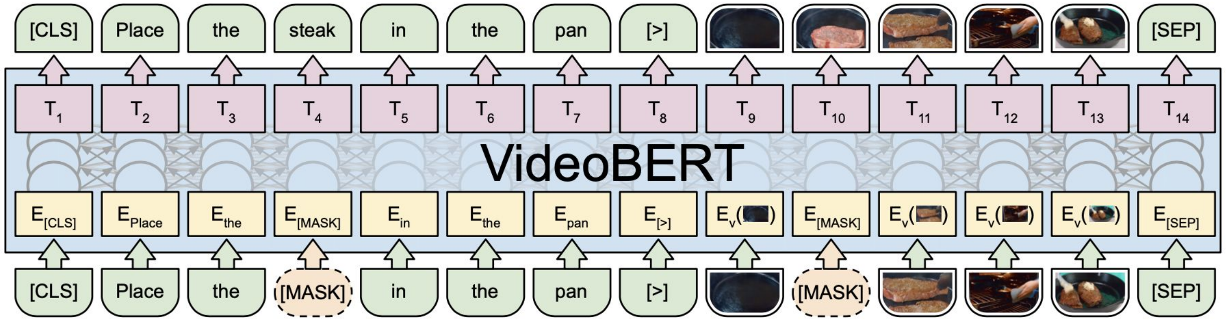
Method


[CLS]
Place
the
steak
in
the
pan
[>]
[SEP]
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
T13
T14
VideoBERT
E[CLS]
EPlace
Ethe
E[MASK]
Ein
Ethe
Epan
E[>]
Ev(
E[MASK]
Ev(
Ev(
Ev(
E[SEP]
[CLS]
Place
the
[MASK]
in
the
pan
[>]
[MASK]
[SEP]

Method

- Video and Language Preprocessing :
  transform the raw visual data into a **discrete sequence of tokens**
  - generate a sequence of "visual words" by applying hierarchical **vector quantization** to features derived from the video using a pretrained model
  - encourages the model to focus on **high level semantics and longer-range temporal dynamics in the video**

Method



- combine the linguistic sentence (derived from the video using ASR) with the visual sentence to generate data
    - e.g., `[CLS] orange chicken with [MASK] sauce [>] v01 [MASK] v08 v72 [SEP]`
    - v01, v08 : visual tokens
    - [>] : 구분자. BERT에서 [SEP]의 역할
      e.g., A [>] B : 자막 [>] 비디오
- **alignment prediciton**

Method

- 3 training region
  - 1. text-only : language-modeling
  - 2. video-only : language model for video
  - 3. video-text : correspondence between the two domains
- After training, downstream tasks : Zero-shot classification, Video Captioning

Experiments

- Dataset
  - We extract a set of publicly available cooking videos from YouTube using the YouTube video annotation system to retrieve **videos with topics related to “cooking” and “recipe”**. We also filter videos by their duration, removing videos longer than 15 minutes, **resulting in a set of 312K videos.** The total duration of this dataset is 23,186 hours, or roughly 966 days
  - To obtain text from the videos, we utilize **YouTube’s automatic speech recognition (ASR) toolkit** provided by the YouTube Data API
  - Evaluation : **YouCook II dataset**, which contains 2000 YouTube videos averaging 5.26 minutes in duration, for a total of 176 hours.

Experiments

- Video and Language Preprocessing
  - extract video feature → S3D
    - pretrain the S3D network on the Kinetics dataset
  - 비디오 프레임과 텍스트 데이터가 모두 토큰화됨
    - 비디오 토큰 = 1.5초(30-frame) 이미지 프레임
    - vector quantization : using hierarchical k-means
    - ASR word sequence
      - tokenizer : WordPieces
      - vocab : BERT 논문에서 사용했던 것과 동일한 것으로 사용

Experiments

- Zero-shot action classification
  - on YouCook 2 Dataset

Method	Supervision	verb top-1 (%)	verb top-5 (%)	object top-1 (%)	object top-5 (%)
S3D [34]	yes	16.1	46.9	13.2	30.9
BERT (language prior)	no	0.0	0.0	0.0	0.0
VideoBERT (language prior)	no	0.4	6.9	7.7	15.3
VideoBERT (cross modal)	no	3.2	43.3	13.1	33.7

Top verbs: make, assemble, prepare
Top nouns: pizza, sauce, pasta

Top verbs: make, do, pour
Top nouns: cocktail, drink, glass

Top verbs: make, prepare, bake
Top nouns: cake, crust, dough

Experiments

- Video Captioning
  - on YouCook 2 Dataset
  - metric : BLEU
  - the best : VideoBERT + S3D

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Zhou *et al.* [39]	7.53	3.84	11.55	27.44	0.38
S3D [34]	6.12	3.24	9.52	26.09	0.31
VideoBERT (video only)	6.33	3.81	10.81	27.14	0.47
VideoBERT	6.80	4.04	11.01	27.50	0.49
VideoBERT + S3D	**7.59**	**4.33**	**11.94**	**28.80**	**0.55**

Table 3: Video captioning performance on YouCook II. We follow the setup from [39] and report captioning performance on the validation set, given ground truth video segments. Higher numbers are better.

GT: add some chopped basil leaves into it
VideoBERT: chop the basil and add to the bowl
S3D: cut the tomatoes into thin slices

GT: cut the top off of a french loaf
VideoBERT: cut the bread into thin slices
S3D: place the bread on the pan

GT: cut yu choy into diagonally medium pieces
VideoBERT: chop the cabbage
S3D: cut the roll into thin slices

GT: remove the calamari and set it on paper towel
VideoBERT: fry the squid in the pan
S3D: add the noodles to the pot

Future work

- we plan to assess our approach on other video understanding tasks, and **on other domains besides cooking.** (For example, we may use the re- cently released COIN dataset of manually labeled instructional videos)
- the future prospects for large scale representation learning from video and language look quite promising.