VideoBERT

A Joint Model for Video and Language Representation Learning

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

Seminar on Current Works in Computer Vision

Presenter: Anna Khatyreva

Advisor: Mohammadreza Zolfaghari

Presentation Overview

- Introduction
- Historical overview
- BERT model
- VideoBERT model
- Summary

Introduction

Motivation

Understand the contents and dynamics of video

Goal

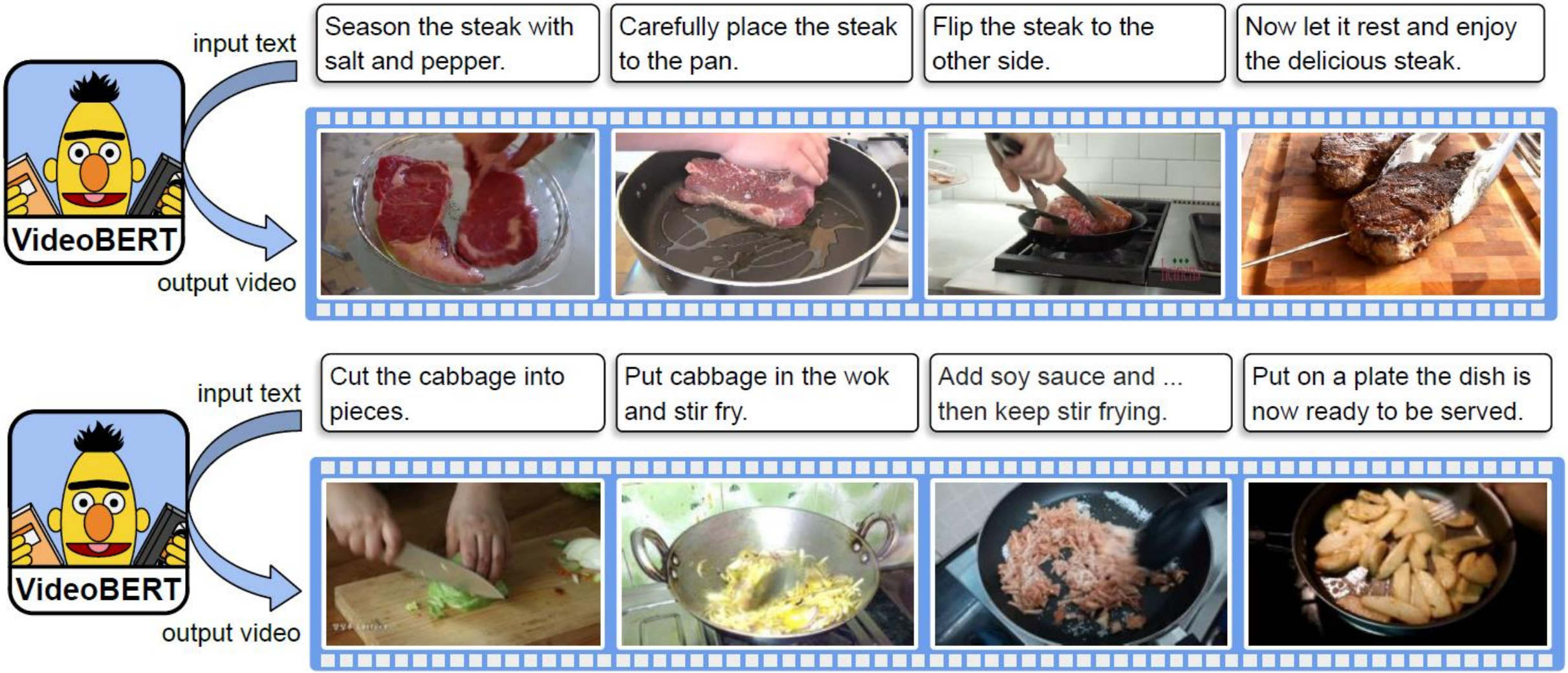
Discover high-level semantic features that correspond to actions and events that unfold over longer time scales

Approach

Learn a bidirectional joint distributions over sequences of visual and linguistic tokens

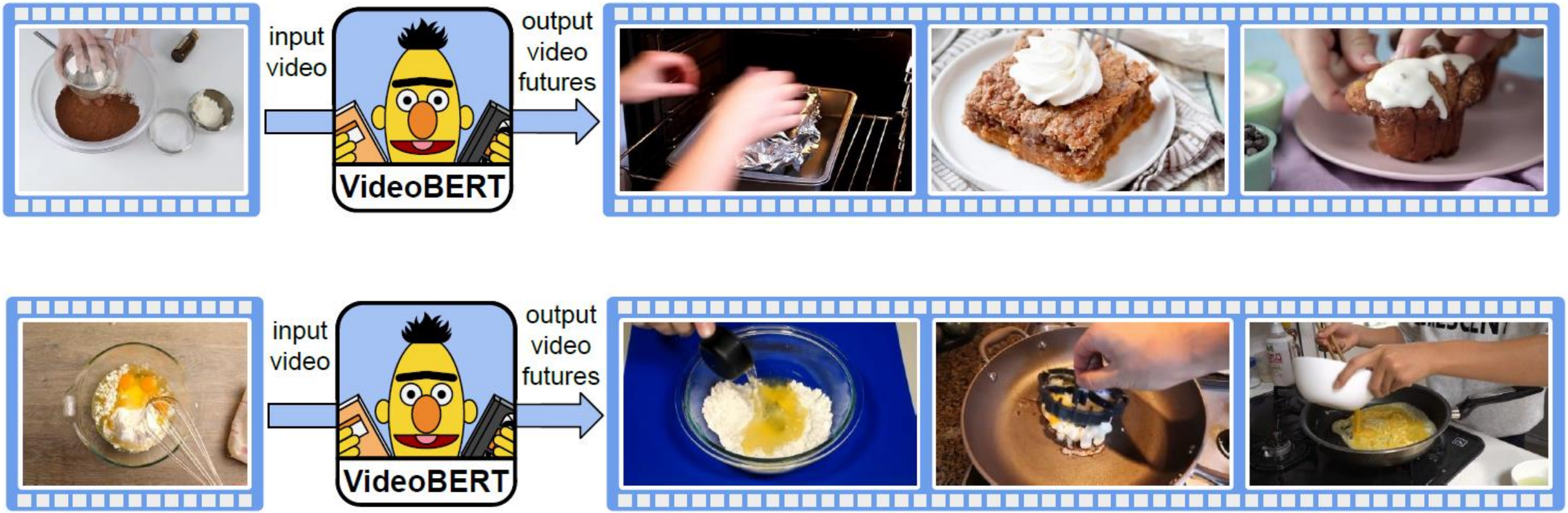
Cut the *lettuce* into pieces

Introduction: Tasks



Source: Sun *et al.* [2019]

Introduction: Tasks



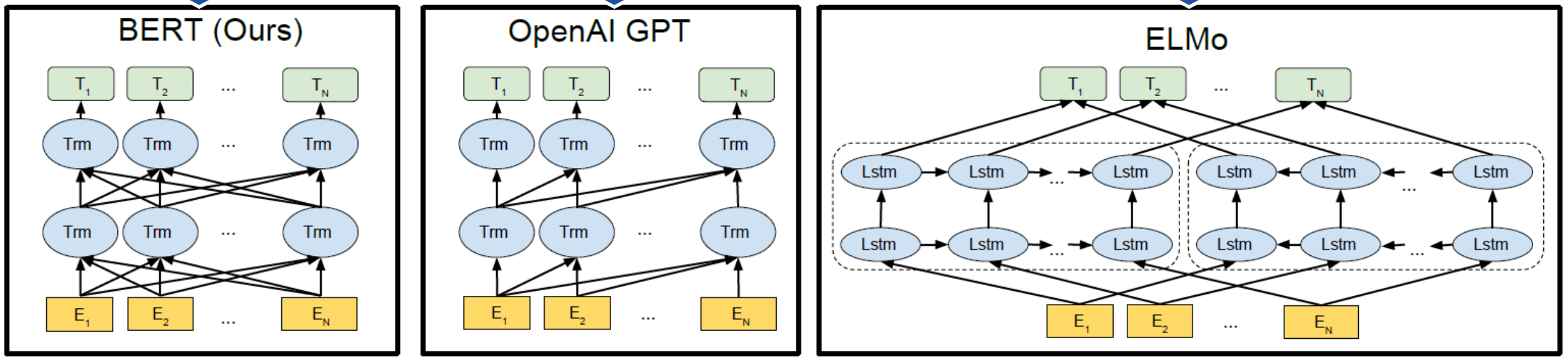
Source: Sun *et al.* [2019]

Historical Overview

Bidirectional Transformer

Left-to-right Transformer

Concatenation of left-to-right and right-to-left LSTMs



Source: Devlin *et al.* [2019]

BERT

Goal: Design a model to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers

Approach: Multi-layer bidirectional Transformer encoder

BERT: Architecture

Architectures

- $BERT_{BASE}$: Number of Transformers layers = 12
  Number of self-attention heads = 12
- $BERT_{LARGE}$: Number of Transformers layers = 24
  Number of self-attention heads = 16

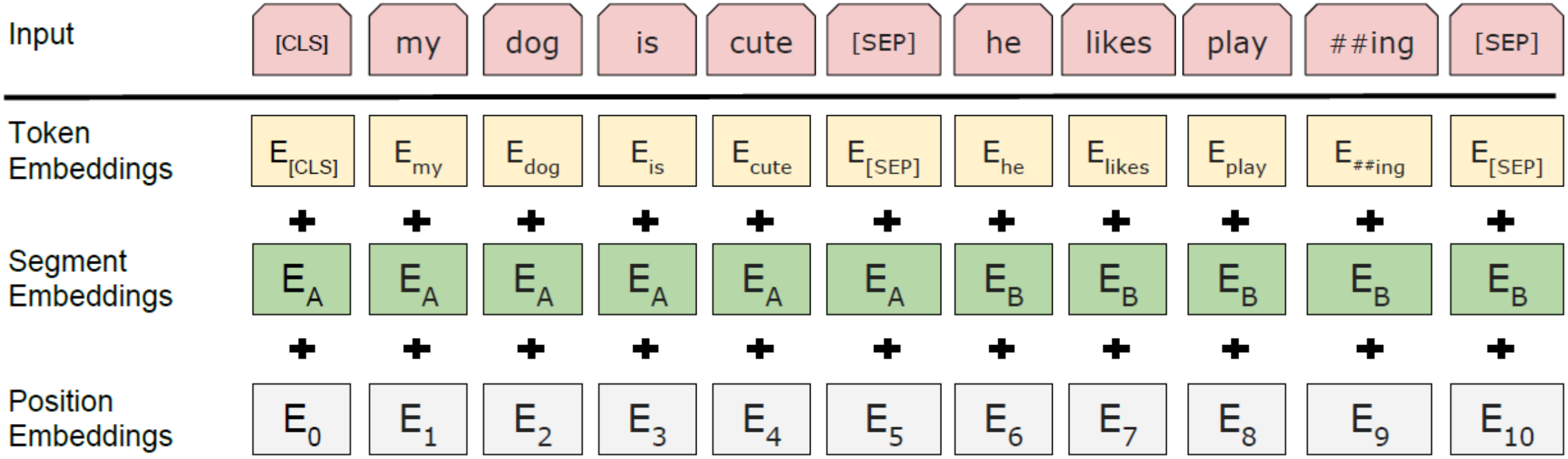
Datasets

- Wikipedia (2.5B words)
- BookCorpus (800M words)

Steps

1. Train a language model on a large unlabeled text corpus
2. Fine-tune this large model to specific NLP tasks

BERT: Input Representation



Source: Devlin *et al.* [2019]

BERT: Masked Language Model

Problem

Bidirectional conditioning would allow each word to trivially predict the target word in a multi-layer context

Solution

Mask out 15% of all WordPiece tokens in each sentence at random

Input:

the man went to [MASK] store

[MASK] – token that replaces “masked” words

BERT: Next Sentence Prediction

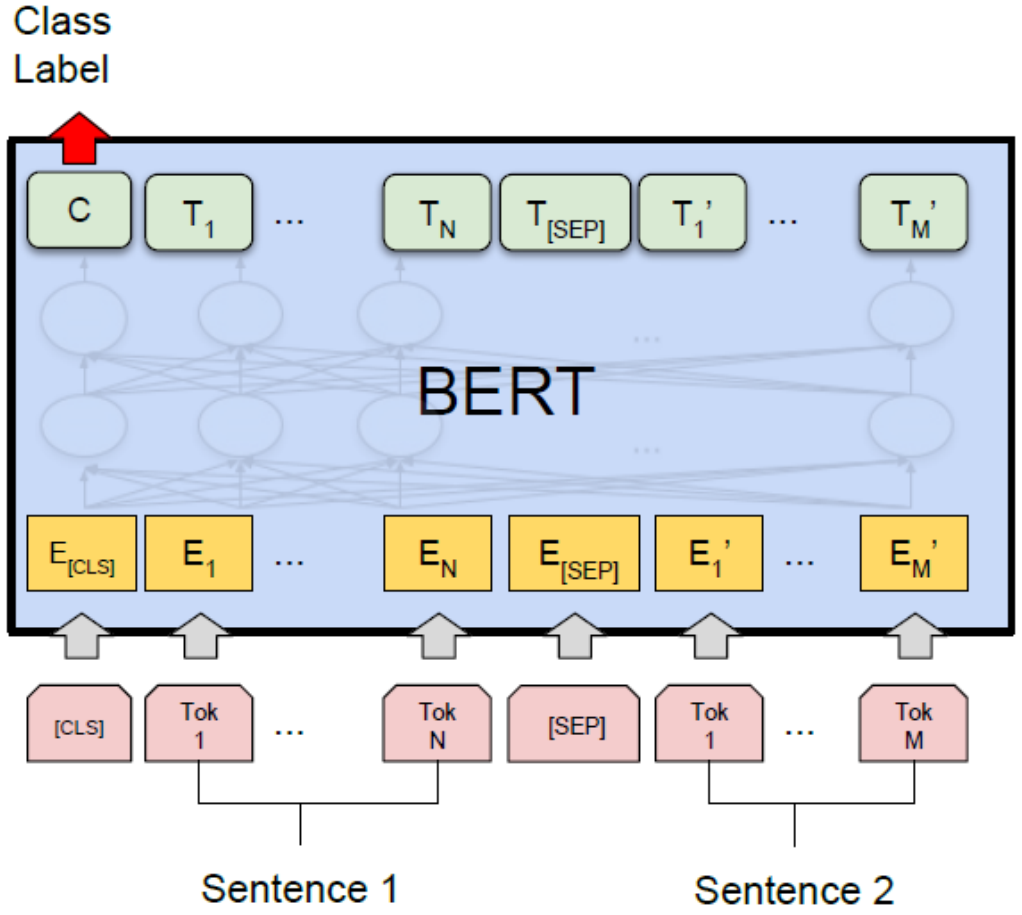
[CLS] – special classification token
[SEP] – special token to separate to different sentences
[MASK] – token that replaces “masked” words
Label = {IsNext, NotNext} – shows if the two sentences are consistent

Input:
[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
Label = IsNext

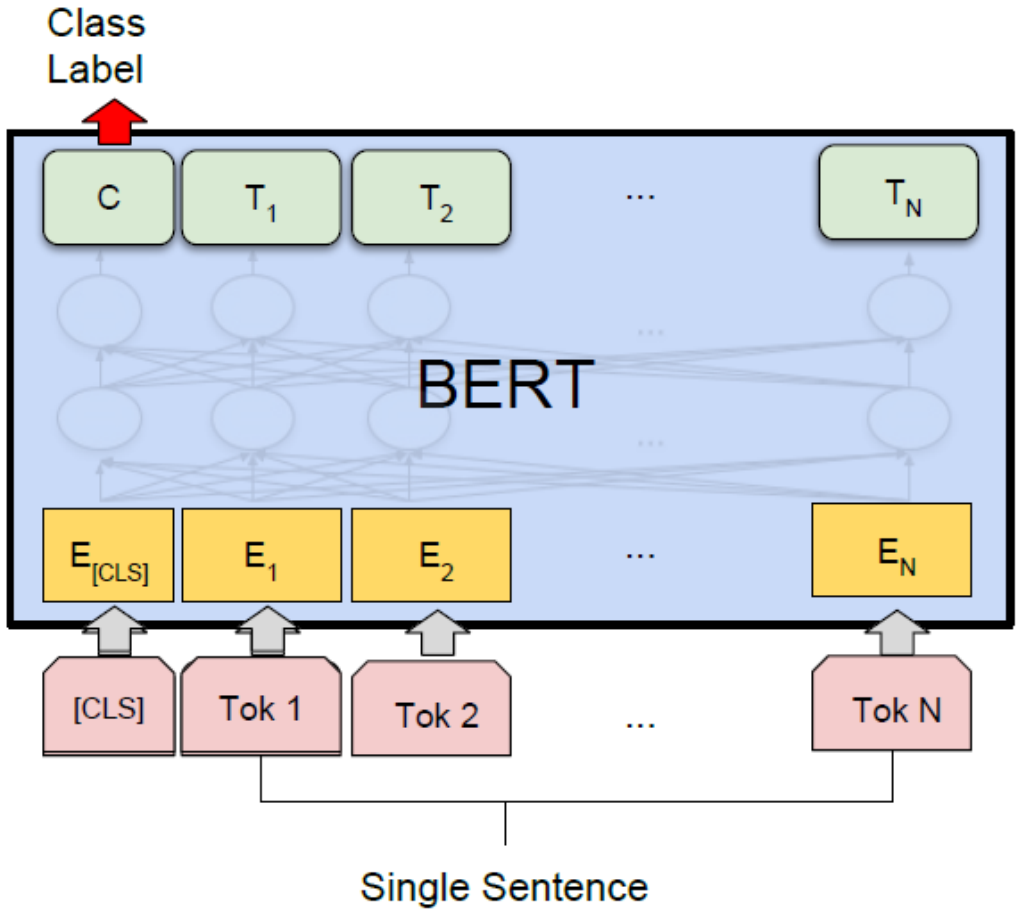
Input:
[CLS] the man [MASK] to the store [SEP]penguin [MASK] are flight ##less birds [SEP]
Label = NotNext

BERT: Fine-tuning

Sentence Pair Classification Tasks



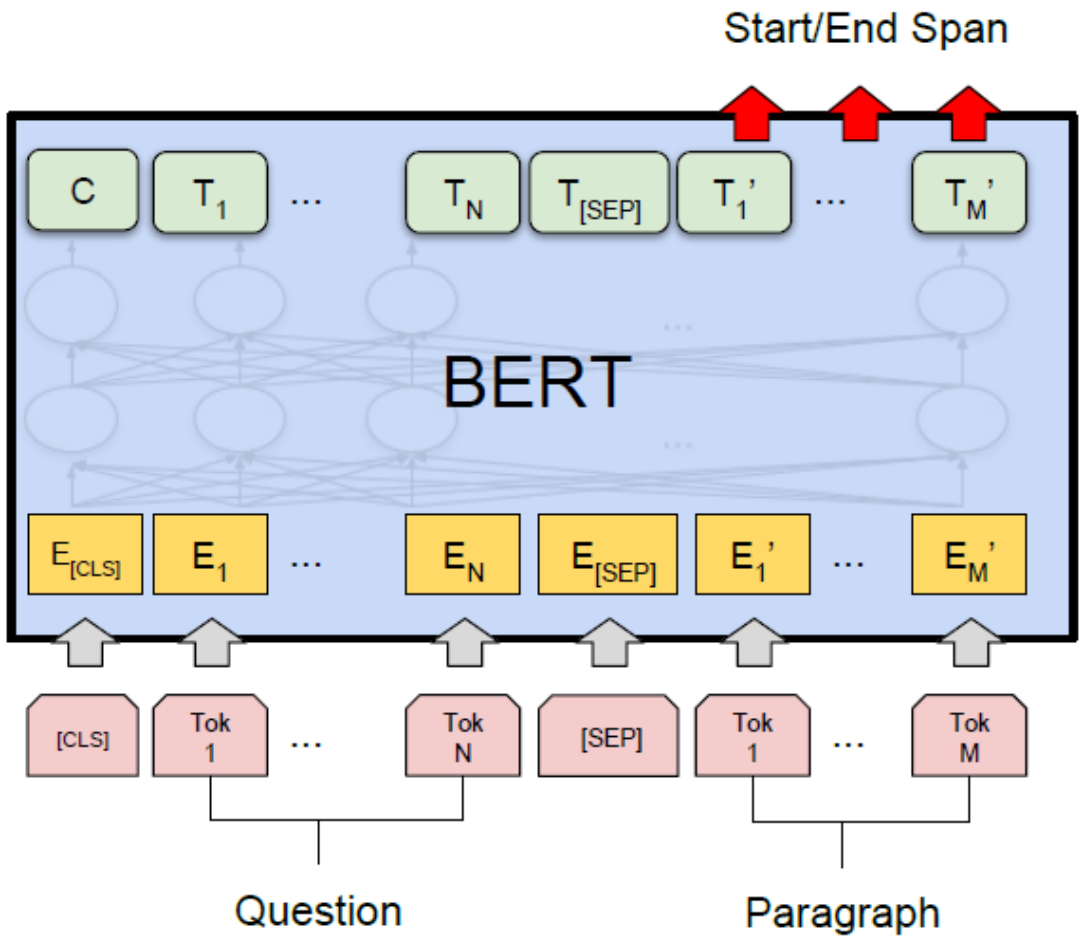
Single Sentence Classification Tasks



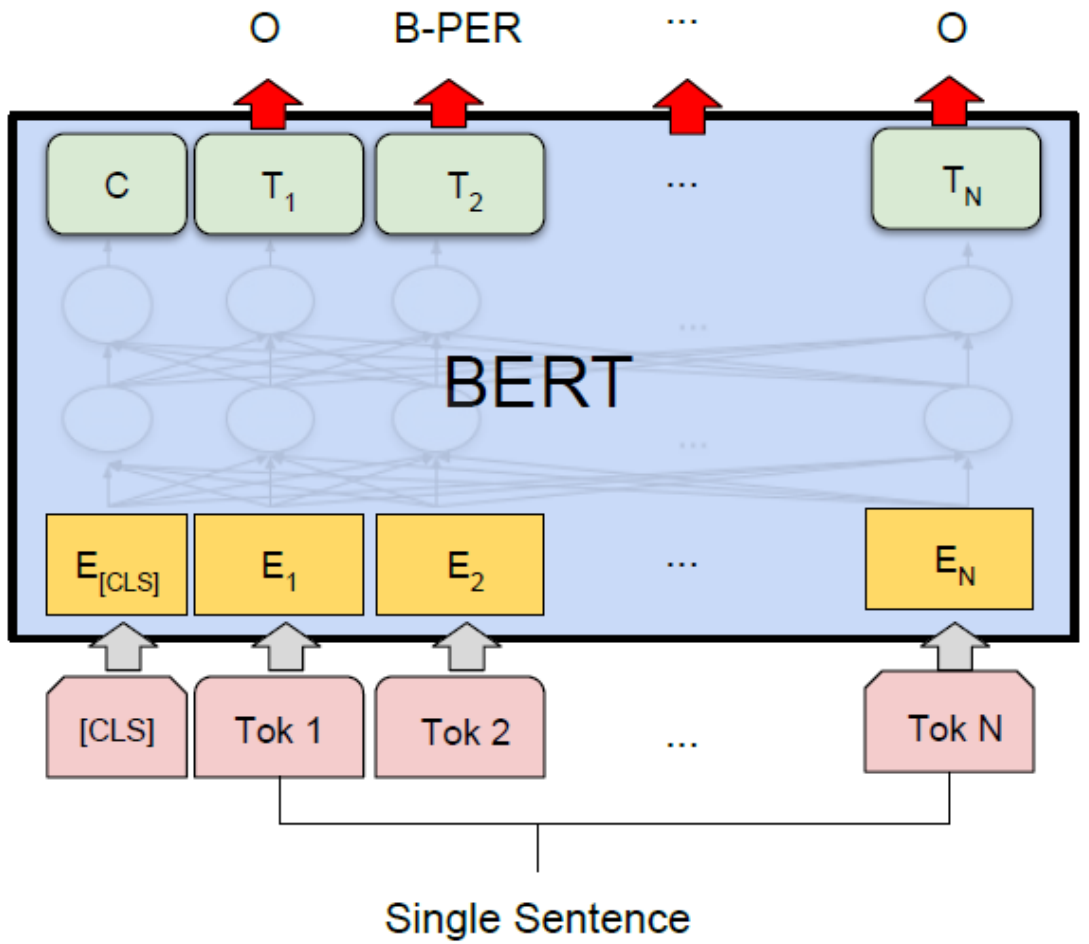
Source: Devlin *et al.* [2019]

BERT: Fine-tuning

Question Answering Tasks



Single Sentence Tagging Tasks



Source: Devlin *et al.* [2019]

VideoBERT

Goal: Find a way to model the relationships between the visual model and the linguistic domain

Approach:

1. Use an automatic speech recognition (ASR) system to convert speech into text
2. Apply vector quantization (VQ) to low-level spatio-temporal visual features derived from pretrained video classification models
3. Use BERT model for learning joint distributions over sequences of discrete tokens

VideoBERT: Datasets

Training

⟹ Focus on cooking videos

⟹ Use videos from YouTube

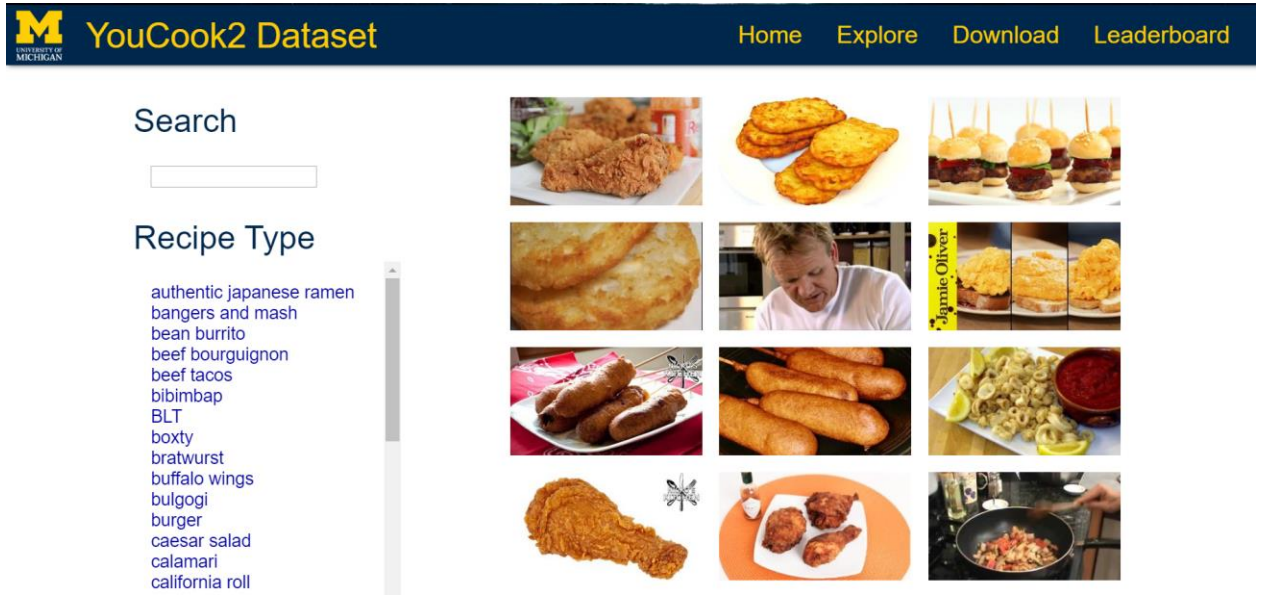
⟹ Use YouTube's ASR toolkit provided by the YouTube Data API to obtain text from videos

Result: 120K videos

Evaluation

⟹ Use YouCook II dataset

Result: 2K videos



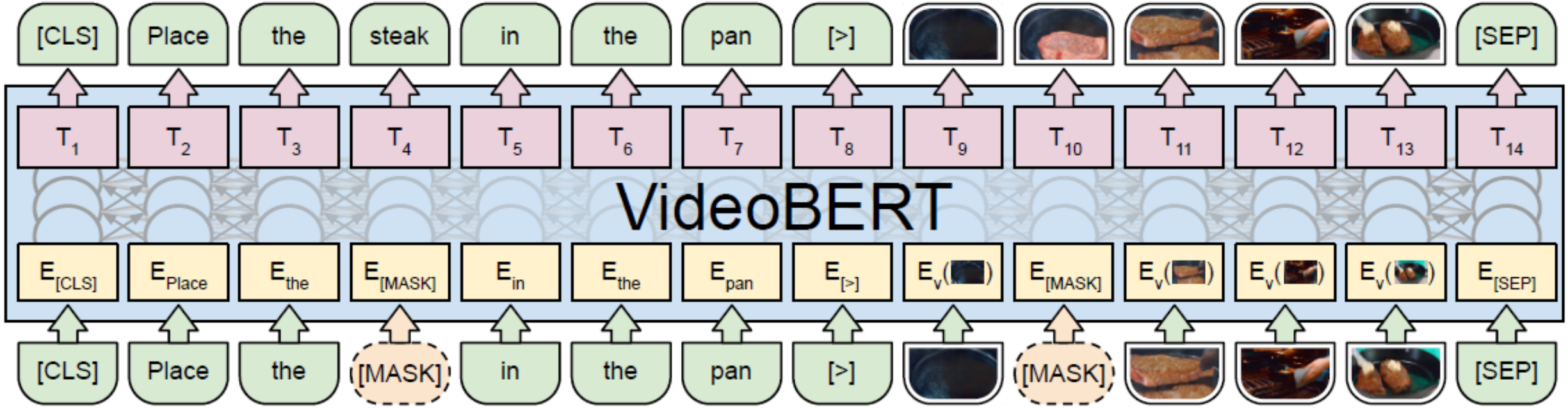
VideoBERT: Masked Language Model

Input:
[CLS]orange chicken with [MASK] sauce [>] v01 [MASK] v08 v72 [SEP]

[CLS] – special classification token
[SEP] – special token to separate to different sentences
[MASK] – token that replaces “masked” words
[>] – special token to combine text and video sentences
v01, v08, v72 – visual tokens

VideoBERT: Masked Language Model

Illustration of VideoBERT in the context of a cloze task



Source: Sun *et al.* [2019]

VideoBERT: Next Sentence Prediction

[CLS] – the first token of a sequence
[SEP] – special token to separate to different sentences
[MASK] – token that replaces “masked” words
$c = \{0, 1\}$ – shows if the two sentences are consistent
v01, v08, v72 – visual tokens
[>] – special token to combine text and video sentences

Input: [CLS] let′s make a traditional [MASK] cuisine [SEP] orange chicken with [MASK] sauce[SEP]

Class label: $c = 1$

VideoBERT: Next Sentence Prediction

Problem

[CLS] token is a noisy indicator, since the speaker may be referring to something that is not visually present

Solution

- Concatenate neighboring sentences into a single long sentence
- Randomly pick a subsampling rate of 1 to 5 steps for the video tokens

VideoBERT: Video and Language Preprocessing

Visual words

1. Sample frames at 20 fps and create clips from 30-frame non-overlapping windows over video

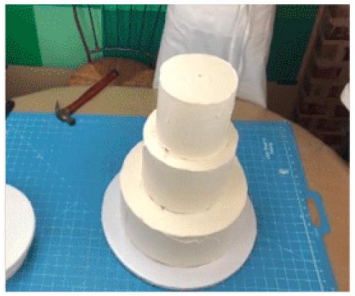
2. For each 30-frame clip, apply S3D network to obtain feature vector

3. Tokenize the features using hierarchical k-means

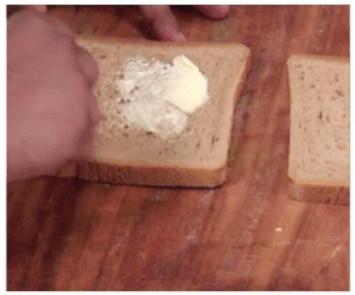
The number of hierarchy levels $d = 4$
The number of clusters per level $k = 12$
$12^4 = 20736$ clusters in total

VideoBERT: Centroids



"but in the meantime, you're just kind of moving around your cake board and you can keep reusing make sure you're working on a clean service so you can just get these all out of your way but it's just a really fun thing to do especially for a birthday party"






"apply a little bit of butter on one side and place a portion of the stuffing and spread evenly cover with another slice of the bread and apply some more butter on top since we're gonna grill the sandwiches"

Source: Sun *et al.* [2019]

VideoBERT: Video and Language Preprocessing

Linguistic sentences

1. Utilize YouTube's ASR toolkit provided by the YouTube Data API to retrieve timestamped speech information
2. Tokenize the text obtained from the video into WordPieces

Visual sentences

1. According to the ASR sentence
2. If ASR sentence is not available, use 16 tokens as a segment

VideoBERT: Zero-shot action classification

Action classification

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


Source: Sun *et al.* [2019]

VideoBERT: Zero-shot action classification

Action classification performance on YouCook II 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

Source: Sun *et al.* [2019]

VideoBERT: Transfer Learning

Idea

Use VideoBERT as a feature extractor

Approach

- Append the video tokens to a template sentence
- Extract the feature for the video tokens and the masked out text tokens

"now let's [MASK] the [MASK] to the [MASK], and then [MASK] the [MASK]"

VideoBERT: Transfer Learning

Video captioning performance on YouCook II

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**

Source: Sun *et al.* [2019]

VideoBERT: Transfer Learning

Source: Sun *et al.* [2019]


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

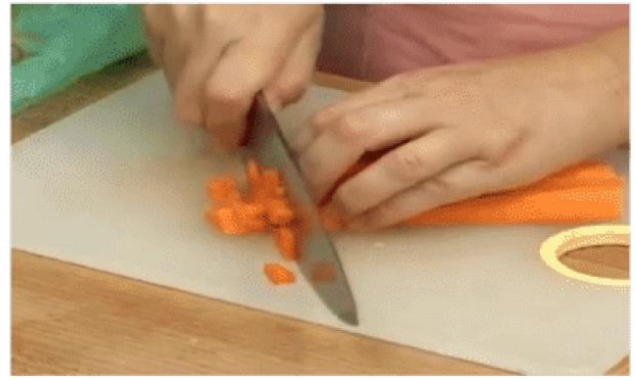

GT: cut the top off of a French loaf
VideoBERT: cut the bread into thin slices
S3D: place the bread on the pan

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

VideoBERT: Tasks

Text-to-video generation

	Retrieved centroid	**Retrieved centroid**
Cut the *steak* into pieces		
Cut the *carrot* into pieces		

Source: Sun *et al.* [2019]

VideoBERT: Tasks

Text-to-video generation

	Retrieved centroid	**Retrieved centroid**
Put the *cookies* into oven		
Put the *pizza* into oven		

Source: Sun *et al.* [2019]

VideoBERT: Tasks

Video-to-text generation

Original



Centroids


ASR:
"This is what happens when you play with dough thinking of yourselves as a kitten who happens to look like Ed Sheeran"

Top verbs: make, shape, roll
Top nouns: dough, filling, chicken

Original



Centroids



ASR:
"I highly recommend that you use a whole sheet just because when you make the smaller sushi rolls, they the fixings tend to fall out"

Top verbs: roll, make, cut
Top nouns: fish, salmon, dough

Source: Sun *et al.* [2019]

VideoBERT: Tasks

Video-to-video generation

Source: Sun *et al.* [2019]

VideoBERT: Impact of the training dataset size

The impact of the training dataset size on the action classification performance on YouCook II dataset

Method	Data size	verb top-1 (%)	verb top-5 (%)	object top-1 (%)	object top-5 (%)
VideoBERT	10K	0.4	15.5	2.9	17.8
VideoBERT	50K	1.1	15.7	8.7	27.3
VideoBERT	100K	2.9	24.5	11.2	30.6
VideoBERT	300K	3.2	43.3	13.1	33.7

Source: Sun *et al.* [2019]

Summary

- VideoBERT can be used in numerous tasks, including action classification and video captioning
- This paper adapts the BERT model to learn a joint visual-linguistic representation for video
- The model is able to learn high level semantic representations and outperforms the state-of-the-art for video captioning on the YouCook II dataset
- The model can be used directly for open-vocabulary classification
- The performance of the model grows monotonically with the size of training set