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

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

\[
F^* = \arg \min_F \mathbb{E}_{d \sim U(D)} L(F, x_i, y_i).
\]

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

Let \( Q : D \mapsto \mathbb{R} \) measure the difficulty of instances. Since CL utilizes a training procedure that favors easy instances in the early stage, we expect \( E_{(x_i, y_i) \sim \varphi_i} Q(x_i, y_i) \) 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 posteriors: \( P(\text{text|scene}) \) (captioning) and \( P(\text{scene|text}) \) (visual retrieval). We introduce how we model \( P(\text{text|scene}) \) for captioning and \( P(\text{scene|text}) \) 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 \( w_i \in \mathcal{V} \) 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,

\[
v = \text{VisualEncoder}(v).
\]

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

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

The final output from LSTM decoder is a conditional distribution over all sentences given \( v \), and we summarize the captioning model \( F \) as \( F(v) \in \)
for sentence-level curriculum training, we minimize the objective function below for step i,

\[ \mathbb{E}_{(v,w_{1:i}) \sim \varphi_i} L(F, v, w_{1:n}) \]
\[ = \mathbb{E}_{(v,w_{1:i}) \sim \varphi_i} - \log P_F(w_{1:n}|v) \]
\[ = \mathbb{E}_{(v,w_{1:i}) \sim \varphi_i} - \sum_{j=1}^{n} \log P_F(w_j|w_{1:j-1}) \]

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

\[ \mathbb{E}_{(v,w_{1:i}) \sim \mathcal{U}(\mathcal{D})} \varphi_i(v, w_j) \log P_F(w_j|w_{1:j-1}) \]

For vanilla training, \( \varphi_1 = \mathcal{U}(\mathcal{D}) \), 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 a Gaussian Mixture model on hidden representations \( \text{v} \) 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

\[ \text{Score}(\text{scene}|\text{text}) = \log P(\text{text}|\text{scene}) \]
\[ + \lambda \log P(\text{scene}) \]

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 benefit 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 \( F \) 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 \( Q(w_{1:n}, v) = -\log P_F(w_{1:n}|v) \); for vision-word pairs, we define \( Q(w_i, v) = -\log P_F(w_i|v, w_{1:i-1}) \).

**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 \( v \) 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

\[ \varphi_i = \begin{cases} \mathcal{U}(\mathcal{D}^E) & i \leq N \\ \mathcal{U}(\mathcal{D}) & i > N \end{cases} \]

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 \( |\mathcal{D}^E| = |\mathcal{D}| \cdot |\mathcal{D}^E| \). 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 \( F \) and corresponding posterior distribution \( P_F \), we define the soft sampling strategy for vision-sentence pairs as

\[ \varphi_i = \begin{cases} \propto P_F(w_{1:n}|v) & i \leq N \\ \mathcal{U}(\mathcal{D}) & i > N \end{cases} \]

and for vision-word pairs as

\[ \varphi_i = \begin{cases} \propto P_F(w_i|v, w_{1:i-1}) & i \leq N \\ \mathcal{U}(\mathcal{D}) & i > N \end{cases} \]

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 ~ 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

$$MR = \frac{1}{|D_{test}|} \cdot \sum_{(v,w_{1:n}) \in D_{test}} \text{Rank}(v|w_{1:n}).$$

Here $D_{test}$ refers to the test dataset and 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 pre-train posteriors are more beneficial for smaller datasets (see results on MS COCO in Table II), while concreteness curriculums bring consistent improvements. 
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of Training Images</th>
<th>4k</th>
<th>8k</th>
<th>16k</th>
<th>32k</th>
<th>Whole</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td></td>
<td>21.51</td>
<td>23.13</td>
<td>24.75</td>
<td>25.83</td>
<td>28.04</td>
</tr>
<tr>
<td>Con L  Con</td>
<td></td>
<td>22.66</td>
<td>24.10</td>
<td>24.88</td>
<td>26.03</td>
<td>29.41</td>
</tr>
<tr>
<td>W Self</td>
<td></td>
<td>21.38</td>
<td>23.38</td>
<td><strong>25.41</strong></td>
<td>26.39</td>
<td>/</td>
</tr>
<tr>
<td>W Held</td>
<td></td>
<td>21.25</td>
<td>23.19</td>
<td>25.05</td>
<td>26.61</td>
<td>/</td>
</tr>
<tr>
<td>S Held</td>
<td></td>
<td>21.76</td>
<td><strong>23.85</strong></td>
<td>24.56</td>
<td>26.07</td>
<td>/</td>
</tr>
<tr>
<td>Whole</td>
<td></td>
<td>21.85</td>
<td>23.41</td>
<td><strong>25.55</strong></td>
<td>25.99</td>
<td>/</td>
</tr>
</tbody>
</table>

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 contains scores with models trained on whole training data (~113k), for which we only consider 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 reflect 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of Training Images</th>
<th>4k</th>
<th>8k</th>
<th>16k</th>
<th>32k</th>
<th>Whole</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td></td>
<td>21.51</td>
<td>23.13</td>
<td>24.75</td>
<td>25.83</td>
<td>28.04</td>
</tr>
<tr>
<td>Con L  Con</td>
<td></td>
<td>22.66</td>
<td>24.10</td>
<td>24.88</td>
<td>26.03</td>
<td>29.41</td>
</tr>
<tr>
<td>W Self</td>
<td></td>
<td>21.38</td>
<td>23.38</td>
<td><strong>25.41</strong></td>
<td>26.39</td>
<td>/</td>
</tr>
<tr>
<td>W Held</td>
<td></td>
<td>21.25</td>
<td>23.19</td>
<td>25.05</td>
<td>26.61</td>
<td>/</td>
</tr>
<tr>
<td>S Held</td>
<td></td>
<td>21.76</td>
<td><strong>23.85</strong></td>
<td>24.56</td>
<td>26.07</td>
<td>/</td>
</tr>
<tr>
<td>Whole</td>
<td></td>
<td>21.85</td>
<td>23.41</td>
<td><strong>25.55</strong></td>
<td>25.99</td>
<td>/</td>
</tr>
</tbody>
</table>

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.
Pretrained posterior is a good measure of vision-word acquisition methods should succeed in both aspects, ing some concrete words. We believe good language semantics into visual scenes by suppress- and natural sentences, but harm the grounding of prior context. These features help produce complete adverbs than verbs since verbs usually have shorter better at capturing nouns, descriptive adjectives, and auto-regressive modeling, word loss curriculums are losses for frequent words like

\[ \log(\text{TrainingSize}) \].

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

<table>
<thead>
<tr>
<th></th>
<th>( a )</th>
<th>( b )</th>
<th>( R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>1.934</td>
<td>5.709</td>
<td>0.996</td>
</tr>
<tr>
<td>Concreteness A</td>
<td>2.028</td>
<td>5.483</td>
<td>0.994</td>
</tr>
<tr>
<td>Concreteness</td>
<td>1.975</td>
<td>6.124</td>
<td>0.992</td>
</tr>
</tbody>
</table>

Examples of vision-word pairs \((v, w_j)\) with higher posterior \( P_k(v, w_j \mid v, w_{1:j-1}) \). 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.

<table>
<thead>
<tr>
<th>No</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1.png" alt="Example 1" /> Lower the right knee.</td>
</tr>
<tr>
<td>2</td>
<td><img src="image2.png" alt="Example 2" /> Relax your arms, and lift your heart.</td>
</tr>
<tr>
<td>3</td>
<td><img src="image3.png" alt="Example 3" /> sweep the arms all the way up and overhead.</td>
</tr>
</tbody>
</table>

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

C. Impact of Multimedia Learning

<table>
<thead>
<tr>
<th>Pairs</th>
<th>Text Vanilla Con L Con</th>
</tr>
</thead>
<tbody>
<tr>
<td>left, right</td>
<td>0.5926 0.4471 0.3681 0.3599</td>
</tr>
<tr>
<td>open, close</td>
<td>0.1422 0.0881 0.0835 0.0724</td>
</tr>
<tr>
<td>up, down</td>
<td>0.2133 0.1528 0.1532 0.1485</td>
</tr>
<tr>
<td>straighten, bend</td>
<td>0.1989 0.1734 0.1195 0.1527</td>
</tr>
<tr>
<td>spread, bend</td>
<td>0.1411 0.0810 0.0374 0.0525</td>
</tr>
</tbody>
</table>

Table V: Cosine similarities between opposite word pairs. We compare embeddings from four language models: 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 curriculum 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 Table 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
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

Pengfei Yu, Heng Ji, Shih-fu Chang, Kevin Duh

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.

ACKNOWLEDGE

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


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 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-collectanguage-taskssocabulary 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 initial 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-innal 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 ~ 2 gold verb-noun pairs for each gold caption.

D. Additional 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 hyperparameter search trials for $\lambda_1, \lambda_2$ and $\lambda$ 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
Table VI: Captioning Results trained on various subsets of MS COCO

<table>
<thead>
<tr>
<th>Method</th>
<th>B-1</th>
<th>B-4</th>
<th>C</th>
<th>R-L</th>
<th>B-1</th>
<th>B-4</th>
<th>C</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>64.22</td>
<td>21.51</td>
<td>63.25</td>
<td>46.79</td>
<td>66.48</td>
<td>23.13</td>
<td>71.08</td>
<td>48.32</td>
</tr>
<tr>
<td>Concreteness</td>
<td>65.55</td>
<td>22.66</td>
<td>68.14</td>
<td>47.86</td>
<td>67.84</td>
<td>24.10</td>
<td>74.30</td>
<td>48.79</td>
</tr>
<tr>
<td>Concreteness A</td>
<td>65.34</td>
<td>21.86</td>
<td>67.36</td>
<td>47.42</td>
<td>67.26</td>
<td>23.87</td>
<td>73.47</td>
<td>48.83</td>
</tr>
<tr>
<td>Self</td>
<td>64.64</td>
<td>21.68</td>
<td>64.81</td>
<td>47.12</td>
<td>69.67</td>
<td>23.66</td>
<td>72.99</td>
<td>48.62</td>
</tr>
<tr>
<td>Word</td>
<td>63.83</td>
<td>21.38</td>
<td>63.25</td>
<td>46.88</td>
<td>66.68</td>
<td>23.38</td>
<td>72.31</td>
<td>48.33</td>
</tr>
<tr>
<td>Whole</td>
<td>64.66</td>
<td>21.25</td>
<td>64.55</td>
<td>46.90</td>
<td>67.28</td>
<td>23.19</td>
<td>73.04</td>
<td>48.66</td>
</tr>
</tbody>
</table>

Table VI: 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, forehead, arm, bottom, rib, ankle, wrist

Table VIII: list of manually selected nouns.

The concreteness of $w$ is computed as

$$c_w = \frac{\sum_{v \in V_w} |N_{k_h}(v) \cap V_w|}{|V_w|^2}$$  (5)

In experiments $k = 50$ and we use Annoy4 library to compute approximate nearest neighbours following [9].

Word level training requires sampling $(v, w)$ 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 $p(v, j)$ is 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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>B-1</th>
<th>B-4</th>
<th>C</th>
<th>R-L</th>
<th>B-1</th>
<th>B-4</th>
<th>C</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>71.36</td>
<td>28.04</td>
<td>90.20</td>
<td>51.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concreteness</td>
<td>72.46</td>
<td>29.41</td>
<td>94.30</td>
<td>52.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concreteness A</td>
<td>72.09</td>
<td>28.78</td>
<td>92.20</td>
<td>52.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word</td>
<td>71.29</td>
<td>27.87</td>
<td>90.38</td>
<td>52.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence</td>
<td>71.02</td>
<td>27.59</td>
<td>87.80</td>
<td>51.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 approximation is effective. We report vanilla performance as average of sentence-level and word level since they are close enough.