MULTIMEDIA CURRICULUM LEARNING FOR LANGUAGE ACQUISITION

Pengfei Yu*, **Heng Ji***, **Shih-fu Chang**†, **Kevin Duh**†,
* University of Illinois Urbana Champaign, United States
† Columbia University, United States
† Johns Hopkins University, United States

Abstract—We explore how curriculum learning impacts language acquisition from multimedia data. We propose a new curriculum learning methods based on word concreteness aiming to strengthen the learning of concrete concepts in images. We construct a new Yoga Videos dataset to evaluate language acquisition and experiment with MS COCO [1] image captioning dataset to show the generalizability of our approaches. Extensive experimental results demonstrate the effectiveness of language curriculum and multimedia learning to accelerate learning and improve data efficiency by achieving the equivalent performance with approximately 40% less training data, especially with small-scale datasets.

Keywords—Language acquisition, Curriculum learning, Multimedia learning.

I. INTRODUCTION

Humans can efficiently acquire their first languages even as children. We note two essential features in human language acquisition (1) exposure to multimedia information including visual, vocal, and textual formats (2) learning from easier materials to more difficult ones. Inspired by these observations, we believe it is important to study language acquisition in the context of multimedia data such as incidentally synchronized video-text pairs in narrated videos or semantically coherent image-caption pairs. We further introduce Curriculum Learning (CL) [2] to language acquisition since the incremental development from easier to more complex concepts coincides with the fundamental idea of CL.

CL trains models with several stages. CL adopts sampling distributions that favor certain instances considered as "easier".[1] in early stages, and gradually smooths sampling distributions to the uniform distribution to take full advantage of the whole training data. In language acquisition, we favor more informative and well-aligned vision-text pairs instead of noisy and overly verbose pairs. We hypothesize that in this way, models can fast learn language ability from less noisy instances and also avoid overfitting to dataset biases, especially when trained without ample training instances. We considered multimedia curriculum learning in two granularities: coarse-grained sentence-level and fine-grained word level. We show the intuition for both granularities in Figure 1, where the model is trained on an easy subset in stage 1, and then on the whole dataset in stage 2. We show the intuition for both granularities in Figure 1, where the model is trained on an easy subset in stage 1, and then on the whole dataset in stage 2.

We explore two curriculum learning methods derived from data without human guidance. The first method is a word-level method based on word concreteness scores, emphasizing concrete concepts such as body parts in Yoga Videos. These words are more closely related to the associated visual scenes and crucial for language acquisition. For comparison, we experiment with another curriculum adapted from transfer-based methods by [3], where we measure vision-text pair "difficulties" according to the corresponding losses from a pretrained captioning model. Extensive experimental results show that both curriculum learning approaches benefit learning for both captioning and visual retrieval tasks. Moreover, we demonstrate that our proposed concreteness-based curriculum brings more consistent improvements compared with transfer-based curriculums.

To summarize, the main contributions of this paper are:

- We explore curriculum learning methods in language acquisition from multimedia data on both sentence-level and word-level and proposed a new concreteness-based word-level curriculum based on the intuition that concrete words are easier to learn because they are better aligned with visual scenes.
- We propose evaluating language acquisition in two aspects: captioning and visual retrieval, and

Contact author: Pengfei Yu,
Email: pengfei4@illinois.edu
Manuscript received: 12/6/2022, revised: 08/7/2022, accepted: 02/8/2022.

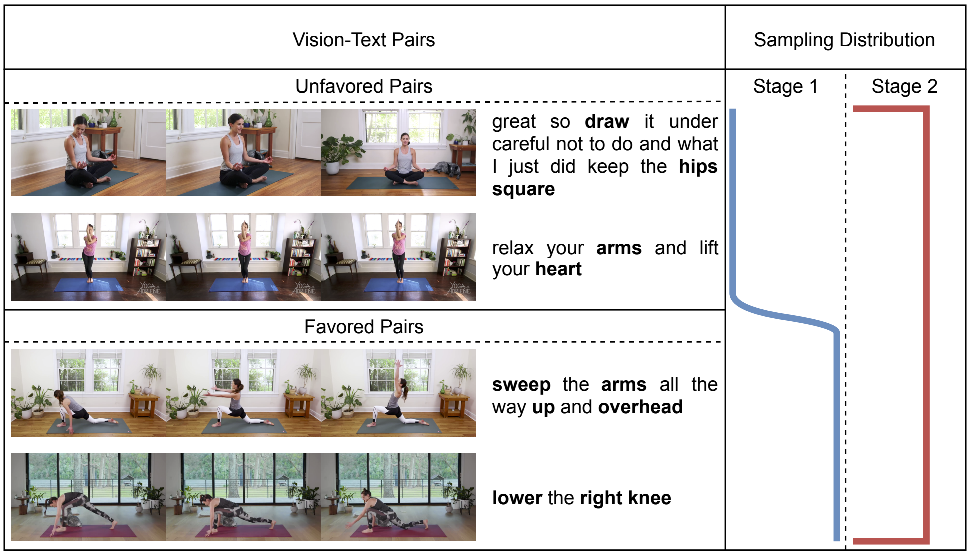


Fig. 1: An example of curriculum learning process. Well-aligned video-sentence pairs are sampled with higher probability in stage 1. Besides, concrete words that are bold in sentences represent central semantics in sentences and it is beneficial to emphasize such vision-word pairs.

collect Yoga Videos as a new demonstrative video-text dataset for language acquisition.
- We show that multimedia features are beneficial in language acquisition to achieve more reliable semantic representation.

II. Approach

A. Curriculum Learning

We give a general introduction to CL in this section. We denote $\mathcal{D} = \{(x_i, y_i)|x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ as the training dataset of data-label pairs, $F \in \mathcal{F} : \mathcal{X} \mapsto \mathcal{Y}$ as the model to be trained, and $L : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ as the loss function. For instance, in captioning, let $\mathcal{W}$ denote the vocabulary set. $\mathcal{X}$ is a set of visual scene inputs (videos or images), $\mathcal{Y} = \biguplus_{n=1}^{\infty} \mathcal{W}^n$ is a set of sentences, and $\mathcal{F}$ is a collection of models that can produce a posterior distribution $P(\text{text}|\text{scene})$ given scenes.

For vanilla Stochastic Gradient Descent (SGD) training on $\mathcal{D}$, as well as optimization methods [4], [5], [6], [7] derived from it, we sample the dataset from uniform distribution $\mathcal{U}(\mathcal{D})$ over $\mathcal{D}$ for each step i. The model F is trained to minimize the loss

$$F^* = \arg\min_{\mathcal{F}} \mathbb{E}_{d \sim \mathcal{U}(\mathcal{D})} L(F, x_i, y_i). \tag{1}$$

On the contrary, curriculum learning uses a sequence of evolving sampling distributions $\varphi_i \in \mathcal{P}(\mathcal{D})$ at each step i. In common practice we adopt $\varphi_i \to \mathcal{U}(\mathcal{D})$ to take full advantage of the training data.

Let $Q : \mathcal{D} \mapsto \mathbb{R}$ measure the difficulty of instances. Since CL utilizes a training procedure that favors easy instances in the early stage, we expect $\mathbb{E}_{(x_i,y_i)\sim\varphi_i} Q(x_i, y_i)$ to be monotonically non-decreasing with respect to i.

B. Basic Settings

We consider two important aspects of human language acquisition in a multimedia setting: describing a new scene and illustrating a descriptive text. We simplify these two tasks as modeling two posterior: $P(\text{text}|\text{scene})$ (captioning) and $P(\text{scene}|\text{text})$ (visual retrieval). We introduce how we model $P(\text{text}|\text{scene})$ for captioning and $P(\text{scene}|\text{text})$ for visual retrieval in this section. We denote each multimedia data sample $d = (v, w_{1:n})$ as a pair of visual input v and textual instruction $w_{1:n}$, where $w_i \in \mathcal{W}$ are words in the vocabulary.

Captioning: For the image/video caption generation problem, we apply a basic encoder-decoder framework to demonstrate the impact of curriculum learning. Each input visual data sample v is first encoded into a hidden representation,

$$\mathbf{v} = \text{VisualEncoder}(v).$$

Then we feed $\mathbf{v}$ as the initial hidden state of a LSTM [8] decoder and decode the hidden representation auto-regressively,

$$P(w_i|v, w_{1:i-1}; \theta) = \text{LSTMDecoder}(\mathbf{v}, w_{1:i-1}).$$

The final output from LSTM decoder is a conditional distribution over all sentences given v, and we summarize the captioning model F as $F(v) \in$

$\mathcal{P}(\cup_{n=1}^{\infty}\mathcal{W}^n)$. For sentence-level curriculum training, we minimize the objective function below for step i,

$$\begin{aligned} &\mathop{\mathbb{E}}_{(v,w_{1:n})\sim\varphi_i} L(F,v,w_{1:n}) \\ =&\mathop{\mathbb{E}}_{(v,w_{1:n})\sim\varphi_i} -\log P_F(w_{1:n}|v) \\ =&\mathop{\mathbb{E}}_{(v,w_{1:n})\sim\varphi_i} -\sum_{j=1}^{n}\log P_F(w_j|v,w_{1:j-1}) \end{aligned} \quad (2)$$

where P_F is the probability density given by $F(v)$. For word-level, we transform sampling vision-word pairs (v, w_j) into weighting over token-level loss

$$\mathop{\mathbb{E}}_{\substack{(v,w_{1:n})\sim\mathcal{U}(D)\\ 1\le j\le n}} \varphi_i(v,w_j)\log P_F(w_j|v,w_{1:j-1}). \quad (3)$$

For vanilla training, $\varphi_i = \mathcal{U}(\mathcal{D})$, and losses in Equation (2) and Equation (3) become equivalent. We use beam search with beam size 10 for decoding.

Visual Retrieval: We model visual retrieval by Bayesian inference

$$P(\text{scene}|\text{text}) \propto P(\text{text}|\text{scene})P(\text{scene}),$$

where $P(\text{scene})$ is roughly estimated by fitting an Gaussian Mixture model on hidden representations $\mathbf{v}$ from captioning model. We leave better modeling of $P(\text{scene})$ for future work. Since our modeling of $P(\text{scene})$ could be sub-optimal, we use

$$\begin{aligned}\text{Score}(\text{scene}|\text{text}) = \log P(\text{text}|\text{scene}) \\ + \lambda\log P(\text{scene})\end{aligned}$$

for retrieval. We set $\lambda \in [0, 2]$ and tune this parameter on validation set.

C. Language Curriculum

To set up a curriculum, we need to select a difficulty function Q to stress easier instances that can benefits training, and a sampling strategy $\{\varphi_i|i = 1, 2, \ldots, \infty\}$ based on Q.

Difficulty Measures: The selection of Q should underscore informative vision-text pairs as illustrated in 1. We explore two measures that can be derived directly from data without human guidance.

Transfer-based Metric. A pretrained captioning model $\hat{F}$ capturing the connection between visual scenes and text semantics should render higher probabilities (or lower losses) for more informative vision-text pairs. For vision-sentence pairs, we define $Q(w_{1:n}, v) = -\log P_{\hat{F}}(w_{1:n}|v)$; for vision-word pairs, we define $Q(w_i, v) = -\log P_{\hat{F}}(w_i|v, w_{1:i-1})$.

Word Concreteness. We assume concrete words are naturally better aligned with visual scenes and compose more informative vision-word pairs. We follow [9] to learn concreteness scores of words from multimedia datasets. We evaluate a word's concreteness by assessing how close its associated visual scene representations $\mathbf{v}$ are to each other. We include details in Appendix. Then $Q(w_i, v)$ is defined as concreteness score of w_i. Word concreteness is also an important concept in linguistics. For additional comparison, we also experiment with a manually constructed word concreteness database [10], which includes concreteness scores for most common words.

Sampling Strategy: Given a specific difficulty measure Q, sampling strategies should favor easier vision-text pairs with smaller Q in early stages. We adopt a simple but effective two-stage curriculum learning strategy [3] based on a single step hyper-parameter N and a difficulty threshold q_0, where

$$\varphi_i = \begin{cases} \mathcal{U}(\mathcal{D}^E) & i \le N \\ \mathcal{U}(\mathcal{D}) & i > N \end{cases},$$

and $\mathcal{D}^E = \{(x_i, y_i)|Q(x_i, y_i) < q_0\}$. Our choice of N is elaborated in Appendix. We choose q_0 to balance the samples such that $|\mathcal{D}^E| = |\mathcal{D}\backslash\mathcal{D}^E|$. Although this method can be extended to multi-step sampling with more steps and threshold parameters, a two-step method is enough to show the effect of curriculum learning since our training dataset is relatively small in scale.

Although collecting an easy subset with hard selection using difficulty scores is common in curriculum learning, we empirically find that for transfer-based methods, switching hard selection to soft selection can be sometimes more beneficial. To be concrete, with pretrained model $\hat{F}$ and corresponding posterior distribution $P_{\hat{F}}$, we define the soft sampling strategy for vision-sentence pairs as

$$\varphi_i = \begin{cases} \propto P_{\hat{F}}(w_{1:n}|v) & i \le N \\ \mathcal{U}(\mathcal{D}) & i > N \end{cases}.$$

and for vision-word pairs as

$$\varphi_i = \begin{cases} \propto P_{\hat{F}}(w_i|v, w_{1:i-1}) & i \le N \\ \mathcal{U}(\mathcal{D}) & i > N \end{cases}.$$

In a sense, the hard sampling strategy is a rectified approximation of the soft sampling.

III. Experiments

A. Dataset and Experiment Setting

We include preprocessing and hyperparameters in Appendix. We experiment with two datasets.

a) Yoga Videos.: We collect the Yoga Videos dataset as a case study from yoga instructional videos, which have realistic yet simple visual scenes of instructors performing yoga with a static background. Instructions are mostly in synchronization with the action of moving body parts in videos. We collect $18,705$ short videos of yoga actions, clipped from 297 yoga instructional videos from YouTube. The average duration of short videos is 3.1s. We use the synchronized transcripts as captions and tokenize them using spaCy [11]. These captions are usually short, informative, and can be easily grounded into some visual scene (for example `straighten arm`, `lift leg`), which is ideal for language acquisition study. We keep only the lemmatized form of each alphabetic token in the dataset, resulting in a vocabulary of 763 words. The average caption length in the dataset is 9.6. We randomly split the dataset into $14,127$ training video-text pairs, $2,246$ validation pairs and $2,332$ test pairs.

b) MS COCO Captioning.: We use MS COCO to explore the generalization ability of curriculums and the effect of the curriculum with respect to dataset size. We follow the validation and test splits released by [12], and combine both `train` split with `restval` split in the original dataset as the full training split. In MS COCO, each image is paired with 5 captions on average. As a result, we have $113,287$ images and $414,113$ captions for training, $5,000$ images and $25,010$ captions for validation and $5,000$ images and $25,010$ captions for testing.

To study curriculum learning with varying data sizes, we randomly sample four expanding training subsets containing 4k, 8k, 16k, and 32k training images respectively. Note that the smallest subset contains around 20k vision-text pairs, which is close to Yoga Videos. We have another held-out subset of $81,287$ training images, which is disjoint with all previous 4 subsets.

c) Methods in Comparison: We compare transfer-based curriculum, concreteness-based curriculum methods to the vanilla model trained without curriculum. Since the concreteness is a novel metric, we add additional comparison for this metric, including reverse concreteness curriculum, random concreteness curriculum, and the linguistic concreteness metric.

Vanilla. The model trained without curriculum learning.

Loss. We use *loss* to denote transfer-based curriculum, since it is based on the losses of a pretrained model. We experiment with this method in both sentence-level and word-level as described in Section II-C. We also empirically study the soft selection methods, denoted as *soft loss*.

Concreteness and **Concreteness L.** Our proposed word-level curriculum using concreteness scores learned from multimedia data (**Concreteness**).We also explore using manually annotated scores by linguistics (**Concreteness L**).

Reverse Concreteness. We reverse the word ranking in concreteness curriculum.

Random Concreteness. We set random concreteness scores for each word. In our experiment we average over 3 random assignments and report the average performance.

B. Captioning and Visual Retrieval

We summarize the main results on Yoga Videos in Table I. We use Yoga Videos dataset to evaluate the impact of curriculum learning on two aspects of language acquisition. For captioning aspects, we present *BLEU-4* [13] scores (CIDEr and ROUGE-L share similar trends, so we leave those results in Appendix for simplicity) .

In addition to classic captioning scores, we evaluate *Verb Noun Recall* of generated captions, i.e., recall of action verbs and body part nouns (e.g. `lift leg`) in generated captions. We assume these words represent the central semantics of yoga instructions. We manually select 37 frequent body-part nouns from all the captions and run part-of-speech tagging and dependency parsing using spaCy [11] on test captions to obtain head verbs for these nouns on the dependency tree. In this way, we collect $1 \sim 2$ verb-noun pairs for each caption. We compute the average recall of these nouns, verbs, and verb-noun pairs for each caption.

For visual retrieval aspects, we have two evaluation metrics: *Hit@20*, as the ratio of the target video ranked top 20 among $2,332$ (1%) test videos. Hit@20 reflects how likely for models to successfully retrieve the target videos; *Mean Rank(MR)*, which is defined as

$$\mathrm{MR} = \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{(v, w_{1:n}) \in \mathcal{D}_{\mathrm{test}}} \mathrm{Rank}(v|w_{1:n}).$$

Here $\mathcal{D}_{\mathrm{test}}$ refers to the test dataset and $\mathrm{Rank}(v|w_{1:n})$ is the rank of target video v given text $w_{1:n}$. MR offers a more general view of retrieval results.

We further use MS COCO image captioning dataset to study the generalization ability of the proposed curriculum and the effects of the curriculum on varying data size. We also explore the impact of curriculum learning in terms of data efficiency. We show BLEU-4 in Table II and leave CIDEr and ROUGE-L in Appendix. For subsets with varying size we experiment with *Concreteness*, *Concreteness L* and *Soft Loss* curriculums. For *Concreteness* curriculum, word concreteness scores are always learned from the corresponding training subset. For *Soft Loss* curriculums, we experiment with three pretrained models to collect losses, pretrained on the corresponding training subsets (**Self**), on held-out subset with $81,287$ images (**Held**) and whole training data (**Whole**) respectively.

In general, curriculum learning methods improve performance over both captioning and visual retrieval settings across multiple datasets. We found that the pretrain posteriors are more beneficial for smaller datasets (see results on MS COCO in Table II), while concreteness curriculums bring consistent improvements.

Curriculum	Captioning BLEU-4	Verb Noun Recall Pair(%)	Verb(%)	Noun(%)	Visual Retrieval Hit@20(%)	MR
Vanilla	8.12	13.45	16.16	21.81	13.95	564.18
Sentence Level						
Loss	8.70	13.46	16.50	23.21	13.94	566.89
Soft Loss	**9.05**	**14.89**	**17.72**	**23.94**	**14.25**	**553.70**
Word Level						
Loss	8.11	12.63	15.57	21.07	12.82	606.06
Soft Loss	**9.35**	13.34	16.62	22.11	10.59	631.95
Concreteness	8.74	**14.19**	**17.00**	**22.24**	**14.51**	**513.17**
Concreteness L	8.46	13.58	16.27	22.16	14.02	561.64
Reverse Con	7.97	13.42	16.13	21.54	13.29	589.47
Random Con	8.26	13.03	15.93	21.74	12.49	595.94

Table I: Results on Yoga Videos. We highlight the best results for sentence-level and word-level respectively. For visual feature retrieval, MR is the mean rank of target videos among 2332 test videos. Concreteness refers to learned concreteness curriculum that automatically compute concreteness scores. Concreteness L refers to linguistic concreteness curriculum. Random Con and Reverse Con are random and reversed baselines for the concreteness curriculum.

Besides, we also notice that curriculums based on word losses show inferior performance in some metrics on Yoga Videos. We present further analysis later by probing into pretrained posterior distribution, together with other interesting observations below.

Method		Number of Training Images 4k	8k	16k	32k	Whole
Vanilla		21.51	23.13	24.75	25.83	28.04
Con L		22.66	24.10	24.88	26.03	29.41
Con		22.08	23.66	25.34	26.89	28.78
W	Self	**21.68**	**23.66**	24.81	25.97	27.87
	Held	21.38	23.38	**25.41**	26.39	/
	Whole	21.25	23.19	25.05	26.61	/
S	Self	**22.06**	23.72	24.81	**26.37**	27.59
	Held	21.76	**23.85**	24.56	26.07	/
	Whole	21.85	23.41	**25.55**	25.99	/

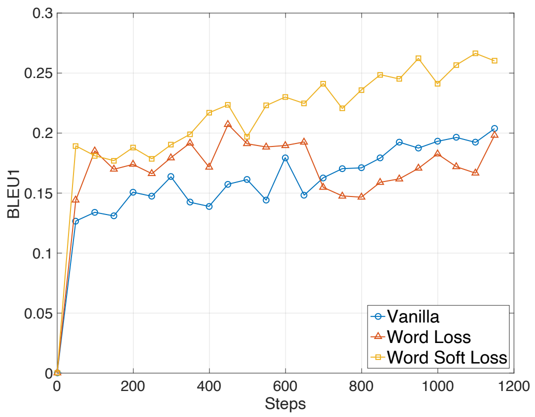
Table II: BLEU-4 scores (%) trained on various subsets of MS COCO with pretrained posteriors obtained from self, held and whole subsets. Column of Whole contain scores with models trained on whole training data ($\sim$ 113k), for which we only considered pretrained posteriors from itself. Con refers to learned concreteness curriculum. Con L refers to linguistic concreteness curriculum. Random Con and W and W are sentence-level and word-level version of soft loss curriculum.

a) Loss-based curriculum accelerates training in early stage.: We show learning curves of loss and soft loss curriculums with respect to gradient steps in Figure 2a-2b for Yoga Videos. We use BLEU-4 for sentence-level experiments, and BLEU-1 for word-level. We use BLEU-1 for word-level because it can better reflects the learning of individual words for the word-level curriculums. Loss-based curriculums learn faster at the beginning of training, and soft sampling methods are even faster than easy subset sampling. We also show learning curves on MS COCO under 4k and 32k training datasets. Here we use BLEU-4 scores to compare both the word-level and sentence-level methods.

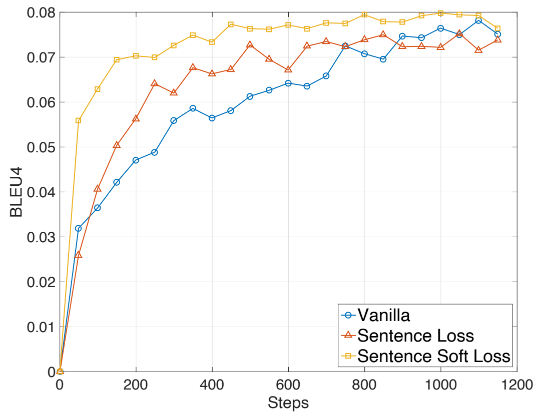
b) Concreteness improves data efficiency.: We show test BLEU-4 scores of proposed concreteness curriculum and linguistic concreteness curriculum with respect to training data size in 4. Having noticed the linearity of these curves when data size is in logscale, we run linear regression as shown in Table III.

Here the incline factor a indicates the data efficiency. In fact, suppose we have n training instances for vanilla methods. If n is relatively large such that bias terms b can be ignored, approximate numbers of training samples for the learned concreteness and linguistic concreteness methods to achieve similar performances are $n_1 = n^{\frac{1.934}{2.028}} = n^{0.954}$ and $n_2 = n^{\frac{1.934}{1.975}} = n^{0.979}$. For example if $n = 50$k, we have $n_1 = 30$k and $n_2 = 40$k, reducing 40% and 20% training instances respectively. We notice that the learned concreteness curriculum has a larger a than the linguistic concreteness curriculum. This is possibly because of the gap between vision-text dataset and human annotated concreteness scores that also consider modality other than vision such as sound, smell and taste.

c) Better Captioning can be worse language acquisition.: Soft word loss achieves good captioning performance but much worse verb-noun recall and visual retrieval performance. This indicates that the method captures good sentence patterns but fail to improve the learning of semantics and concrete concepts in multimedia data. In Table IV we have ex-



(a) BLEU-1 score on Yoga Videos test set with respect to gradient steps.



(b) BLEU-4 score on Yoga Videos test set with respect to gradient steps.

$y = a \log x + b$	a	b	R
Vanilla	1.934	5.709	0.996
Concreteness A	2.028	5.483	0.994
Concreteness	1.975	6.124	0.992

Table III: Linear regression of BLEU-4 with respect to log(TrainingSize). R is correlation coefficient.

amples of vision-word pairs (v, w_j) with higher posterior $P_{\hat{F}}(w_j|v, w_{1:j-1})$. Word pretrained posteriors are affected by the word frequency and render lower losses for frequent words like `the`, `your`. Due to auto-regressive modeling, word loss curriculums are better at capturing nouns, descriptive adjectives, and adverbs than verbs since verbs usually have shorter prior context. These features help produce complete and natural sentences, but harm the grounding of language semantics into visual scenes by suppressing some concrete words. We believe good language acquisition methods should succeed in both aspects, and captioning cannot represent language acquisition. Pretrained posterior is a good measure of vision-word pair difficulty for captioning but less effective for language acquisition.

No	Examples
1	 Lower **the right knee**.
2	 Relax **your** arms, and lift **your heart**.
3	 **sweep the** arms **all the way up** and **overhead**.

Table IV: Examples of vision-word pairs. We bold the words that have lower losses from the pretrained model.

C. Impact of Multimedia Learning

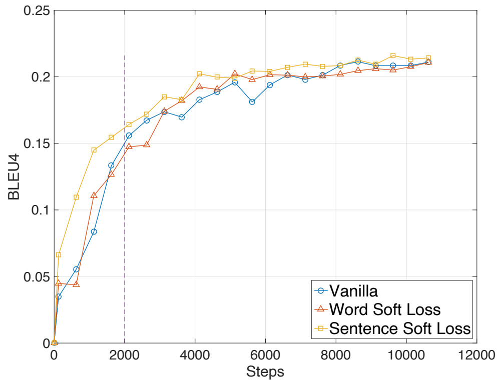
Pairs	Text	Vanilla	Con L	Con
left, right	0.5926	0.4471	0.3681	**0.3599**
open, close	0.1422	0.0881	0.0835	**0.0724**
up, down	0.2133	0.1528	0.1532	**0.1485**
straighten, bend	0.1989	0.1734	**0.1195**	0.1527
spread, bend	0.1411	0.0810	**0.0374**	0.0525

Table V: Cosine similarities between opposite word pairs. We compare embeddings from four language mdoels: Text refers to a language model trained on text corpus of Yoga Videos with the same architecture as captioning decoder; Vanilla is the vanilla captioning decoder; Con refers to learned concreteness curricu-lum that automatically compute concreteness scores. Con L refers to linguistic concreteness curriculum. We highlight the lowest similarity score for each pair.

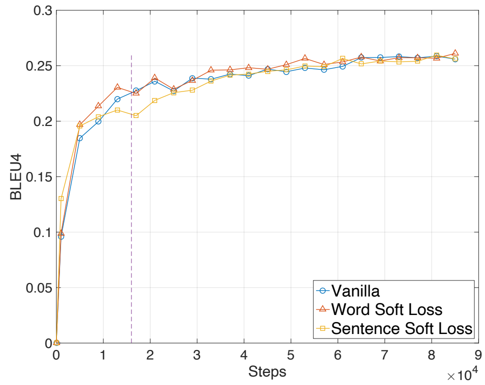
Some concepts are hard to learn only from textual context. For instance distinguishing words sharing highly similar textual context with opposite meaning (e.g. `left` and `right`) can be difficult. However, learning from multimedia dataset can compensate this deficiency. We use Yoga Videos for qualitative analysis and take the parameter of decoder output layers as word embeddings. Cosine word similarities between 5 pairs of opposite words are shown in Tabel V. We can see that multimedia models learn less similar embeddings for these opposite words. Besides, concreteness curriculums stress on these concrete words and are very helpful with distinguish these opposite concepts.

IV. Related Work

Language Acquisition: Classical theories about children language acquisition include the nativist



(a) BLEU-4 score on MS COCO test set with respect to gradient steps using 4k training images and curriculum from held-out data.



(b) BLEU-4 score on MS COCO test set with respect to gradient steps using 8k training images and curriculum from held-out data.

Fig. 3: Learning curves with various number of training images on MS COCO.

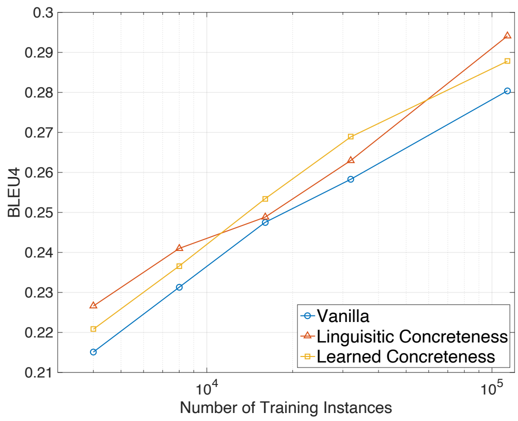


Fig. 4: Test Bleu-4 on MS COCO using varying number of training instances

theory [14], the learning theory [15], and the social interactionist theory [16]. There is also work in lan-guage acquisition from the NLP community [17], [18],[19], [20], [21], [22], [23], [24], [25], but they all focus on a specific language phenomenon or task. On the contrary, we formulate multimedia language acquisition in a general way as modeling two aspects and explore impact of curriculum learning.

Curriculum Learning: Various human language learning theories [26], [27], [28] all state that humans learn language much better when the learning materials are organized in increasing order of difficulty instead of random order. [29], [30], [31] bootstrap language acquisition by learning words first and then grammatical structures later. Inspired from these theories, some recent work applies curriculum learning [2] to NLP applications including name tagging [32], question answering [33], [34], neural machine translation[35], [36], [37], [38], coreference resolution [39], semantic parsing [40] and dialog systems [41], [42]. In addition to NLP applications, how curriculum learning can help on noisy dataset [43] or smaller dataset [44] is also studied. [45] and [3] propose curriculum learning by transfer and theoretically analyze the effect of loss-based curriculum learning methods under certain conditions.

Image Captioning, Video Captioning and Multimodal Language Modeling: Our task is related to image captioning [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59] and video captioning [23], [60], [61], [62], [63], [64], [65], [66]. However, we focus on curriculum learning for language acquisition. Therefore, we only use simpler encoder-decoder framework to demonstrate the power of our proposed curriculum learning methods, and our evaluation is not only based on captioning. Also it is worth mentioning that our proposed CL methods can be naturally adapted to most state-of-the-art captioning methods. Another line of work on multimodal language modeling [67], [68], [69], [70], [71], [72], [73], [74], [75], [76] focuses on use large-scale multimodal resources for universal pre-training that can benefit downstream tasks, while our work aims at effectively using smaller multimodal corpora for language acquisition.

V. CONCLUSIONS AND FUTURE WORK

We explore transfer-based and concreteness-based curriculum learning, both of which can be derived from multimedia data alone without additional human guidance. We observe that both transfer-based methods are effective in improving learning speed for captioning in early stage, and our proposed concreteness curriculum is a more effective framework in acquisition of reliable language knowledge with more consistent final performance across various settings in both directions. Concreteness curriculum also improves data efficiency. We also found that multimedia features can compensate contextual bias in small text data for language acquisition.

Our work explores curriculum learning in language acquisition. We model visual retrieval as Bayesian inference based on captioning model such that the captioning curriculum learning can be directly used to compare on visual retrieval, but its performance is less desirable than training a specialized model for visual retrieval. Besides, more advanced curriculum learning strategies may be applied to our concreteness metric, such as better sampling schedule instead of simple two-stage curriculum , and adding homework into our curriculum so that the learner's performance on homework can be exploited to dynamically adjust future learning materials. We also plan to explore more visual features such as motion dynamics, temporal action compositions to further enhance our curriculum.

ACCKNOWLEGDE

This research is based upon work supported in part by U.S. DARPA GAILA Program HR00111990058. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, ser. Lecture Notes in Computer Science, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8693. Springer, 2014, pp. 740–755. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_48

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, ser. ACM International Conference Proceeding Series, A. P. Danyluk, L. Bottou, and M. L. Littman, Eds., vol. 382. ACM, 2009, pp. 41–48. [Online]. Available: https://doi.org/10.1145/1553374.1553380

[3] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 2535–2544. [Online]. Available: http://proceedings.mlr.press/v97/hacohen19a.html

[4] J. C. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2021068

[5] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012. [Online]. Available: http://arxiv.org/abs/1212.5701

[6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[7] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[9] J. Hessel, D. Mimno, and L. Lee, "Quantifying the visual concreteness of words and topics in multimodal datasets," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2194–2205. [Online]. Available: https://www.aclweb.org/anthology/N18-1199

[10] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known english word lemmas," *Behavior Research Methods*, vol. 46, no. 3, pp. 904–911, 2014. [Online]. Available: https://doi.org/10.3758/s13428-013-0403-5

[11] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[12] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664–676, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2598339

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://www.aclweb.org/anthology/P02-1040

[14] N. Chomsky, *Knowledge of Language: Its Nature, Origin and Use*. New York, NY: Praeger, 1986.

[15] J. R. Saffran, "Statistical language learning: mechanisms and constraints," *Current Directions in Psychological Science. 12 (4): 110–114*, 2003.

[16] C. Gallaway and B. Richard, *Input and Interaction in Language Acquisition*. Cambridge University Press, 1994.

[17] B. Pedersen, S. Edelman, Z. Solan, D. Horn, and E. Ruppin, "Some tests of an unsupervised model of language acquisition," in *Proceedings of the Workshop on Psycho-Computational Models of Human Language Acquisition*, 2004.

[18] A. Clark, "Grammatical inference and first language acquisition," in *Proceedings of the Workshop on Psycho-Computational Models of Human Language Acquisition*, 2004.

[19] M. Connor, Y. Gertner, C. Fisher, and D. Roth, "Baby srl: Modeling early language acquisition," in *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL)*, 2008.

[20] ——, "Minimally supervised model of early language acquisition," in *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL)*, 2009.

[21] G. Aimetti, "Modelling early language acquisition skills: Towards a general statistical learning mechanism," in *Proceedings of the Student Research Workshop at EACL 2009*. Athens, Greece: Association for Computational Linguistics, Apr. 2009, pp. 1–9. [Online]. Available: https://www.aclweb.org/anthology/E09-3001

[22] D. Chen, "Fast online lexicon learning for grounded language acquisition," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, Jul. 2012, pp. 430–439. [Online]. Available: https://www.aclweb.org/anthology/P12-1045

[23] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, Portland, OR, USA, June 23-28, 2013*. IEEE Computer Society, 2013, pp. 2634–2641. [Online]. Available: https://doi.org/10.1109/CVPR.2013.340

[24] S. Kottur, J. Moura, S. Lee, and D. Batra, "Natural language does not emerge 'naturally' in multi-agent dialog," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2962–2967. [Online]. Available: https://www.aclweb.org/anthology/D17-1321

[25] J. Kodner and C. Cerezo Falco, "A framework for representing language acquisition in a population setting," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1149–1159. [Online]. Available: https://www.aclweb.org/anthology/P18-1106

[26] B. F. Skinner, *Verbal Behavior*. Acton, MA: Copley Publishing Group, 1957.

[27] G. B. Peterson, "A day of great illumination: B. f. skinner's discovery of shaping," *Journal of the Experimental Analysis of Behavior, 82, 317–328*, 2004.

[28] K. A. Krueger and P. Dayan, "Flexible shaping: How learning in small steps helps," *Cognition*, vol. 110, no. 3, pp. 380 – 394, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010027708002850

[29] N. Z. Kirkham, A. J. Slemmer, and S. P. Johnson, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, p. 781–799, 1993.

[30] ——, "Visual statistical learning in infancy: Evidence for a domain general learning mechanism," *Cognition*, vol. 83, pp. B35–B42, 2002.

[31] O. Abend, T. Kwiatkowski, N. J. Smith, S. Goldwater, and M. Steedman, "Bootstrapping language acquisition," *Cognition*, vol. 164, pp. 116–143, Jul. 2017.

[32] C. Cardellino, M. Teruel, L. Alonso Alemany, and S. Villata, "Legal NERC with ontologies, Wikipedia and curriculum learning," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 254–259. [Online]. Available: https://www.aclweb.org/anthology/E17-2041

[33] M. Sachan and E. P. Xing, "Easy questions first? a case study on curriculum learning for question answering," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.

[34] C. Liu, S. He, K. Liu, and J. Zhao, "Curriculum learning for natural answer generation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, J. Lang, Ed. ijcai.org, 2018, pp. 4223–4229. [Online]. Available: https://doi.org/10.24963/ijcai.2018/587

[35] T. Kocmi and O. Bojar, "Curriculum learning and minibatch bucketing in neural machine translation," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., Sep. 2017, pp. 379–386. [Online]. Available: https://doi.org/10.26615/978-954-452-049-6_050

[36] X. Zhang, P. Shapiro, G. Kumar, P. McNamee, M. Carpuat, and K. Duh, "Curriculum learning for domain adaptation in neural machine translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1903–1915. [Online]. Available: https://www.aclweb.org/anthology/N19-1189

[37] E. A. Platanios, O. Stretcu, G. Neubig, B. Poczos, and T. Mitchell, "Competence-based curriculum learning for neural machine translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1162–1172. [Online]. Available: https://www.aclweb.org/anthology/N19-1119

[38] X. Liu, H. Lai, D. F. Wong, and L. S. Chao, "Norm-based curriculum learning for neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 427–436. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.41

[39] D. Huang, S. Buch, L. M. Dery, A. Garg, L. Fei-Fei, and J. C. Niebles, "Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 5948–5957. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Huang_Finding_It_Weakly-Supervised_CVPR_2018_paper.html

[40] C. Ross, A. Barbu, Y. Berzak, B. Myanganbayar, and B. Katz, "Grounding language acquisition by training semantic parsers using captioned videos," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2647–2656. [Online]. Available: https://www.aclweb.org/anthology/D18-1285

[41] A. Saito, "Curriculum learning based on reward sparseness for deep reinforcement learning of task completion dialogue management," in *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 46–51. [Online]. Available: https://www.aclweb.org/anthology/W18-5707

[42] L. Shen and Y. Feng, "CDL: Curriculum dual learning for emotion-controllable response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 556–566. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.52

[43] L. Jiang, Z. Zhou, T. Leung, L. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 2309–2318. [Online]. Available: http://proceedings.mlr.press/v80/jiang18c.html

[44] Y. Fan, F. Tian, T. Qin, X. Li, and T. Liu, "Learning to teach," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=HJewuJWCZ

[45] D. Weinshall, G. Cohen, and D. Amir, "Curriculum learning by transfer learning: Theory and experiments with deep networks," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 5235–5243. [Online]. Available: http://proceedings.mlr.press/v80/weinshall18a.html

[46] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6632

[47] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE

Computer Society, 2015, pp. 3156–3164. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298935

[48] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 2048–2057. [Online]. Available: http://proceedings.mlr.press/v37/xuc15.html

[49] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 2422–2431. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298856

[50] Q. Wu, C. Shen, L. Liu, A. R. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 203–212. [Online]. Available: https://doi.org/10.1109/CVPR.2016.29

[51] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton key: Image captioning by skeleton-attribute decomposition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 7378–7387. [Online]. Available: https://doi.org/10.1109/CVPR.2017.780

[52] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1141–1150. [Online]. Available: https://doi.org/10.1109/CVPR.2017.127

[53] J. Gu, G. Wang, J. Cai, and T. Chen, "An empirical study of language CNN for image captioning," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 1231–1240. [Online]. Available: https://doi.org/10.1109/ICCV.2017.138

[54] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, New Orleans, Louisiana, USA*. AAAI Press, 2018, pp. 6837–6844. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16465

[55] J. Liu, K. Wang, C. Xu, Z. Zhao, R. Xu, Y. Shen, and M. Yang, "Interactive dual generative adversarial networks for image captioning," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, New York, New York, USA*. AAAI Press, 2020.

[56] P. H. Seo, P. Sharma, T. Levinboim, B. Han, and R. Soricut, "Reinforcing an image caption generator using off-line human feedback," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, New York, New York, USA*. AAAI Press, 2020.

[57] L. Wang, Z. Bai, Y. Zhang, and H. Lu, "Show, recall, and tell: Image captioning with recall mechanism," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, New York, New York, USA*. AAAI Press, 2020.

[58] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, New York, New York, USA*. AAAI Press, 2020.

[59] W. Zhao, X. Wu, and X. Zhang, "Memcap: Memorizing style knowledge for image captioning," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, New York, New York, USA*. AAAI Press, 2020.

[60] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko, "Improving lstm-based video description with linguistic knowledge mined from text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[61] N. Laokulrat, S. Phan, N. Nishida, R. Shu, Y. Ehara, N. Okazaki, Y. Miyao, and H. Nakayama, "Generating video description using sequence-to-sequence model with temporal attention," in *Proceedings of the 26th International Conference on Computational Linguistics*, 2016.

[62] R. Pasunuru and M. Bansal, "Multi-task video captioning with video and entailment generation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1273–1283. [Online]. Available: https://www.aclweb.org/anthology/P17-1117

[63] ——, "Reinforced video captioning with entailment rewards," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 979–985. [Online]. Available: https://www.aclweb.org/anthology/D17-1103

[64] S. Whitehead, H. Ji, M. Bansal, S.-F. Chang, and C. Voss, "Incorporating background knowledge into video description generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3992–4001. [Online]. Available: https://www.aclweb.org/anthology/D18-1433

[65] X. Long, C. Gan, and G. de Melo, "Video captioning with multi-faceted attention," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 173–184, 2018. [Online]. Available: https://www.aclweb.org/anthology/Q18-1013

[66] X. Xiao, L. Wang, B. Fan, S. Xiang, and C. Pan, "Guiding the flowing of semantics: Interpretable video captioning via POS tag," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2068–2077. [Online]. Available: https://www.aclweb.org/anthology/D19-1213

[67] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 7463–7472. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00756

[68] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5100–5111. [Online]. Available: https://www.aclweb.org/anthology/D19-1514

[69] G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," *CoRR*, vol. abs/1908.06066, 2019. [Online]. Available: http://arxiv.org/abs/1908.06066

[70] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: pre-training of generic visual-linguistic representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SygXPaEYvH

[71] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang, "Visualbert: A simple and performant baseline for vision and language," *CoRR*, vol. abs/1908.03557, 2019. [Online]. Available: http://arxiv.org/abs/1908.03557

[72] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: learning universal image-text representations," *CoRR*, vol. abs/1909.11740, 2019. [Online]. Available: http://arxiv.org/abs/1909.11740

[73] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural*

Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 13–23. [Online]. Available: http://papers.nips.cc/paper/8297-vilbert-pretraining-task-agnostic-visiolinguistic-representatio ns-for-vision-languege-task

[74] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Contrastive bidirectional transformer for temporal representation learning," *CoRR*, vol. abs/1906.05743, 2019. [Online]. Available: http://arxiv.org/abs/1906.05743

[75] W. Rahman, M. K. Hasan, A. Zadeh, L. Morency, and M. E. Hoque, "M-BERT: injecting multimodal information in the BERT structure," *CoRR*, vol. abs/1908.05787, 2019. [Online]. Available: http://arxiv.org/abs/1908.05787

[76] A. Ororbia, A. Mali, M. Kelly, and D. Reitter, "Like a baby: Visually situated neural language acquisition," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5127–5136. [Online]. Available: https://www.aclweb.org/anthology/P19-1506

[77] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 7753–7762. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Pavllo_3D_Human_Pose_Estimation_in_Video_With_Temporal_Convolutions_and_CVPR_2019_paper.html

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

APPENDIX

A. Other Captioning Metrics

We show other captioning metrics for MS COCO in Table VI and Table VII.

B. Preprocessing and Hyperparameters

For Yoga Videos, we preprocess the original videos into 2D poses using AlphaPose [2] and transform 2D poses into 3D using VideoPose3D[77][3]. For each frame, output 3D pose is a 51-dimensional coordinate vector of 17 joints. We encode pose sequences into $\mathbf{v}$ as final hidden states of a 2-layer bidirectional GRU with hidden size 768 (384 for each direction). For decoder, we use another 2-layer unidirectional GRU with hidden size 768 and an output layer to map hidden states into word distribution. We used 768-dimensional input word embeddings. For training, we use AdamW [7] with learning rate $1e-3$. For sentence level training, we use batch size 64. Word level training batch size selection is elaborated in Appendix D. Maximum number of training epoch is 60. We evaluate models on validation set every 50 gradient steps and stop training if performance is not improved in consecutive 10 evaluations. For Yoga Videos we run all methods with 3 random seeds and report results using average scores over 3 runs.

For MS COCO[1], we adapt the implementation in https://github.com/ruotianluo/self-critical.pytorch for curriculum learning. Following their parameter settings,or-vision-weand-collectlanguagve-tasksocabulary with words appear at least 5 times, resulting in 9488 words. We use the image features from last layer of pretrained ResNet101 [78] as v. We use single-layer LSTM[8] with hidden size 512 as decoder. We use Adam [6] with inital learning rate 0.0005, which is decayed with factor 0.8 every 3 epochs. In MS COCO each image is associated with multiple captions. For sentence level training, we use batch size 10 to sample images, and sample 5 captions for each image. To make sure each vision-sentence pairs are sampled uniformly, we sample images with probability proportional to number of associated captions and sample captions for each image uniformly. Word level training batch size selection is elaborated in Appendix D. We train models for 30 epochs and best models are selected according to performance on validation set.

For curriculum learning methods, we set the max-imal number of curriculum training epochs as 5. The corresponding maximal number of training steps N may vary since the number of training instances varies across datasets. We add early-stop mechanism in curriculum training, which will stop the curricu-lum when the training losses converge. For concrete-ness curriculums, we experiment with $(\lambda_1, \lambda_2) \in \{(0, 1), (0.5, 0.5), (1, 0)\}$ on Yoga Videos and select the best combination $(0, 1)$ for all the experiments.

C. Collection of Verb-Noun Pairs for Verb-Noun Recall

We manually select 38 (see Table VIII) frequent body-part nouns from all gold captions, and run part-of-speech tagging using spaCy [11] on gold captions to obtain head verbs for these nouns. In this way we collect $1 \sim 2$ gold verb-noun pairs for each gold caption.

D. Additonal Dataset and Experiment Details

We use single Nvidia Tesla V100 with 16GB DRAM for all experiments. Numbers of parameters for models on Yoga Videos are all $11,946,247$, and for models on MS COCO are $13,400,848$ (not including pretrained resnet). We use implementation in https://github.com/ruotianluo/self-critical.pytorch to compute BLEU, CIDEr and ROUGE scores. Number of hyper-parameter search trials for λ_1, λ_2 and λ are 3, and we select the ones with best results as stated previously. Yoga Videos dataset is collected automatically from yoga videos on YouTube, and textual captions are closed captions provided by YouTube. We collect short

[2] https://github.com/MVIG-SJTU/AlphaPose

[3] https://github.com/facebookresearch/VideoPose3D

Method		B-1	B-4	C	R-L	B-1	B-4	C	R-L
		4k training images				8k training images			
Vanilla		64.22	21.51	63.25	46.79	66.48	23.13	71.08	48.32
Concreteness		65.55	22.66	68.14	47.86	67.84	24.10	74.30	48.79
Concreteness A		65.34	21.86	67.36	47.42	67.26	23.87	73.47	48.83
Word	Self	64.64	**21.68**	**64.81**	**47.12**	66.97	**23.66**	72.99	48.62
	Held	63.83	21.38	63.25	46.88	66.68	23.38	72.31	48.33
	Whole	**64.66**	21.25	64.55	46.90	**67.28**	23.19	**73.04**	**48.66**
Sentence	Self	**64.75**	**22.06**	66.23	47.26	66.85	23.72	**73.03**	48.70
	Held	64.48	21.76	65.92	47.06	66.67	**23.85**	72.89	**48.76**
	Whole	64.46	21.85	**66.56**	**47.33**	**66.96**	23.41	72.96	48.64
		16k training images				32k training images			
Vanilla		68.26	24.75	77.94	49.48	69.62	25.83	82.49	50.42
Concreteness		68.28	24.88	79.05	49.58	70.05	26.30	84.89	51.00
Concreteness A		68.35	25.03	80.01	49.86	70.03	26.48	85.80	50.84
Word	Self	68.51	24.81	79.27	49.61	69.63	25.97	83.30	50.59
	Held	68.69	**25.41**	79.13	**49.97**	69.96	26.39	83.93	50.77
	Whole	**68.85**	25.05	**79.52**	49.86	**70.18**	**26.61**	**84.57**	**50.86**
Sentence	Self	68.19	24.81	78.02	49.64	69.56	**26.37**	**82.90**	**50.58**
	Held	68.18	24.56	77.61	49.58	69.33	26.07	81.67	50.42
	Whole	**68.59**	**25.55**	**79.18**	**49.77**	**69.72**	25.99	82.69	50.43

Table VI: Captioning Results trained on various subsets of MS COCO

Method	B-1	B-4	C	R-L
Vanilla	71.36	28.04	90.20	51.87
Concreteness	**72.46**	**29.41**	**94.30**	**52.54**
Concreteness A	72.09	28.78	92.20	52.16
Word	71.29	27.87	90.38	52.02
Sentence	71.02	27.59	87.80	51.66

Table VII: Captioning Results trained on whole MS COCO

waist, tongue, mouth, nose, thigh, elbow, ear, thumb, forearm, neck, foot, cheek, hand, lip, eyelash, fist, fingertip, leg, back, knee, bum, head, belly, calf, forehead, hair, toe, eye, shoulder, hip, finger, chin, nostril, arm, bottom, rib, ankle, wrist

Table VIII: list of manually selected nouns.

clips that presents yoga actions by selecting video segments whose captions contain the nouns listed in table VIII.

We compute word concreteness scores using presentation $\mathbf{v}$ following [9]. For Yoga Videos we use $\mathbf{v}$ from pretrained vanilla models. We first compute cosine similarity between two videos/images $S(v, v') = \cos(\mathbf{v}, \mathbf{v}')$. We find k nearest neighbouring video clips for every clip v using this similarity measure, denoted as $\mathrm{NN}_k(v)$. We denote the set of video clips whose paired captions contain w as V_w:

$$V_w = \{v | \exists (v, w_{1:n}) \in \mathcal{D} \text{ and } j \in [1, n], w_j = w\}. \quad (4)$$

The concreteness of w is computed as

$$c_w = \frac{\sum_{v \in V_w} |\mathrm{NN}_k(v) \cap V_w|}{|V_w|^2} \quad (5)$$

In experiments $k = 50$ and we use Annoy[4] library to compute approximate neareast neighbours following [9].

Word level training requires sampling (v, w_j) from training data. However, due to the sequential computation of LSTMs and GRUs, it is highly inefficient to train on only one word in sentences by minimizing $-\log P(w_j | v, w_{1:j})$. To improve the sampling efficiency of word-level training, we approximate this training process by associating weights to the sentence level losses as follows

$$L(v, w_{1:n}, F) = \sum_{j=1}^{n} -p(v, j) \log P_F(w_j | v, w_{1:j})$$

where $p(v, j)$s are weights associated with (v, w_j). For vanilla training, all $p(v, j) = 1$. For easy subset sampling curriculums, $p(v, j) = 1$ for pairs in the easy subset and $p(v, j) = 0$ otherwise. For soft sampling, $p(v, j)$ is proportional to sampling probability of (v, w_j). In this way, if we sample each vision-sentence pair uniformly, the overall training objective is equivalent to sampling vision-word pairs with corresponding sampling distributions. We use the same batch size to sample vision-sentence pairs for soft sampling. For

[4]https://github.com/spotify/annoy

easy subset sampling, since we have half of instances in the easy subset, the expected number of vision-word pairs with $p(v, j) = 1$ in each sentence is also half of the sentence length. We therefore doubled sentence batch size for easy subset sampling for fair comparison of learning pace with respect to gradient steps, although we notice similar trends in performance without doubling the batch size. After we sampled a batch of vision-word pairs following the above procedure, we normalize the loss weights $p(v, j)$ to sum 1 within the batch to balance the learning rate.

We notice close performance for vanilla sentence level training and vanilla word level training with above approximation of sampling (note that vanilla training objectives are the same for sentence level and word level), which shows that above appoximation is effective. We report vanilla performance as average of sentence-level and word level since they are close enough.