Video-aided Unsupervised Grammar Induction

Songyang Zhang[1]*, **Linfeng Song**[2], **Lifeng Jin**[2], **Kun Xu**[2], **Dong Yu**[2] and **Jiebo Luo**[1]
[1]University of Rochester, Rochester, NY, USA
szhang83@ur.rochester.edu, jluo@cs.rochester.edu
[2]Tencent AI Lab, Bellevue, WA, USA
{lfsong,lifengjin,kxkunxu,dyu}@tencent.com

Abstract

We investigate video-aided grammar induction, which learns a constituency parser from both unlabeled text and its corresponding video. Existing methods of multi-modal grammar induction focus on learning syntactic grammars from text-image pairs, with promising results showing that the information from static images is useful in induction. However, videos provide even richer information, including not only static objects but also actions and state changes useful for inducing verb phrases. In this paper, we explore rich features (*e.g.* action, object, scene, audio, face, OCR and speech) from videos, taking the recent Compound PCFG model (Kim et al., 2019) as the baseline. We further propose a Multi-Modal Compound PCFG model (MMC-PCFG) to effectively aggregate these rich features from different modalities. Our proposed MMC-PCFG is trained end-to-end and outperforms each individual modality and previous state-of-the-art systems on three benchmarks, *i.e.* DiDeMo, YouCook2 and MSRVTT, confirming the effectiveness of leveraging video information for unsupervised grammar induction.

1 Introduction

Constituency parsing is an important task in natural language processing, which aims to capture syntactic information in sentences in the form of constituency parsing trees. Many conventional approaches learn constituency parser from human-annotated datasets such as Penn Treebank (Marcus et al., 1993). However, annotating syntactic trees by human language experts is expensive and time-consuming, while the supervised approaches are limited to several major languages. In addition, the treebanks for training these supervised parsers are small in size and restricted to the newswire domain, thus their performances tend to be worse when applying to other domains (Fried et al., 2019). To address these issues, recent approaches (Shen et al., 2018b; Jin et al., 2018; Drozdov et al., 2019; Kim et al., 2019) design unsupervised constituency parsers and grammar inducers, since they can be trained on large-scale unlabeled data. In particular, there has been growing interests in exploiting visual information for unsupervised grammar induction because visual information can capture important knowledge required for language learning that is ignored by text (Gleitman, 1990; Pinker and MacWhinney, 1987; Tomasello, 2003). This task aims to learn a constituency parser from raw unlabeled text aided by its visual context.

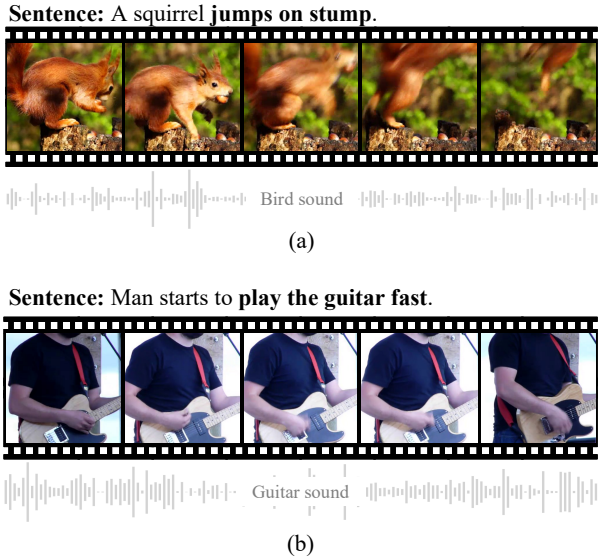

Figure 1: Examples of video aided unsupervised grammar induction. We aim to improve the constituency parser by leveraging aligned video-sentence pairs.

Previous methods (Shi et al., 2019; Kojima et al., 2020; Zhao and Titov, 2020; Jin and Schuler, 2020) learn to parse sentences by exploiting object information from images. However, images are static and cannot present the dynamic interactions among visual objects, which usually correspond to verb phrases that carry important information. There-

* This work was done when Songyang Zhang was an intern at Tencent AI Lab.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1513–1524
June 6–11, 2021.

fore, images and their descriptions may not be fully-representative of all linguistic phenomena encountered in learning, especially when action verbs are involved. For example, as shown in Figure 1(a), when parsing a sentence "A squirrel **jumps on stump**", a single image cannot present the verb phrase "**jumps on stump**" accurately. Moreover, as shown in Figure 1(b), the guitar sound and the moving fingers clearly indicate the speed of music playing, while it is impossible to present only with a static image as well. Therefore, it is difficult for previous methods to learn these constituents, as static images they consider lack dynamic visual and audio information.

In this paper, we address this problem by leveraging video content to improve an unsupervised grammar induction model. In particular, we exploit the current state-of-the-art techniques in both video and audio understanding, domains of which include object, motion, scene, face, optical character, sound, and speech recognition. We extract features from their corresponding state-of-the-art models and analyze their usefulness with the VC-PCFG model (Zhao and Titov, 2020). Since different modalities may correlate with each other, independently modeling each of them may be sub-optimal. We also propose a novel model, Multi-Modal Compound Probabilistic Context-Free Grammars (MMC-PCFG), to better model the correlation among these modalities.

Experiments on three benchmarks show substantial improvements when using each modality of the video content. Moreover, our MMC-PCFG model that integrates information from different modalities further improves the overall performance. Our code is available at https://github.com/Sy-Zhang/MMC-PCFG.

The main contributions of this paper are:

- We are the first to address video aided unsupervised grammar induction and demonstrate that verb related features extracted from videos are beneficial to parsing.
- We perform a thorough analysis on different modalities of video content and propose a model to effectively integrate these important modalities to train better constituency parsers.
- Experiments results demonstrate the effectiveness of our model over the previous state-of-the-art methods.

2 Background and Motivation

Our model is motivated by C-PCFG (Kim et al., 2019) and its variant of the image-aided unsupervised grammar induction model, VC-PCFG (Zhao and Titov, 2020). We will first review the evolution of these two frameworks in Sections 2.1–2.2, and then discuss their limitations in Section 2.3.

2.1 Compound PCFGs

A probabilistic context-free grammar (PCFG) in Chomsky normal form can be defined as a 6-tuple $(S, \mathcal{N}, \mathcal{P}, \Sigma, \mathcal{R}, \Pi)$, where S is the start symbol, $\mathcal{N}, \mathcal{P}$ and Σ are the set of nonterminals, preterminals and terminals, respectively. $\mathcal{R}$ is a set of production rules with their probabilities stored in Π, where the rules include binary nonterminal expansions and unary terminal expansions. Given a certain number of nonterminal and preterminal categories, a PCFG induction model tries to estimate rule probabilities. By imposing a sentence-specific prior on the distribution of possible PCFGs, the compound PCFG model (Kim et al., 2019) uses a mixture of PCFGs to model individual sentences in contrast to previous models (Jin et al., 2018) where a corpus-level prior is used. Specifically in the generative story, the rule probability π_r is estimated by the model g with a latent representation $\mathbf{z}$ for each sentence σ, which is in turn drawn from a prior $p(\mathbf{z})$:

$$\pi_r = g_r(\mathbf{z}; \theta), \qquad \mathbf{z} \sim p(\mathbf{z}). \tag{1}$$

The probabilities for the CFG initial expansion rules $S \rightarrow A$, nonterminal expansion rules $A \rightarrow B\ C$ and preterminal expansion rules $T \rightarrow w$ can be estimated by calculating scores of each combination of a parent category in the left hand side of a rule and all possible child categories in the right hand side of a rule:

$$\begin{aligned}
\pi_{S\rightarrow A} &= \frac{\exp(\mathbf{u}_A^\top f_s([\mathbf{w}_S; \mathbf{z}]))}{\sum_{A' \in \mathcal{N}} \exp(\mathbf{u}_{A'} f_s([\mathbf{w}_S; \mathbf{z}]))}, \\
\pi_{A\rightarrow BC} &= \frac{\exp(\mathbf{u}_{BC}^\top [\mathbf{w}_A; \mathbf{z}])}{\sum_{B',C' \in \mathcal{N} \cup \mathcal{P}} \exp(\mathbf{u}_{B'C'}^\top [\mathbf{w}_A; \mathbf{z}]))}, \\
\pi_{T\rightarrow w} &= \frac{\exp(\mathbf{u}_w^\top f_t([\mathbf{w}_T; \mathbf{z}]))}{\sum_{w' \in \Sigma} \exp(\mathbf{u}_{w'}^T f_t([\mathbf{w}_T; \mathbf{z}]))},
\end{aligned} \tag{2}$$

where $A, B, C \in \mathcal{N}$, $T \in \mathcal{P}$, $w \in \Sigma$, $\mathbf{w}$ and $\mathbf{u}$ vectorial representations of words and categories, and f_t and f_s are encoding functions such as neural networks.

Optimization of the PCFG induction model usually involves maximizing the marginal likelihood of a training sentence $p(\sigma)$ for all sentences in a corpus. In the case of compound PCFGs:

$$\log p_\theta(\sigma) = \log \int_{\mathbf{z}} \sum_{t \in \mathcal{T}_\mathcal{G}(\sigma)} p_\theta(t|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}, \quad (3)$$

where t is a possible binary branching parse tree of σ among all possible trees $\mathcal{T}$ under a grammar $\mathcal{G}$. Since computing the integral over $\mathbf{z}$ is intractable, $\log p_\theta(\sigma)$ can be optimized by maximizing its evidence lower bound $\mathrm{ELBO}(\sigma; \phi, \theta)$:

$$\begin{aligned} \mathrm{ELBO}(\sigma; \phi, \theta) &= \mathbb{E}_{q_\phi(\mathbf{z}|\sigma)}[\log p_\theta(\sigma|\mathbf{z})] \\ &\quad - \mathrm{KL}[q_\phi(\mathbf{z}|\sigma)||p(\mathbf{z})], \end{aligned} \quad (4)$$

where $q_\phi(\mathbf{z}|\sigma)$ is a variational posterior, a neural network parameterized with ϕ. The sample log likelihood can be computed with the inside algorithm, while the KL term can be computed analytically when both prior $p(\mathbf{z})$ and the posterior approximation $q_\phi(\mathbf{z}|\sigma)$ are Gaussian (Kingma and Welling, 2014).

2.2 Visualy Grounded Compound PCFGs

The visually grounded compound PCFGs (VC-PCFG) extends the compound PCFG model (C-PCFG) by including a matching model between images and text. The goal of the vision model is to match the representation of an image $\mathbf{v}$ to the representation of a span $\mathbf{c}$ in a parse tree t of a sentence σ. The word representation $\mathbf{h}_i$ for the ith word is calculated by a BiLSTM network. Given a particular span $c = w_i, \ldots, w_j (0 < i < j \leq n)]$, we then compute its representation $\mathbf{c}$. We first compute the probabilities of its phrasal labels $\{p(k|c, \sigma)|1 \leq k \leq K, K = |\mathcal{N}|\}$, as described in Section 2.1. The representation $\mathbf{c}$ is the sum of all label-specific span representations weighted by the probabilities we predicted:

$$\mathbf{c} = \sum_{k=1}^{K} p(k|c, \sigma) f_k\left(\frac{1}{j-i+1} \sum_{l=i}^{j} \mathbf{h}_l\right), \quad (5)$$

Finally, the matching loss between a sentence σ and an image representation $\mathbf{v}$ can be calculated as a sum over all matching losses between a span and the image representation, weighted by the marginal of a span from the parser:

$$s_{img}(\mathbf{v}, \sigma) = \sum_{c \in \sigma} p(c|\sigma) h_{img}(\mathbf{c}, \mathbf{v}), \quad (6)$$

where $h_{img}(\mathbf{c}, \mathbf{v})$ is a hinge loss between the distances from the image representation $\mathbf{v}$ to the matching and unmatching (*i.e.* sampled from a different sentence) spans $\mathbf{c}$ and $\mathbf{c}'$, and the distances from the span $\mathbf{c}$ to the matching and unmatching (*i.e.* sampled from a different image) image representations $\mathbf{v}$ and $\mathbf{v}'$:

$$\begin{aligned} h_{img}(\mathbf{c},\mathbf{v}) &= \mathbb{E}_{\mathbf{c}'}[\cos(\mathbf{c}', \mathbf{v}) - \cos(\mathbf{c}, \mathbf{v})) + \epsilon]_+ \\ &\quad + \mathbb{E}_{\mathbf{v}'}[\cos(\mathbf{c}, \mathbf{v}') - \cos(\mathbf{c}, \mathbf{v}) + \epsilon]_+, \end{aligned} \quad (7)$$

where ϵ is a positive margin, and the expectations are approximated with one sample drawn from the training data. During training, ELBO and the image-text matching loss are jointly optimized.

2.3 Limitation

VC-PCFG improves C-PCFG by leveraging the visual information from paired images. In their experiments (Zhao and Titov, 2020), comparing to C-PCFG, the largest improvement comes from NPs (+11.9% recall), while recall values of other frequent phrase types (VP, PP, SBAR, ADJP and ADVP) are fairly similar. The performance gain on NPs is also observed with another multi-modal induction model, VG-NSL (Shi et al., 2019; Kojima et al., 2020). Intuitively, image representations from image encoders trained on classification tasks very likely contain accurate information about objects in images, which is most relevant to identifying NPs[1]. However, they provide limited information for phrase types that mainly involve action and change, such as verb phrases. Representations of dynamic scenes may help the induction model to identify verbs, and also contain information about the argument structure of the verbs and nouns based on features of actions and participants extracted from videos. Therefore, we propose a model that induces PCFGs from raw text aided by the multi-modal information extracted from videos, and expect to see accuracy gains on such places in comparison to the baseline systems.

3 Multi-Modal Compound PCFGs

In this section, we introduce the proposed multi-modal compound PCFGs (MMC-PCFG). Instead

[1] Jin and Schuler (2020) reports no improvement on English when incorporating visual information into a similar neural network-based PCFG induction model, which may be because Zhao and Titov (2020) removes punctuation from the training data, which removes a reliable source of phrasal boundary information. This loss is compensated by the induction model with image representations. We leave the study of evaluation configuration on induction results for future work.

of purely relying on object information from images, we generalize VC-PCFG into the video domain, where multi-modal video information is considered. We first introduce the video representation in Section 3.1. We then describe the procedure for matching the multi-modal video representation with each span in Section 3.2. After that we introduce the training and inference details in Section 3.3.

3.1 Video Representation

A video contains a sequence of frames, denoted as $V = \{v_i\}_{i=1}^{L^0}$, where v_i represents a frame in a video and L^0 indicates the total number of frames. We extract video representation from M models trained on different tasks, which are called *experts*. Each expert focuses on extracting a sequence of features of one type. In order to project different expert features into the same dimension, their feature sequences are feed into linear layers (one per expert) with same output dimension. We denote the outputs of the mth expert after projection as $\mathbf{F}^m = \{\mathbf{f}_i^m\}_{i=1}^{L^m}$, where $\mathbf{f}_i^m$ and L^m represent the ith feature and the total number of features of the mth expert, respectively.

A simple method would average each feature along the temporal dimension and then concatenating them together. However, this would ignore the relations among different modalities and the temporal ordering within each modality. In this paper, we use a multi-modal transformer to collect video representations (Gabeur et al., 2020; Lei et al., 2020).

The multi-modal transformer expects a sequence as input, hence we concatenate all feature sequences together and take the form:

$$\mathbf{X} = [\mathbf{f}_{avg}^1, \mathbf{f}_1^1, ..., \mathbf{f}_{L_1}^1, ...\mathbf{f}_{avg}^M, \mathbf{f}_1^M, ..., \mathbf{f}_{L_M}^M], \quad (8)$$

where $\mathbf{f}_{avg}^m$ is the averaged feature of $\{\mathbf{f}_i^m\}_{i=1}^{L_m}$. Each transformer layer has a standard architecture and consists of multi-head self-attention module and a feed forward network (FFN). Since this architecture is permutation-invariant, we supplement it with expert type embeddings $\mathbf{E}$ and positional encoding $\mathbf{P}$ that are added to the input of each attention layer. The expert type embeddings indicate the expert type for input features and take the form:

$$\mathbf{E} = [\mathbf{e}^1, \mathbf{e}^1, ..., \mathbf{e}^1, ..., \mathbf{e}^M, \mathbf{e}^M, ..., \mathbf{e}^M], \quad (9)$$

where $\mathbf{e}^m$ is a learned embedding for the mth expert. The positional encodings indicate the location of each feature within the video and take the form:

$$\mathbf{P} = [\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_{L_1}, ..., \mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_{L_M}], \quad (10)$$

where fixed encodings are used (Vaswani et al., 2017). After that, we collect the output of transformer that corresponds to the averaged features as the final video representation, *i.e.*, $\mathbf{\Psi} = \{\boldsymbol{\psi}_{avg}^i\}_{i=1}^M$. In this way, we can learn more effective video representation by modeling the correlations of features from different modalities and different timestamps.

3.2 Video-Text Matching

To compute the similarity between a video V and a particular span c, a span representation $\mathbf{c}$ is obtained following Section 2.2 and projected to M separate expert embeddings via gated embedding modules (one per expert) (Miech et al., 2018):

$$\begin{aligned} \boldsymbol{\xi}_1^i &= \boldsymbol{W}_1^i \boldsymbol{c} + \boldsymbol{b}_1^i, \\ \boldsymbol{\xi}_2^i &= \boldsymbol{\xi}_1^i \circ sigmoid(\boldsymbol{W}_2^i \boldsymbol{\xi}_1^i + \boldsymbol{b}_2^i), \\ \boldsymbol{\xi}^i &= \frac{\boldsymbol{\xi}_2^i}{\|\boldsymbol{\xi}_2^i\|_2}, \end{aligned} \quad (11)$$

where i is the index of expert, $\boldsymbol{W}_1^i$, $\boldsymbol{W}_2^i$, $\boldsymbol{b}_1^i$, $\boldsymbol{b}_2^i$ are learnable parameters, $sigmoid$ is an element-wise sigmoid activation and $\circ$ is the element-wise multiplication. We denote the set of expert embeddings as $\mathbf{\Xi} = \{\boldsymbol{\xi}^i\}_{i=1}^M$. The video-span similarity is computed as following,

$$\begin{aligned} \omega_i(\mathbf{c}) &= \frac{\exp(\mathbf{u}_i^\top \mathbf{c})}{\sum_{j=1}^M \exp(\mathbf{u}_j^\top \mathbf{c})}, \\ o(\mathbf{\Xi}, \mathbf{\Psi}) &= \sum_{i=1}^M \omega_i(\mathbf{c}) \cos(\boldsymbol{\xi}^i, \boldsymbol{\psi}^i), \end{aligned} \quad (12)$$

where $\{\mathbf{u}_i\}_{i=1}^M$ are learned weights. Given $\mathbf{\Xi'}$, an unmatched span expert embeddings of $\mathbf{\Psi}$, and $\mathbf{\Psi'}$, an unmatched video representation of $\mathbf{\Xi}$, the hinge loss for video is given by:

$$\begin{aligned} h_{vid}(\mathbf{\Xi}, \mathbf{\Psi}) &= \mathbb{E}_{\mathbf{c}'}[o(\mathbf{\Xi'}, \mathbf{\Psi}) - o(\mathbf{\Xi}, \mathbf{\Psi})) + \epsilon]_+ \\ &+ \mathbb{E}_{\mathbf{\Psi}'}[o(\mathbf{\Xi}, \mathbf{\Psi'}) - o(\mathbf{\Xi}, \mathbf{\Psi}) + \epsilon]_+, \end{aligned} \quad (13)$$

where ϵ is a positive margin. Finally the video-text matching loss is defined as:

$$s_{vid}(V, \sigma) = \sum_{c \in \sigma} p(c|\sigma) h_{vid}(\mathbf{\Xi}, \mathbf{\Psi}). \quad (14)$$

Noted that s_{vid} can be regarded as a generalized form of s_{img} in Equation 6, where features from different timestamps and modalities are considered.

3.3 Training and Inference

During training, our model is optimized by the ELBO and the video-text matching loss:

$$\mathcal{L}(\phi,\theta) = \sum_{(V,\sigma)\in\Omega} -\text{ELBO}(\sigma;\phi,\theta)+\alpha s_{vid}(V,\sigma), \tag{15}$$

where α is a hyper-parameter balancing these two loss terms and Ω is a video-sentence pair.

During inference, we predict the most likely tree t^* given a sentence σ without accessing videos. Since computing the integral over $\mathbf{z}$ is intractable, t^* is estimated with the following approximation,

$$\begin{aligned} t^* &= \arg\max_t \int_{\mathbf{z}} p_\theta(t|\mathbf{z}) p_\theta(\mathbf{z}|\sigma) d\mathbf{z} \\ &\approx \arg\max_t p_\theta(t|\sigma, \boldsymbol{\mu}_\phi(\sigma)), \end{aligned} \tag{16}$$

where $\boldsymbol{\mu}_\phi(\sigma)$ is the mean vector of the variational posterior $q_\phi(\mathbf{z}|\sigma)$ and t^* can be obtained using the CYK algorithm (Cocke, 1969; Younger, 1967; Kasami, 1966).

4 Experiments

4.1 Datasets

DiDeMo (Hendricks et al., 2017) collects 10K unedited, personal videos from Flickr with roughly $3-5$ pairs of descriptions and distinct moments per video. There are 32 994, 4 180 and 4 021 video-sentence pairs, validation and testing split.

YouCook2 (Zhou et al., 2018) includes 2K long untrimmed videos from 89 cooking recipes. On average, each video has 6 procedure steps described by imperative sentences. There are 8 713, 969 and 3 310 video-sentence pairs in the training, validation and testing sets.

MSRVTT (Xu et al., 2016) contains 10K videos sourced from YouTube which are accompanied by 200K descriptive captions. There are 130 260, 9 940 and 59 794 video-sentence pairs in the training, validation and testing sets.

4.2 Evaluation

Following the evaluation practice in Zhao and Titov (2020), we discard punctuation and ignore trivial single-word and sentence-level spans at test time. The gold parse trees are obtained by applying a state-of-the-art constituency parser, Benepar (Kitaev and Klein, 2018), on the testing set. All models are run 4 times for 10 epochs with different random seeds. We evaluate both averaged corpus-level F1 (C-F1) and averaged sentence-level F1 (S-F1) numbers as well as their standard deviations.

4.3 Expert Features

In order to capture the rich content from videos, we extract features from the state-of-the-art models of different tasks, including object, action, scene, sound, face, speech, and optical character recognition (OCR). For object and action recognition, we explore multiple models with different architectures and pre-trained dataset. Details are as follows:

Object features are extracted by two models: ResNeXt-101 (Xie et al., 2017), pre-trained on Instagram hashtags (Mahajan et al., 2018) and fine-tuned on ImageNet (Krizhevsky et al., 2012), and SENet-154 (Hu et al., 2018), trained on ImageNet. These datasets include images of common objects, such as, "cock", "kite", and "goose", etc. We use the predicted logits as object features for both models, where the dimension is 1000.

Action features are extracted by three models: I3D trained on Kinetics-400 (Carreira and Zisserman, 2017), R2P1D (Tran et al., 2018) trained on IG-65M (Ghadiyaram et al., 2019) and S3DG (Miech et al., 2020) trained on HowTo100M (Miech et al., 2019). These datasets include videos of human actions, such as "playing guitar", "ski jumping", and "jogging", etc. Following the same processing steps in their original work, we extract the predicted logits as action features, where the dimension is 400 (I3D), 359 (R2P1D) and 512 (S3DG), respectively.

Scene features are extracted by DenseNet-161 (Huang et al., 2017) trained on Places365 (Zhou et al., 2017). Places365 contains images of different scenes, such as "library", "valley", and "rainforest", etc. The predicted logits are used as scene features, where the feature dimension is 365.

Audio features are extracted by VGGish trained on YouTube-$8M$ (Hershey et al., 2017), where the feature dimension is 128. YouTube-$8M$ is a video dataset where different types of sound are involved, such as "piano", "drum", and "violin".

OCR features are extracted by two steps: characters are first recognized by combining text detector Pixel Link (Deng et al., 2018) and text recognizer SSFL (Liu et al., 2018). The characters are then converted to word embeddings through $word2vec$ (Mikolov et al., 2013) as the final OCR features, where the feature dimension is 300.

Face features are extracted by combining face de-

tector SSD (Liu et al., 2016) and face recognizer ResNet50 (He et al., 2016). The feature dimension is 512.

Speech features are extracted by two steps: transcripts are first obtained via Google Cloud Speech to Text API. The transcripts are then converted to word embeddings through $word2vec$ (Mikolov et al., 2013) as the final speech features, where the dimension is 300.

4.4 Implementation Details

We keep sentences with fewer than 20 words in the training set due to the computational limitation. After filtering, the training sets cover 99.4%, 98.5% and 97.1% samples of their original splits in DiDeMo, YouCook2 and MSRVTT.

We train baseline models, C-PCFG and VC-PCFG, with same hyper parameters suggested in Kim et al. (2019); Zhao and Titov (2020). Our MMC-PCFG is composed of a parsing model and a video-text matching model. The parsing model has the same parameters as VC-PCFG (please refer to their paper for details). For video-text matching model, all extracted expert features are projected to 512-dimensional vectors. The transformer has 2 layers, a dropout probability of 10%, a hidden size of 512 and an intermediate size of 2048. We select the top-2000 most common words as vocabulary for all datasets. All the baseline methods and our models are optimized using Adam (Kingma and Ba, 2015) with the learning rate set to 0.001, $\beta_1 = 0.75$ and $\beta_2 = 0.999$. All parameters are initialized with Xavier uniform initializer (Glorot and Bengio, 2010). The batch size is set to 16.

Due to the long video durations, it is infeasible to feed all features into the multi-modal transformer. Therefore, each feature from object, motion and scene categories is partitioned into 8 chunks and then average-pooled within each chunk. For features from other categories, global average pooling is applied. In this way, the coarse-grained temporal information is preserved. Noted that some videos do not have audio and some videos do not have detected faces or text characters. For these missing features, we pad them with zeros. All the aforementioned expert features are obtained from Albanie et al. (2020).

4.5 Main Results

We evaluate the proposed MMC-PCFG approach on three datasets, and compare it with recently proposed state-of-the-art methods, C-PCFG (Kim et al., 2019) and VC-PCFG (Zhao and Titov, 2020). The results are summarized in Table 1. The values high-lighted by **bold** and *italic* fonts indicate the top-2 methods, respectively. All results are reported in percentage (%). LBranch, RBranch and Random represent left branching trees, right branching trees and random trees, respectively. Since VC-PCFG is originally designed for images, it is not directly comparable with our method. In order to allow VC-PCFG to accept videos as input, we average video features in the temporal dimension first and then feed them into the model. We evaluate VC-PCFG with 10, 7, and 10 expert features for DiDeMo, YouCook2 and MSRVTT, respectively. In addition, we also include the concatenated averaged features (Concat). Since object and action categories involve more than one expert, we directly use experts' names instead of their categories in Table 1.

Overall performance comparison. We first compare the overall performance, *i.e.*, C-F1 and S-F1, among all models, as shown in Table 1. The right branching model serves as a strong baseline, since English is a largely right-branching language. C-PCFG learns parsing purely based on text. Compared to C-PCFG, the better overall performance of VC-PCFG demonstrates the effectiveness of leveraging video information. Compared within VC-PCFG, concatenating all features together may not even outperform a model trained on a single expert (R2P1D *v.s.* Concat in DiDeMo and MSRVTT). The reason is that each expert is learned independently, where their correlations are not considered. In contrast, our MMC-PCFG outperforms all baselines on C-F1 and S-F1 in all datasets. The superior performance indicates that our model can leverage the benefits from all the experts[2]. Moreover, the superior performance over Concat demonstrates the importance of modeling relations among different experts and different timestamps.

Performance comparison among different phrase types. We compare the models' recalls on top-3 frequent phrase types (NP, VP and PP). These three types cover 77.4%, 80.1% and 82.4% spans of gold trees on DiDeMo, YouCook2 and MSRVTT, respectively. In the following, we compare their performance on DiDeMo, as shown in Table 1. Comparing VC-PCFG trained

[2]The larger improvement on DiDeMo may be caused by the diversity of the video content. Videos in DiDeMo are more diverse in scenes, actions and objects, which provide a great opportunity for leveraging video information.

	Method	DiDeMo					YouCook2		MSRVTT	
		NP	VP	PP	C-F1	S-F1	C-F1	S-F1	C-F1	S-F1
	LBranch	41.7	0.1	0.1	16.2	18.5	6.8	5.9	14.4	16.8
	RBranch	32.8	**91.5**	**66.5**	*53.6*	*57.5*	35.0	41.6	54.2	58.6
	Random	$36.5_{\pm 0.6}$	$30.5_{\pm 0.5}$	$30.1_{\pm 0.5}$	$29.4_{\pm 0.3}$	$32.7_{\pm 0.5}$	$21.2_{\pm 0.2}$	$24.0_{\pm 0.2}$	$27.2_{\pm 0.1}$	$30.5_{\pm 0.1}$
	C-PCFG	$\mathbf{72.9}_{\pm 5.5}$	$16.5_{\pm 6.2}$	$23.4_{\pm 16.9}$	$38.2_{\pm 5.0}$	$40.4_{\pm 4.1}$	$37.8_{\pm 6.7}$	$41.4_{\pm 6.6}$	$50.7_{\pm 3.2}$	$55.0_{\pm 3.2}$
VC-PCFG	ResNeXt	$64.4_{\pm 21.4}$	$25.7_{\pm 17.7}$	$34.6_{\pm 25.0}$	$40.0_{\pm 13.7}$	$41.8_{\pm 14.0}$	$38.2_{\pm 8.3}$	$42.8_{\pm 8.4}$	$50.7_{\pm 1.7}$	$54.9_{\pm 2.2}$
VC-PCFG	SENet	$70.5_{\pm 15.3}$	$25.7_{\pm 15.9}$	$36.5_{\pm 24.6}$	$42.6_{\pm 10.4}$	$44.0_{\pm 10.4}$	$39.9_{\pm 8.7}$	$44.9_{\pm 8.3}$	$52.2_{\pm 1.2}$	$56.0_{\pm 1.6}$
VC-PCFG	I3D	$57.9_{\pm 13.5}$	$45.7_{\pm 14.1}$	$45.8_{\pm 17.2}$	$45.1_{\pm 6.0}$	$49.2_{\pm 6.0}$	$40.6_{\pm 3.6}$	$45.7_{\pm 3.2}$	$54.5_{\pm 1.6}$	$59.1_{\pm 1.7}$
VC-PCFG	R2P1D	$61.2_{\pm 8.5}$	$38.1_{\pm 5.4}$	$62.1_{\pm 4.1}$	$48.1_{\pm 4.4}$	$50.7_{\pm 4.2}$	$39.4_{\pm 8.1}$	$44.4_{\pm 8.3}$	$54.0_{\pm 2.5}$	$58.0_{\pm 2.3}$
VC-PCFG	S3DG	$61.3_{\pm 13.4}$	$31.7_{\pm 16.7}$	$51.8_{\pm 8.0}$	$44.0_{\pm 2.7}$	$46.5_{\pm 5.1}$	$39.3_{\pm 6.5}$	$44.1_{\pm 6.6}$	$50.7_{\pm 3.2}$	$54.7_{\pm 2.9}$
VC-PCFG	Scene	$62.2_{\pm 9.6}$	$30.6_{\pm 12.3}$	$41.1_{\pm 24.8}$	$41.7_{\pm 6.5}$	$44.9_{\pm 7.4}$	–	–	$54.6_{\pm 1.5}$	$58.4_{\pm 1.3}$
VC-PCFG	Audio	$64.2_{\pm 18.6}$	$21.3_{\pm 26.5}$	$34.7_{\pm 11.0}$	$38.7_{\pm 3.7}$	$39.5_{\pm 5.2}$	$39.2_{\pm 4.7}$	$43.3_{\pm 4.9}$	$52.8_{\pm 1.3}$	$56.7_{\pm 1.4}$
VC-PCFG	OCR	$64.4_{\pm 15.0}$	$27.4_{\pm 19.5}$	$42.8_{\pm 31.2}$	$41.9_{\pm 16.9}$	$44.6_{\pm 17.5}$	$38.6_{\pm 5.5}$	$43.2_{\pm 5.6}$	$51.0_{\pm 3.0}$	$55.5_{\pm 3.0}$
VC-PCFG	Face	$60.8_{\pm 16.0}$	$31.5_{\pm 17.0}$	$52.8_{\pm 9.8}$	$43.9_{\pm 4.5}$	$46.3_{\pm 5.5}$	–	–	$50.5_{\pm 2.6}$	$54.5_{\pm 2.6}$
VC-PCFG	Speech	$61.8_{\pm 12.8}$	$26.6_{\pm 17.6}$	$43.8_{\pm 34.5}$	$40.9_{\pm 16.0}$	$43.1_{\pm 16.1}$	–	–	$51.7_{\pm 2.6}$	$56.2_{\pm 2.5}$
VC-PCFG	Concat	$68.6_{\pm 8.6}$	$24.9_{\pm 19.9}$	$39.7_{\pm 19.5}$	$42.2_{\pm 12.3}$	$43.2_{\pm 14.2}$	$42.3_{\pm 5.7}$	$47.0_{\pm 5.6}$	$49.8_{\pm 4.1}$	$54.2_{\pm 4.0}$
	MMC-PCFG	$67.9_{\pm 9.8}$	$52.3_{\pm 9.0}$	$63.5_{\pm 8.6}$	$\mathbf{55.0}_{\pm 3.7}$	$\mathbf{58.9}_{\pm 3.4}$	$\mathbf{44.7}_{\pm 5.2}$	$\mathbf{48.9}_{\pm 5.7}$	$\mathbf{56.0}_{\pm 1.4}$	$\mathbf{60.0}_{\pm 1.2}$

Table 1: Performance comparison on three benchmark datasets.

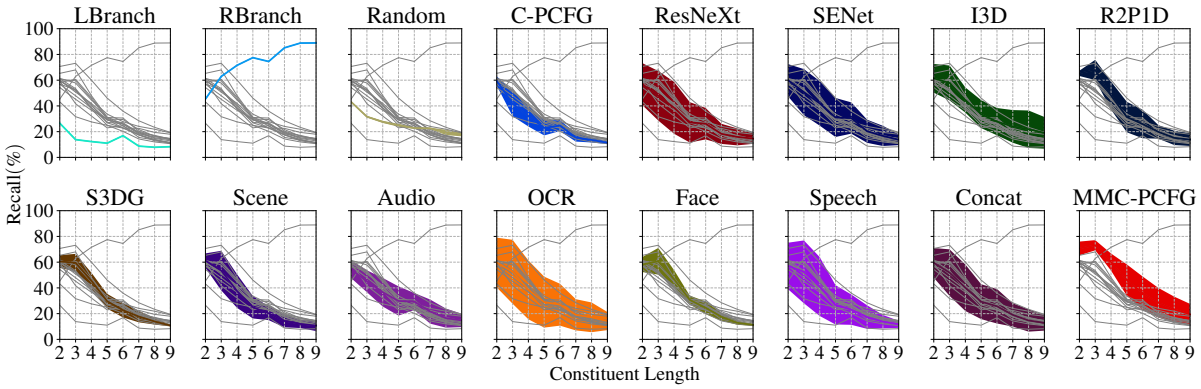

Figure 2: Recall comparison over constituent length on DiDeMo. Methods are differentiated with colors. For easy comparison, we additional draw gray lines in each figure to indicate the average recall shown by other figures.

with a single expert, we find that object features (ResNeXt and SENet) achieve top-2 recalls on NPs, while action features (I3D, R2P1D and S3DG) achieve the top-3 recalls on VPs and PPs. It indicates that different experts help parser learn syntactic structures from different aspects. Meanwhile, action features improve C-PCFG[3] on VPs and PPs by a large margin, which once again verifies the benefits of using video information.

Comparing our MMC-PCFG with VC-PCFG, our model achieves the top-2 recall and is smaller in variance in NP, VP and PP. It demonstrates that our model can take the advantages of different experts and learn consistent grammar induction.

[3]The low performance of C-PCFG on DiDeMo in terms of VP recall may be caused by it attaching a high attaching PP to the rest of the sentence instead of the rest of the verb phrase, which breaks the whole VP. For PPs, C-PCFG attaches prepositions to the word in front, which may be caused by confusion between prepositions in PPs and phrasal verbs.

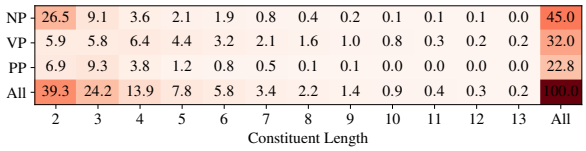

Figure 3: Label distributions over the constituent length on DiDeMo. *All* represent frequencies of constituent lengths.

4.6 Ablation Study

In this section, we conduct several ablation studies on DiDeMo, shown in Figures 2–4. All results are reported in percentage (%).

Performance comparison over constituent length. We first demonstrate the model performance for constituents at different lengths in Figure 2. As constituent length becomes longer, the recall of all models (except RBranch) decreases as expected (Kim et al., 2019; Zhao and Titov,

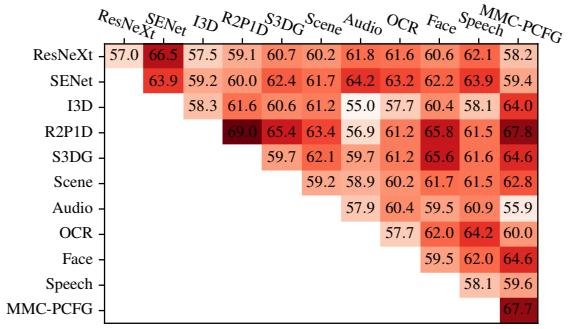

Figure 4: Consistency scores for different models on DiDeMo.

2020). MMC-PCFG outperforms C-PCFG and VC-PCFG under all constituent lengths. We further illustrate the label distribution over constituent length in Figure 3. We find that approximately 98.1% of the constituents have fewer than 9 words and most of them are NPs, VPs and PPs. This suggests that the improvement on NPs, VPs and PPs can strongly affect the overall performance.

Consistency between different models. Next, we analyze the consistency of these different models. The consistency between two models is measured by averaging sentence-level F1 scores over all possible pairings of different runs[4] (Williams et al., 2018). We plot the consistency for each pair of models in Figure 4 and call it consistency matrix. Comparing the self F1 of all the models (the diagonal in the matrix), R2P1D has the highest score, suggesting that R2P1D is the most reliable feature that can help parser to converge to a specific grammar. Comparing the models trained with different single experts, ResNeXt *v.s.* SENet reaches the highest non-self F1, since they are both object features trained on ImageNet and have similar effects to the parser. We also find that the lowest non-self F1 comes from Audio *v.s.* I3D, since they are extracted from different modalities (video *v.s.* sound). Compared with other models, our model is most consistent with R2P1D, indicating that R2P1D contributes most to our final prediction.

Contribution of different modalities. We also evaluate how different modalities contribute to the performance of MMC-PCFG. We divide current experts into three groups, video (objects, action, scene and face), audio (audio) and text (OCR and ASR). By ablating one group during training, we find that the model without video experts has the largest performance drops (see Table 2). Therefore, videos contribute most to the performance among all modalities.

[4]Different runs represent models trained with different seeds.

Model	NP	VP	PP	C-F1	S-F1
full	$68.0_{\pm 9.9}$	$52.7_{\pm 9.0}$	$63.8_{\pm 8.7}$	$55.3_{\pm 3.3}$	$59.0_{\pm 3.4}$
w/o audio	$69.3_{\pm 8.0}$	$41.8_{\pm 11.0}$	$45.3_{\pm 20.2}$	$48.7_{\pm 6.2}$	$52.0_{\pm 6.5}$
w/o text	$68.5_{\pm 13.7}$	$38.8_{\pm 16.9}$	$57.0_{\pm 20.4}$	$49.6_{\pm 10.4}$	$52.0_{\pm 11.1}$
w/o video	$64.3_{\pm 4.4}$	$28.1_{\pm 7.5}$	$38.9_{\pm 25.6}$	$41.4_{\pm 6.0}$	$44.8_{\pm 5.9}$

Table 2: Performance comparison over modalities on MMC-PCFG on DiDeMo.

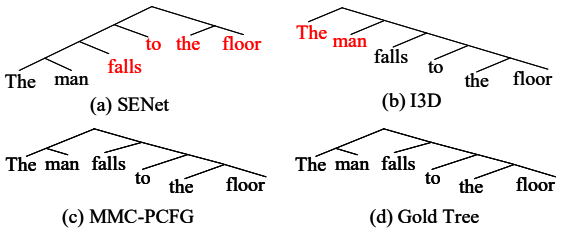

Figure 5: Parse trees predicted by different models for the sentence *The man falls to the floor*.

4.7 Qualitative Analysis

In Figure 5, we visualize a parse tree predicted by the best run of SENet154, I3D and MMC-PCFG. We can observe that SENet identifies all NPs but fails at the VP. I3D correctly predicts the VP but fails at recognizing a NP, "the man". Our MMC-PCFG can take advantages of all experts and produce the correct prediction.

5 Related Work

Grammar Induction Grammar induction and unsupervised parsing has been a long-standing problem in computational linguistics (Carroll and Charniak, 1992). Recent work utilized neural networks in predicting constituency structures with no supervision (Shen et al., 2018a; Drozdov et al., 2019; Shen et al., 2018b; Kim et al., 2019; Jin et al., 2019a) and showed promising results. In addition to learning purely from text, there is a growing interest to use image information to improve accuracy of induced constituency trees (Shi et al., 2019; Kojima et al., 2020; Zhao and Titov, 2020; Jin and Schuler, 2020). Different from previous work, our work improves the constituency parser by using videos containing richer information than images.

Video-Text Matching Video-text matching has been widely studied in various tasks, such as video retrieval (Liu et al., 2019; Gabeur et al., 2020), moment localization with natural language (Zhang

et al., 2019, 2020) and video question and answering (Xu et al., 2017; Jin et al., 2019b). It aims to learn video-semantic representation in a joint embedding space. Recent works (Liu et al., 2019; Gabeur et al., 2020; Chen et al., 2020) focus on learning video's multi-modal representation to match with text. In this work, we borrow this idea to match video and textual representations.

6 Conclusion

In this work, we have presented a new task referred to as video-aided unsupervised grammar induction. This task aims to improve grammar induction models by using aligned video-sentence pairs as an effective way to address the limitation of current image-based methods where only object information from static images is considered and important verb related information from vision is missing. Moreover, we present Multi-Modal Compound Probabilistic Context-Free Grammars (MMC-PCFG) to effectively integrate video features extracted from different modalities to induce more accurate grammars. Experiments on three datasets demonstrate the effectiveness of our method.

7 Acknowledgement

We thank the support of NSF awards IIS-1704337, IIS-1722847, IIS-1813709, and the generous gift from our corporate sponsors.

References

Samuel Albanie, Yang Liu, Arsha Nagrani, Antoine Miech, Ernesto Coto, Ivan Laptev, Rahul Sukthankar, Bernard Ghanem, Andrew Zisserman, Valentin Gabeur, et al. 2020. The end-of-end-to-end: A video understanding pentathlon challenge (2020). *arXiv preprint arXiv:2008.00744*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Glenn Carroll and Eugene Charniak. 1992. *Two experiments on learning probabilistic dependency grammars from corpora*. Department of Computer Science, Univ.

Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. 2020. Learning modality interaction for temporal sentence localization and event captioning in videos. In *ECCV*.

John Cocke. 1969. *Programming languages and their compilers: Preliminary notes*. New York University.

Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. 2018. Pixellink: Detecting scene text via instance segmentation. In *AAAI*.

Andrew Drozdov, Patrick Verga, Mohit Yadav, Mohit Iyyer, and Andrew McCallum. 2019. Unsupervised latent tree induction with deep inside-outside recursive auto-encoders. In *NAACL*.

Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-domain generalization of neural constituency parsers. In *ACL*.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *ECCV*.

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. 2019. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*.

Lila Gleitman. 1990. The Structural Sources of Verb Meanings. *Language Acquisition*, 1(1):3–55.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *ECCV*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *ICASSP*.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *CVPR*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018. Unsupervised grammar induction with depth-bounded pcfg. *TACL*.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, Lane Schwartz, and William Schuler. 2019a. Unsupervised learning of PCFGs with normalizing flow. In *ACL*.

Lifeng Jin and William Schuler. 2020. Grounded pcfg induction with images. In *AACL-IJCNLP*.

Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. 2019b. Multi-interaction network with object relation for video question answering. In *ACM Multimedia*.

Tadao Kasami. 1966. An efficient recognition and syntax-analysis algorithm for context-free languages. *Coordinated Science Laboratory Report no. R-257*.

Yoon Kim, Chris Dyer, and Alexander M Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *ACL*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *ACL*.

Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M Rush, and Yoav Artzi. 2020. What is learned in visually grounded neural syntax acquisition. In *ACL*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. SSD: Single shot multibox detector. In *ECCV*.

Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. In *arXiv preprint arxiv:1907.13487*.

Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. 2018. Synthetically supervised feature learning for scene text recognition. In *ECCV*.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *ECCV*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*.

Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Steven Pinker and B MacWhinney. 1987. The bootstrapping problem in language acquisition. *Mechanisms of language acquisition*, pages 399–441.

Yikang Shen, Zhouhan Lin, Chin wei Huang, and Aaron Courville. 2018a. Neural language modeling by jointly learning syntax and lexicon. In *ICLR*.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2018b. Ordered neurons: Integrating tree structures into recurrent neural networks. In *ICLR*.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *ACL*.

Michael Tomasello. 2003. *Constructing a language: A usage-based theory of language acquisition.* Harvard University Press, Cambridge, MA, US.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Adina Williams, Andrew Drozdov*, and Samuel R Bowman. 2018. Do latent tree learning models identify meaningful structure in sentences? *TACL*.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.

Daniel H Younger. 1967. Recognition and parsing of context-free languages in time n3. *Information and control*, 10(2):189–208.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks formoment localization with natural language. In *AAAI*.

Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019. Exploiting temporal relationships in video moment localization with natural language. In *ACM Multimedia*.

Yanpeng Zhao and Ivan Titov. 2020. Visually grounded compound PCFGs. In *EMNLP*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *TPAMI*.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI*.

A Performance Comparison - Full Tables

	Method	NP	VP	PP	SBAR	ADJP	ADVP	C-F1	S-F1
	LBranch	41.7	0.1	0.1	0.7	7.2	0.0	16.2	18.5
	RBranch	32.8	**91.5**	**66.5**	**88.2**	**36.9**	**63.6**	*53.6*	*57.5*
	Random	$36.5_{\pm 0.6}$	$30.5_{\pm 0.5}$	$30.1_{\pm 0.5}$	$25.7_{\pm 2.8}$	$29.5_{\pm 2.3}$	$28.5_{\pm 4.8}$	$29.4_{\pm 0.3}$	$32.7_{\pm 0.5}$
	C-PCFG	$\mathbf{72.9}_{\pm 5.5}$	$16.5_{\pm 6.2}$	$23.4_{\pm 16.9}$	$26.6_{\pm 15.9}$	$25.0_{\pm 11.6}$	$14.7_{\pm 12.8}$	$38.2_{\pm 5.0}$	$40.4_{\pm 4.1}$
VC-PCFG	ResNeXt	$64.4_{\pm 21.4}$	$25.7_{\pm 17.7}$	$34.6_{\pm 25.0}$	$40.5_{\pm 26.3}$	$16.7_{\pm 9.5}$	$28.4_{\pm 21.3}$	$40.0_{\pm 13.7}$	$41.8_{\pm 14.0}$
VC-PCFG	SENet	$70.5_{\pm 15.3}$	$25.7_{\pm 15.9}$	$36.5_{\pm 24.6}$	$36.8_{\pm 25.9}$	$21.2_{\pm 12.5}$	$23.6_{\pm 16.8}$	$42.6_{\pm 10.4}$	$44.0_{\pm 10.4}$
VC-PCFG	I3D	$57.9_{\pm 13.5}$	$45.7_{\pm 14.1}$	$45.8_{\pm 17.2}$	$38.2_{\pm 14.8}$	$28.4_{\pm 9.2}$	$22.0_{\pm 9.3}$	$45.1_{\pm 6.0}$	$49.2_{\pm 6.0}$
VC-PCFG	R2P1D	$61.2_{\pm 8.5}$	$38.1_{\pm 5.4}$	$62.1_{\pm 4.1}$	$61.5_{\pm 5.1}$	$21.4_{\pm 11.4}$	$40.8_{\pm 7.3}$	$48.1_{\pm 4.4}$	$50.7_{\pm 4.2}$
VC-PCFG	S3DG	$61.3_{\pm 13.4}$	$31.7_{\pm 16.7}$	$51.8_{\pm 8.0}$	$50.3_{\pm 6.5}$	$18.0_{\pm 4.5}$	$35.2_{\pm 11.4}$	$44.0_{\pm 2.7}$	$46.5_{\pm 5.1}$
VC-PCFG	Scene	$62.2_{\pm 9.6}$	$30.6_{\pm 12.3}$	$41.1_{\pm 24.8}$	$35.2_{\pm 21.9}$	$21.4_{\pm 14.0}$	$27.6_{\pm 17.1}$	$41.7_{\pm 6.5}$	$44.9_{\pm 7.4}$
VC-PCFG	Audio	$64.2_{\pm 18.6}$	$21.3_{\pm 26.5}$	$34.7_{\pm 11.0}$	$37.3_{\pm 19.6}$	$26.1_{\pm 4.9}$	$18.2_{\pm 11.6}$	$38.7_{\pm 3.7}$	$39.5_{\pm 5.2}$
VC-PCFG	OCR	$64.4_{\pm 15.0}$	$27.4_{\pm 19.5}$	$42.8_{\pm 31.2}$	$35.9_{\pm 20.7}$	$14.6_{\pm 1.7}$	$23.2_{\pm 24.0}$	$41.9_{\pm 16.9}$	$44.6_{\pm 17.5}$
VC-PCFG	Face	$60.8_{\pm 16.0}$	$31.5_{\pm 17.0}$	$52.8_{\pm 9.8}$	$49.3_{\pm 5.6}$	$12.6_{\pm 3.3}$	$32.9_{\pm 14.6}$	$43.9_{\pm 4.5}$	$46.3_{\pm 5.5}$
VC-PCFG	Speech	$61.8_{\pm 12.8}$	$26.6_{\pm 17.6}$	$43.8_{\pm 34.5}$	$34.2_{\pm 20.6}$	$14.4_{\pm 4.8}$	$12.9_{\pm 9.6}$	$40.9_{\pm 16.0}$	$43.1_{\pm 16.1}$
VC-PCFG	Concat	$68.6_{\pm 8.6}$	$24.9_{\pm 19.9}$	$39.7_{\pm 19.5}$	$39.3_{\pm 19.8}$	$10.8_{\pm 2.8}$	$18.3_{\pm 18.1}$	$42.2_{\pm 12.3}$	$43.2_{\pm 14.2}$
	MMC-PCFG	$\mathit{67.9}_{\pm 9.8}$	$\mathit{52.3}_{\pm 9.0}$	$\mathit{63.5}_{\pm 8.6}$	$\mathit{60.7}_{\pm 10.8}$	$\mathit{34.7}_{\pm 17.0}$	$\mathit{50.4}_{\pm 8.3}$	$\mathbf{55.0}_{\pm 3.7}$	$\mathbf{58.9}_{\pm 3.4}$

Table 3: Performance Comparison on DiDeMo.

	Method	NP	VP	PP	SBAR	ADJP	ADVP	C-F1	S-F1
	LBranch	1.7	42.8	0.4	8.1	1.5	0.0	6.8	5.9
	RBranch	35.6	*47.5*	67.0	**88.9**	**33.9**	**65.0**	35.0	41.6
	Random	$27.2_{\pm 0.3}$	$27.1_{\pm 1.4}$	$29.9_{\pm 0.5}$	$31.3_{\pm 5.2}$	$26.9_{\pm 7.7}$	$26.2_{\pm 11.9}$	$21.2_{\pm 0.2}$	$24.0_{\pm 0.2}$
	C-PCFG	$47.4_{\pm 18.4}$	$\mathbf{49.4}_{\pm 11.9}$	$58.0_{\pm 22.6}$	$45.7_{\pm 6.0}$	$\mathit{27.7}_{\pm 15.1}$	$36.2_{\pm 7.4}$	$37.8_{\pm 6.7}$	$41.4_{\pm 6.6}$
VC-PCFG	ResNeXt	$46.5_{\pm 13.7}$	$40.8_{\pm 9.8}$	$67.9_{\pm 12.7}$	$50.5_{\pm 13.3}$	$22.3_{\pm 6.7}$	$38.8_{\pm 21.3}$	$38.2_{\pm 8.3}$	$42.8_{\pm 8.4}$
VC-PCFG	SENet	$48.3_{\pm 14.4}$	$40.7_{\pm 9.2}$	$73.6_{\pm 11.2}$	$45.5_{\pm 17.0}$	$26.9_{\pm 13.6}$	$41.2_{\pm 17.5}$	$39.9_{\pm 8.7}$	$44.9_{\pm 8.3}$
VC-PCFG	I3D	$48.1_{\pm 10.7}$	$39.0_{\pm 8.0}$	$\mathit{79.4}_{\pm 8.4}$	$50.0_{\pm 14.9}$	$18.5_{\pm 7.0}$	$41.2_{\pm 4.1}$	$40.6_{\pm 3.6}$	$45.7_{\pm 3.2}$
VC-PCFG	R2P1D	$\mathit{52.4}_{\pm 10.9}$	$33.7_{\pm 16.4}$	$66.7_{\pm 10.7}$	$49.5_{\pm 13.8}$	$25.8_{\pm 10.6}$	$33.8_{\pm 12.4}$	$39.4_{\pm 8.1}$	$44.4_{\pm 8.3}$
VC-PCFG	S3DG	$50.4_{\pm 13.1}$	$32.6_{\pm 16.3}$	$71.7_{\pm 7.5}$	$33.3_{\pm 5.9}$	$30.8_{\pm 17.5}$	$40.0_{\pm 7.1}$	$39.3_{\pm 6.5}$	$44.1_{\pm 6.6}$
VC-PCFG	Audio	$51.2_{\pm 3.1}$	$42.0_{\pm 7.2}$	$61.5_{\pm 18.0}$	$\mathit{51.0}_{\pm 14.8}$	$23.5_{\pm 16.8}$	$48.8_{\pm 8.2}$	$39.2_{\pm 4.7}$	$43.3_{\pm 4.9}$
VC-PCFG	OCR	$48.6_{\pm 8.1}$	$41.5_{\pm 4.1}$	$65.5_{\pm 17.4}$	$39.9_{\pm 4.4}$	$18.5_{\pm 6.6}$	$\mathit{53.8}_{\pm 14.7}$	$38.6_{\pm 5.5}$	$43.2_{\pm 5.6}$
VC-PCFG	Concat	$50.3_{\pm 10.3}$	$42.3_{\pm 2.9}$	$\mathbf{81.6}_{\pm 8.7}$	$40.1_{\pm 3.9}$	$17.7_{\pm 8.2}$	$52.5_{\pm 5.6}$	$\mathit{42.3}_{\pm 5.7}$	$\mathit{47.0}_{\pm 5.6}$
	MMC-PCFG	$\mathbf{62.7}_{\pm 9.8}$	$45.3_{\pm 2.8}$	$63.4_{\pm 17.7}$	$43.9_{\pm 4.8}$	$26.2_{\pm 7.5}$	$35.0_{\pm 3.5}$	$\mathbf{44.7}_{\pm 5.2}$	$\mathbf{48.9}_{\pm 5.7}$

Table 4: Performance Comparison on YouCook2.

	Method	NP	VP	PP	SBAR	ADJP	ADVP	C-F1	S-F1
	LBranch	34.6	0.1	0.9	0.2	3.8	0.3	14.4	16.8
	RBranch	34.6	**90.9**	67.5	**94.8**	25.4	54.8	54.2	58.6
	Random	$34.6_{\pm 0.1}$	$26.8_{\pm 0.1}$	$28.1_{\pm 0.2}$	$24.6_{\pm 0.3}$	$24.8_{\pm 1.0}$	$28.1_{\pm 1.4}$	$27.2_{\pm 0.1}$	$30.5_{\pm 0.1}$
	C-PCFG	$46.6_{\pm 3.2}$	$61.1_{\pm 3.3}$	$72.5_{\pm 8.3}$	$63.7_{\pm 4.0}$	$33.1_{\pm 7.1}$	$67.1_{\pm 4.7}$	$50.7_{\pm 3.2}$	$55.0_{\pm 3.2}$
VC-PCFG	ResNeXt	$48.6_{\pm 3.0}$	$59.0_{\pm 6.0}$	$72.0_{\pm 3.6}$	$62.1_{\pm 5.2}$	$32.6_{\pm 2.5}$	$70.4_{\pm 6.4}$	$50.7_{\pm 1.7}$	$54.9_{\pm 2.2}$
VC-PCFG	SENet	$49.0_{\pm 4.4}$	$63.5_{\pm 6.4}$	$71.7_{\pm 4.8}$	$60.9_{\pm 10.6}$	$34.0_{\pm 6.4}$	$\mathit{74.1}_{\pm 1.9}$	$52.2_{\pm 1.2}$	$56.0_{\pm 1.6}$
VC-PCFG	I3D	$\mathbf{53.9}_{\pm 10.5}$	$63.2_{\pm 9.1}$	$73.7_{\pm 2.9}$	$65.3_{\pm 9.1}$	$\mathbf{35.0}_{\pm 6.8}$	$73.8_{\pm 4.1}$	$54.5_{\pm 1.6}$	$\mathit{59.1}_{\pm 1.7}$
VC-PCFG	R2P1D	$52.8_{\pm 3.6}$	$63.3_{\pm 4.6}$	$73.1_{\pm 10.1}$	$66.9_{\pm 2.0}$	$34.0_{\pm 2.2}$	$72.5_{\pm 4.2}$	$54.0_{\pm 2.5}$	$58.0_{\pm 2.3}$
VC-PCFG	S3DG	$48.2_{\pm 4.4}$	$60.4_{\pm 3.9}$	$71.4_{\pm 6.4}$	$58.1_{\pm 8.2}$	$25.3_{\pm 2.2}$	$61.8_{\pm 8.4}$	$50.7_{\pm 3.2}$	$54.7_{\pm 2.9}$
VC-PCFG	Scene	$50.7_{\pm 1.6}$	$65.0_{\pm 4.7}$	$\mathbf{78.6}_{\pm 3.6}$	$\mathit{67.3}_{\pm 3.9}$	$\mathit{34.5}_{\pm 4.6}$	$71.7_{\pm 1.8}$	$\mathit{54.6}_{\pm 1.5}$	$58.4_{\pm 1.3}$
VC-PCFG	Audio	$50.0_{\pm 1.1}$	$63.7_{\pm 6.1}$	$72.7_{\pm 3.0}$	$61.9_{\pm 6.5}$	$\mathit{34.5}_{\pm 2.3}$	$68.0_{\pm 5.9}$	$52.8_{\pm 1.3}$	$56.7_{\pm 1.4}$
VC-PCFG	OCR	$48.3_{\pm 8.3}$	$57.1_{\pm 4.6}$	$76.9_{\pm 0.6}$	$60.7_{\pm 4.9}$	$33.9_{\pm 8.3}$	$72.1_{\pm 4.4}$	$51.0_{\pm 3.0}$	$55.5_{\pm 3.0}$
VC-PCFG	Face	$46.5_{\pm 6.8}$	$61.3_{\pm 3.6}$	$71.5_{\pm 7.1}$	$60.8_{\pm 11.0}$	$30.9_{\pm 3.4}$	$68.4_{\pm 6.0}$	$50.5_{\pm 2.6}$	$54.5_{\pm 2.6}$
VC-PCFG	Speech	$48.5_{\pm 7.6}$	$60.7_{\pm 3.5}$	$74.5_{\pm 5.7}$	$62.6_{\pm 6.2}$	$27.3_{\pm 1.8}$	$74.0_{\pm 3.1}$	$51.7_{\pm 2.6}$	$56.2_{\pm 2.5}$
VC-PCFG	Concat	$43.6_{\pm 6.0}$	$64.7_{\pm 3.0}$	$68.5_{\pm 8.0}$	$63.8_{\pm 3.8}$	$32.0_{\pm 5.5}$	$70.4_{\pm 5.9}$	$49.8_{\pm 4.1}$	$54.2_{\pm 4.0}$
	MMC-PCFG	$\mathit{52.3}_{\pm 5.1}$	$\mathit{68.1}_{\pm 2.9}$	$\mathit{78.2}_{\pm 1.9}$	$65.8_{\pm 2.4}$	$32.0_{\pm 2.0}$	$\mathbf{74.7}_{\pm 2.3}$	$\mathbf{56.0}_{\pm 1.4}$	$\mathbf{60.0}_{\pm 1.2}$

Table 5: Performance Comparison on MSRVTT.