VideoBERT: A Joint Model for Video and Language Representation Learning

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

Key Points/ Takeaways

- Joint visual-linguistic model.
- Model learns high-level semantic features.
- Outperformed the state-of-the art on video captioning.

Architecture

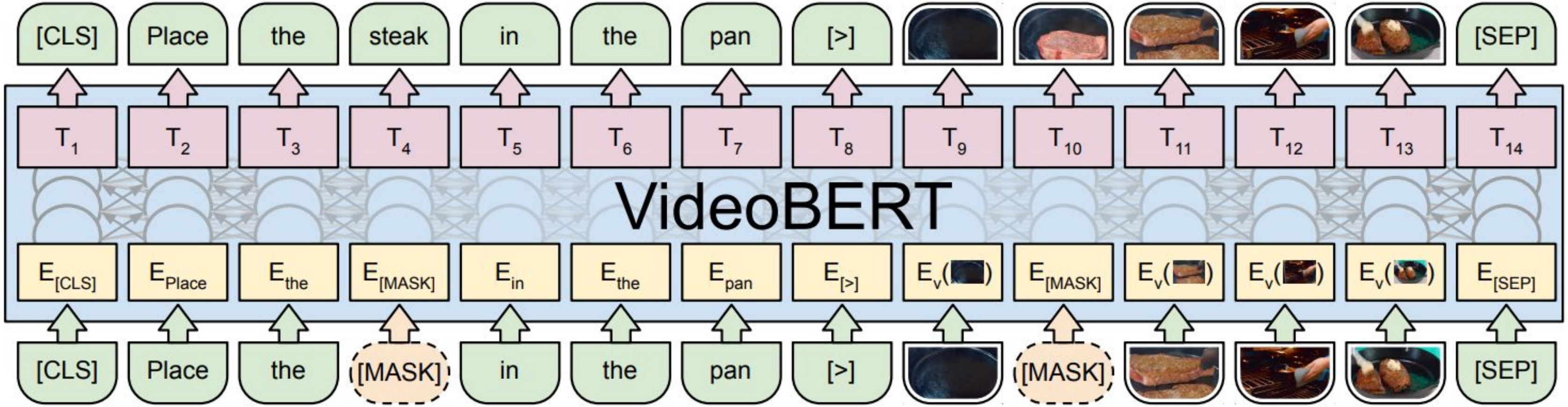


Figure 3: Illustration of VideoBERT in the context of a video and text masked token prediction, or *cloze*, task. This task also allows for training with text-only and video-only data, and VideoBERT can furthermore be trained using a linguistic-visual alignment classification objective (not shown here, see text for details).

Input features

- Sequence of “visual words” obtained by applying hierarchical vector quantization to features derived from the video using a pretrained model.
  - Create 30-frame clips from video
  - Apply a pretrained video ConvNet to extract its features S3D
  - Tokenize the visual features using hierarchical k-means
- For each ASR word sequence,
  - break the stream of words into sentences by adding punctuation using an off-the-shelf LSTM-based language model
  - ASR sentence is associated with starting and ending timestamps

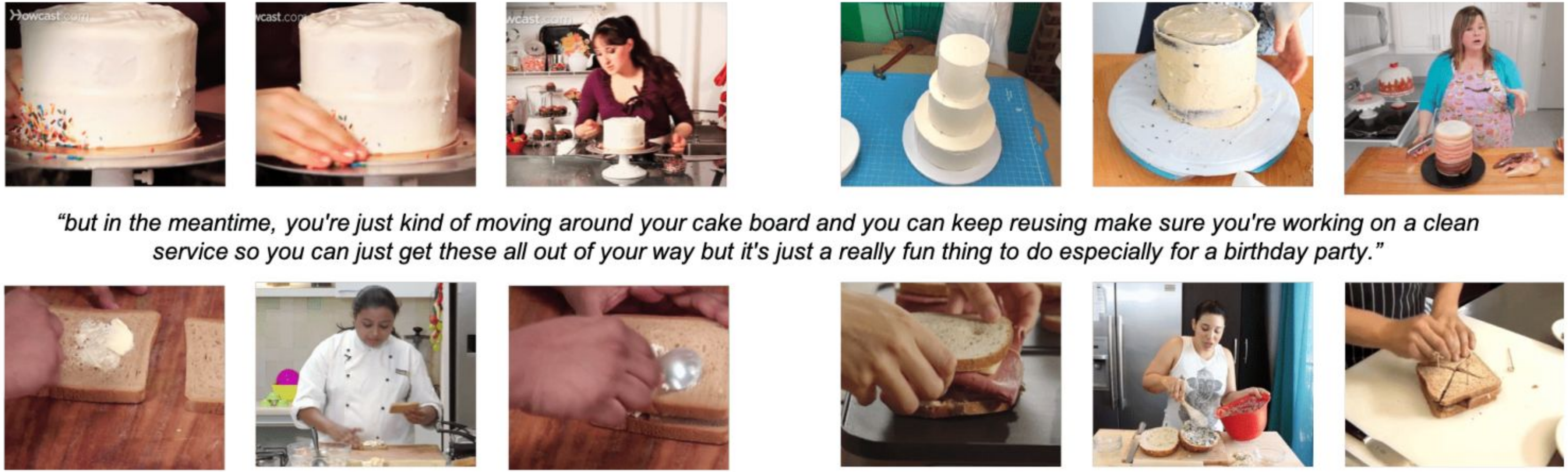
Pretraining

Three training regimes corresponding to the different input data modalities: text-only, video-only and video-text.

For text-only and video-only, the standard mask-completion objectives are used for training the model.

For text-video, we use the linguistic-visual alignment classification objective

Pretraining

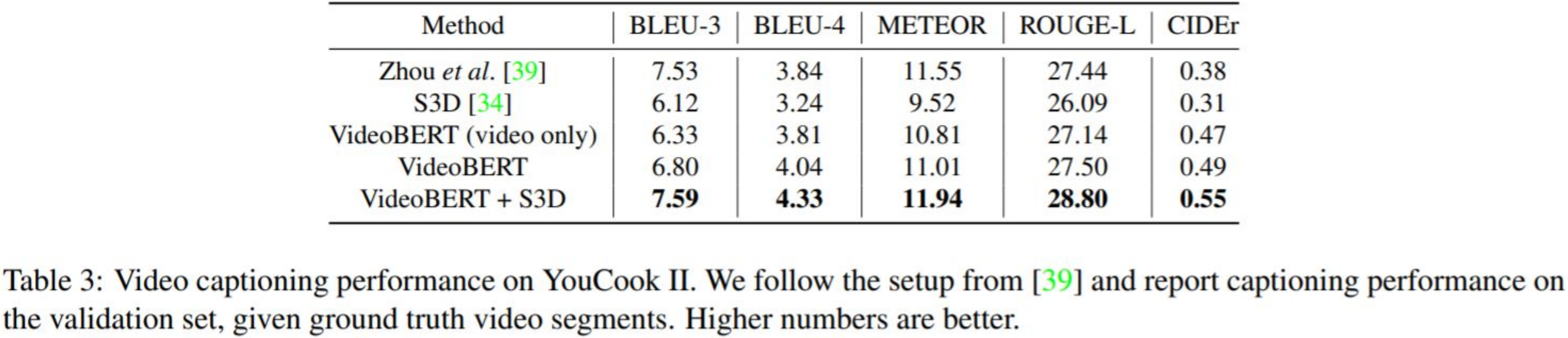


“but in the meantime, you're just kind of moving around your cake board and you can keep reusing make sure you're working on a clean service so you can just get these all out of your way but it's just a really fun thing to do especially for a birthday party.”

“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.”

Figure 4: Examples of video sentence pairs from the pretraining videos. We quantize each video segment into a token, and then represent it by the corresponding visual centroid. For each row, we show the original frames (left) and visual centroids (right). We can see that the tokenization process preserves semantic information rather than low-level visual appearance.

Results



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.

Working of VideoBERT

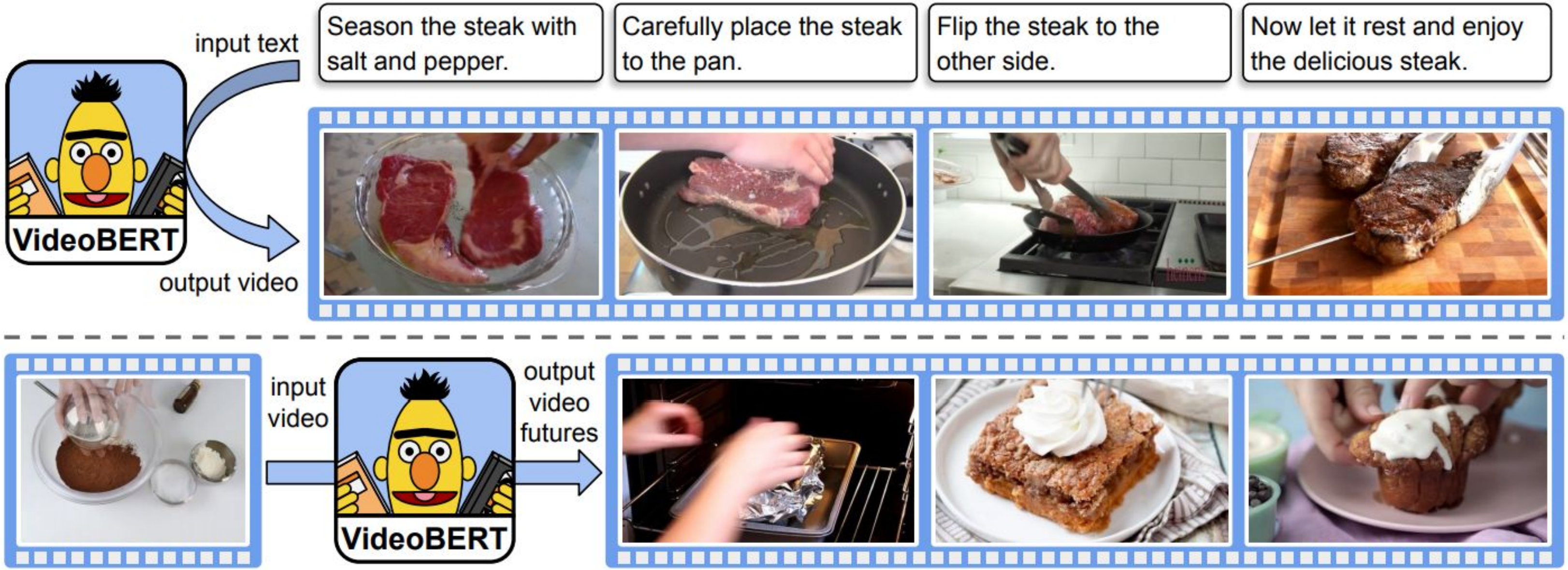

input text
VideoBERT
output video
Season the steak with salt and pepper.
Carefully place the steak to the pan.
Flip the steak to the other side.
Now let it rest and enjoy the delicious steak.
input video
VideoBERT
output video futures

Figure 1: **VideoBERT text-to-video generation and future forecasting.** (Above) Given some recipe text divided into sentences, $y = y_{1:T}$, we generate a sequence of video tokens $x = x_{1:T}$ by computing $x_t^* = \arg\max_k p(x_t = k|y)$ using VideoBERT. (Below) Given a video token, we show the top three future tokens forecasted by VideoBERT at different time scales. In this case, VideoBERT predicts that a bowl of flour and cocoa powder may be baked in an oven, and may become a brownie or cupcake. We visualize video tokens using the images from the training set closest to centroids in feature space.

Training statistics

- Use 4 Cloud TPUs in the Pod configuration with a total batch size of 128
- Train the model for 0.5 million iterations, or roughly 8 epochs.
- Use the Adam optimizer with an initial learning rate of 1e-5, and a linear decay learning rate schedule.
- The training process takes around 2 days.