Hidden State Guidance:
Improving Image Captioning Using an Image Conditioned Autoencoder

Jialin Wu and Raymond J. Mooney
Department of Computer Science
University of Texas at Austin
{jalinwu,mooney}@cs.utexas.edu

Abstract
Most RNN-based image captioning models receive supervision on the output words to mimic human captions. Therefore, the hidden states can only receive noisy gradient signals via layers of back-propagation through time, leading to less accurate generated captions. Consequently, we propose a novel framework, Hidden State Guidance (HSG), that matches the hidden states in the caption decoder to those in a teacher decoder trained on an easier task of autoencoding the captions conditioned on the image. During training with the REINFORCE algorithm, the conventional rewards are sentence-based evaluation metrics equally distributed to each generated word, no matter their relevance. HSG provides a word-level reward that helps the model learn better hidden representations. Experimental results demonstrate that HSG clearly outperforms various state-of-the-art caption decoders using either raw images, detected objects, or scene graph features as inputs.

1 Introduction
In recent years, image captioning has been widely studied in both the vision and NLP communities. Most recent research (Xu et al. 2015; Donahue et al. 2015; Karpathy and Fei-Fei 2015; Vinyals et al. 2015; Anderson et al. 2018; Yao et al. 2018; Yang et al. 2018) trains an RNN-based decoder to learn the word probabilities conditioned on the previous hidden state and various visual features. These methods improve results by incorporating richer visual inputs from object detection (Anderson et al. 2018) and relationship detection (Yao et al. 2018; Yang et al. 2018).

By contrast, we focus on improving the hidden state representation learned during training. Most current image captioners are trained using maximum log-likelihood or REINFORCE with CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) or BLEU (Papineni et al. 2002) rewards, where only the final word probabilities receive supervision. Therefore, the hidden states can only access noisy training signals from layers of backpropagation through time. Especially when training using REINFORCE, rewards are delayed till the end and equally distributed to each word in the caption, regardless of whether or not the words are descriptive, making the training signals even noisier.

We present a new framework, called Hidden State Guidance (HSG), that treats the RNN caption decoder as a student network (Romero et al. 2014) and directly guides its hidden-state learning. However, this requires a teacher to provide hidden state supervision. We use a caption autoencoder as the teacher, giving it the same image as input. Its decoder has the same architecture as the caption decoder, allowing matching of the hidden states. Since the teacher has access to all of the human captions and visual inputs, its hidden states are expected to encode a richer representation that generates better captions, and therefore, provides useful hidden state supervision.

In order to align the initial states of the teacher and student, we insert a small state transformation network that uses the visual features to estimate the initial state of the teacher decoder. During testing, run-time only slightly increases due to the light-weight state transformation network.

HSG plays a particularly helpful role when training using REINFORCE since it also provides a word-level intermediate reward that highlights the important words. For example, Figure 1 shows generated captions with the rewards indicated by the words’ background intensity. These are generated using the model trained with maximum likelihood. HSG recognizes descriptive words like “purple outfit” and “glove,” and rewards them more than digressive words like...
“waiting” and “swinging”.

Our general framework can be used in almost any RNN-based image captioner. Experimental results show significant improvements over three recent caption decoders, FC (Rennie et al. 2017) using image features, Up-Down (Anderson et al. 2018) using object detection features, and SGAE (Yang et al. 2018) using scene graph features.

2 Related Work

2.1 Image Captioning

Most recent image captioning models use RNNs (i.e. GRUs (Cho et al. 2014), LSTMs (Hochreiter and Schmidhuber 1997)) as caption decoders (Donahue et al. 2015; Karpathy and Fei-Fei 2015; Vinyals et al. 2015; Luo et al. 2018; Liu et al. 2018). The output words’ probabilities at each step are trained to maximize the human captions’ log-likelihood or some end evaluation metric (e.g. CIDEr) directly using REINFORCE.

However, less research provides hidden-state supervision. Professor forcing (Lamb et al. 2016) trains the hidden states of an RNN to be indistinguishable whether the network is trained with its inputs clamped to a training sequence or whether its inputs are self-generated. However, the professor hidden states do not necessarily carry richer information to guide the student hidden states. Our image caption framework uses richer hidden states from an easier task, which generates the human captions given both the visual inputs and all of its human captions, to advise the hidden states in the student caption decoders.

2.2 Autoencoders

Autoencoders (Hinton and Salakhutdinov 2006) learn to compress input data into dimension-reduced features using an encoder and learn a decoder to reconstruct the original input data from the the hidden state. For image captioning, the most relevant research (Yang et al. 2018) uses an autoencoder to construct scene graphs using relations between pairs of objects to represent the complex structure of both image and sentence before generating the caption. In contrast, we utilize autoencoders to learn richer teacher hidden states conditioned on both the image and human captions. Finally, we train the student caption decoder to produce similar hidden states without the human captions as input.

2.3 Teacher-Student Networks

The teacher-student method (Romero et al. 2014) transfers knowledge from a shallower and wider teacher CNN to a deeper and thinner student CNN by minimizing the divergence between the two output probabilities. Kim and Rush (2016) extend this method to train RNNs for machine translation. In our framework, instead of using the same training data, we allow the teacher network (an autoencoder) access to additional information (i.e. human captions), resulting in a richer representation to advise the student caption decoder.

3 Approach

This section presents the details of HSG. We first present the overall architecture in Sec. 3.1, then describe three student caption decoders in Sec. 3.2 to illustrate that HSG can be applied to almost any RNN-based decoder. After that, we explain the teacher autoencoder in Sec. 3.3 and the state transformation network to estimate the initial teacher hidden state only from the visual inputs in Sec. 3.4.

3.1 Overview

The goal of HSG is to provide hidden state guidance to any conventional RNN-based caption decoder, which we regard as a student network, as shown in Figure 2. In order to collect the guidance, we first train a teacher on an easier task that uses images to help autoencode human captions, which shares the same architecture as the student decoder. Then, we utilize a state transformation network to estimate the teacher decoder’s initial hidden states (t=0) using only the visual input. These approximations are used to initialize the student decoder’s hidden states so that it is capable of directly generating captions from images.

3.2 Student Caption Decoder

In this section, we briefly present three RNN-based student caption decoders.

FC. This model (Vinyals et al. 2015) adopts a single layer LSTM as the caption decoder. For the visual input features, we first feed the full image to a deep CNN (i.e. ResNet-101), and then average-pool the features from the final convolutional layer, resulting in a 2048-d vector for each image. The words are encoded using a trainable embedding matrix. At each time step, the LSTM receives the previous hidden states, generated words, and the visual features to generate the current word. Please refer to (Vinyals et al. 2015) for details.

Up-Down. This model (Anderson et al. 2018) incorporates object detection features into image captioning. The caption decoder operates on features of detected objects extracted using Faster RCNN (Ren et al. 2015) with a ResNet-101 (He et al. 2016) base network. It consists of a two-layer LSTM, where the first LSTM learns to distinguish important objects for generating the current word using an attention mechanism, and the second LSTM encodes the sequential information from the attended features to compute the output word probabilities.

SGAE. This model (Yang et al. 2018) incorporates scene graphs for both sentences and images into image captioning. Specifically, they first encode a sentence scene graph to reconstruct human captions and learn an additional memory matrix. Then, they use visual scene graph features with the learned memory to generate captions via a two-layer LSTM with attention similar to Up-Down.

3.3 Teacher Autoencoder

Our teacher autoencoder is trained to generate captions using not only visual input features, but also the set of human captions for the image. The teacher has two components: (1) a caption encoder to learn the joint representation of both the human captions and the visual inputs, and (2) a
decoder to generate captions from the output of the encoder and the visual features.

### Caption Encoder

Our caption encoder takes as input the image feature set \( \mathbf{V} = \{ \mathbf{v}_1, \ldots, \mathbf{v}_K \} \) consisting of \( K \) vectors for \( K \) detected objects, \( C \) human captions \( \mathbf{W}_c^i = \{ w_{c,1}^i, w_{c,2}^i, \ldots, w_{c,T}^i \} \), where \( T \) denotes the length of the captions and \( i = 1, \ldots, C \) are the caption indices.

Figure 3: Overview of the caption encoder. The Word LSTM is used to generate attention to identify the key words in each caption, and the Caption LSTM generates the final caption embedding. Blue arrows denote \( f \epsilon \) layers and yellow arrows attention embedding.

Inspired by (Wu, Hu, and Mooney 2019), we use a two-layer LSTM architecture to encode human captions as illustrated in Figure 3. The first-layer LSTM (called the Word LSTM) sequentially encodes the words in a caption \( \mathbf{W}_c^i \) at each time step as \( h_{i,t} \).

\[
\begin{align*}
    h_{i,t}^{c,1} & = \text{LSTM}( \mathbf{W}_c \Pi_{i,t}^{c}, h_{i,t-1}^{c,1}, c_{i,t-1}^{c,1} ) \\
    c_{i,t}^{c,1} & = h_{i,t-1}^{c,1}
\end{align*}
\]  

(1)

where \( \mathbf{W}_c \) is the 300-d word embedding matrix, and \( \Pi_{i,t}^{c} \) is the one-hot embedding for the word \( w_{c,t}^i \).

Then, we design a caption attention module \( \mathbf{A}^c \) which utilizes the image feature set \( \mathbf{V} \), and \( h_{i,t}^{c,1} \) to generate the attention weight on the current word in order to indicate its importance. Specifically, the Word LSTM first encodes the word embedding \( \Pi_{i,t}^{c} \) in Eq. 1. Then we feed the outputs \( h_{i,t}^{c,1} \) and \( \mathbf{V} \) to the attention module \( \mathbf{A}^c \) as shown in Eq. 2.

\[
\alpha_{i,j,t} = \text{softmax}( h_{i,t}^{c,1} \circ \mathcal{f}( \mathbf{v}_j ))
\]  

(2)

where the softmax function is over the \( K \) objects in visual feature set \( \mathbf{V} \).

Next, the attended word representations \( w_{c} \) in the caption are used to produce the final caption representation in Eq. 4 via the Caption LSTM.

\[
\begin{align*}
    w_{c} & = \max_j \{ \alpha_{i,j,t} \} \mathbf{W}_c \Pi_{i,t}^{c} \\
    h_{i,t}^{c,2} & = \text{LSTM}( w_{c}, h_{i,t-1}^{c,2}, c_{i,t-1}^{c,2} )
\end{align*}
\]  

(3)

(4)

where \( \max \) denotes the element-wise max pooling over the attention weights for the \( K \) objects in the image.

### Caption Decoder

Since the goal of the teacher caption decoder is to provide hidden state guidance to the student decoder, we require
these two decoders to have the same architecture. However, the differences between these two decoders lie in the initial hidden states. The teacher decoder is initialized with the outputs from the encoder while the student caption decoder is initialized with an estimated version as detailed in Sec. 3.4.

For the FC caption decoder, we use the max pooling of the final hidden state from the second LSTM in the caption encoder as the initial state, i.e., $\tilde{h}_0^d = \max_c \{ h_{c,T}^d \}$. Similarly, we max pool the final hidden states from both layers to initialize the hidden state for the two LSTMs in the Up-Down and SGAE caption decoders. Specifically, the teacher initial hidden state is computed as $\tilde{h}_0^d = [\max_c \{ h_{c,T}^{d,1} \}; \max_c \{ h_{c,T}^{d,2} \}]$. After initialization of the initial states $\tilde{h}_0^d$, the student decoder is trained to match the teacher states $\tilde{h}_t^d$ for each time step $t$.

### 3.4 State Transformation Network

The state transformation network uses the visual features to estimate the initial teacher hidden states $\tilde{h}_0^d$ so that the student caption decoder is capable of using estimated hidden states to start a sentence purely from the visual inputs alone. For efficiency, we simply use a two-layer $fc$ network for state transformation. For the FC decoder, we directly apply the two-layer networks to the visual feature vector to estimate the initial hidden states $h_0^d = f(f(v))$, where the $v$ is the $2,048$-d features for the image. For the Up-Down and the SGAE decoder, the first $fc$ layer adapts the visual features for each object and the second $fc$ layer estimates the hidden states using the sum of the first layer’s output over all objects or nodes in the scene graph. Specifically the hidden states are computed as $h_0^d = f(\sum_{j=1}^{K} f(v_j))$.

### 4 Training

Training involves pre-training the teacher autoencoder and then training the student caption generator using maximum likelihood or REINFORCE. We use $\theta_a$ to denote the parameters in the autoencoder (i.e., the caption grounding encoder and the teacher caption decoder), and $\theta_g$ to denote the parameters in the state transformation network and the student decoder. We use $c$ to denote the entire caption, $c_t$ to denote the $t$-th word in the caption, and $c_{\leq t}$ to denote the first $t$ words in the caption. We omit the visual features $v$ in all of probabilities in this section for simplicity. We denote the maximum likelihood loss using parameters $\theta$ as $L_{ll}(\theta)$ defined in Eq. 5:

$$L_{ll}(\theta) = - \sum_{t=1}^{T} \log(p(c_{\leq t-1}; \theta))$$  

(5)

#### 4.1 Pretraining the Teacher Autoencoder

We pre-train the teacher autoencoder using cross-entropy loss, minimizing $L_{ll}(\theta_a)$ defined in Eq. 5. After pre-training, the parameters $\theta_a$ are fixed.

Additionally, we pre-train the state transformation network using $L_{s,t}(\theta_g)$, $t = 0$ as defined in Eq. 6 using L2 distance.

$$L_{s,t}(\theta_g) = \|h_t^d - \tilde{h}_t^d\|_2^2$$  

(6)

In particular, the generated captions from the student decoder are fed to the teacher autoencoder to compute the teacher hidden states at each time ($t$) as shown in Fig 2. We will omit “$(\theta_g)$” from $L_{s,t}(\theta_g)$ for simplicity.

### 4.2 Training the Student Decoder

We tested two different approaches to training the student decoders using either maximum likelihood or REINFORCE (with various evaluation metrics as rewards). The student decoder is initialized with the teacher decoder’s parameters.

#### Maximum Likelihood Training

Maximum likelihood trains the student decoder to maximize the word-level log-likelihood, where human captions are fed into the decoder to compute the next word’s probability distribution. We use the joint loss in Eq. 7:

$$L = L_{ll}(\theta_g) + \lambda \sum_{t=0}^{T} L_{s,t}$$  

(7)

With human captions as input to the teacher autoencoder, we compute its hidden state, which is needed to calculate $L_{s,t}$. The $\lambda$ parameter controls the weights of the state loss.

#### REINFORCE

An alternative to log-likelihood maximization is to fine-tune the model to directly maximize the expected evaluation metric using REINFORCE. Negative rewards, such as BLEU (Papineni et al. 2002) or CIDER (Vedantam, Lawrence Zitnick, and Parikh 2015), are minimized using $L = -\mathbb{E}_{c \sim p_g}[\hat{r}(\hat{c})]$ where $\hat{r}(\hat{c}) = r(\hat{c}) - r(c^*)$ denotes the variance-reduced rewards (Rennie et al. 2017), $\hat{c}$ denotes the sampled captions using the probabilities over the vocabulary, and $c^*$ denotes greedily sampled captions using the word with the maximum probability. We will omit “$(\theta_g)$” from $p_{\theta_g}$ for simplicity. The parameters in the student caption decoder are updated using the policy gradients

$$\nabla_{\theta_g} L = -\mathbb{E}_{c \sim p}[\hat{r}(\hat{c})\nabla_{\theta_g} \log p(\hat{c})].$$

However, one remaining problem with this approach is that the sentence-level reward $\hat{r}(\hat{c})$ is equally distributed over each word in the sampled captions, no matter how relevant the word is. Therefore, some desired words will not get enough credit because of the presence of some unrelated or inaccurate words in the sentence. To address this issue, we propose to use our hidden state loss as an intermediate reward to encourage the student decoder to produce hidden states that match the hidden states of the high-performing teacher decoder. We add a reward objective $\check{R}$ that is the accumulated expectation of the negative hidden state losses over time ($t$) as shown in Eq. 8:

$$\check{R} = -\sum_{t=0}^{T} \mathbb{E}_{c_{\leq t} \sim p}[L_{s,t}]$$  

(8)

Therefore, the new policy gradients can be written as Eq. 9, and we provide detailed derivation in the supplementary
Table 1: Automatic evaluation comparisons with various baseline caption decoders on the Karpathy test set. “HSG” denotes trained with hidden state guidance, B-4, M, R-L, C and S are short hands for BLEU-4, METEOR, ROUGE-L, CIDEr and SPICE. All captions are generated with beam search (beam size=5) except for SGAE decoders that use beam size = 2.

| Model                  | B-4 | M   | R-L | C   | S   | B-4 | M   | R-L | C   | S   |
|------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| LSTM-A (Yao et al. 2017) | 35.2 | 26.9 | 55.8 | 108.8 | 20.0 | 35.5 | 27.3 | 56.8 | 118.3 | 20.8 |
| StackCap (Gu et al. 2018) | 35.2 | 26.5 | -   | 109.1 | -   | 36.1 | 27.4 | 56.9 | 120.4 | 20.9 |
| FC (Vinyals et al. 2015) | 32.9 | 25.0 | 54.0 | 95.4  | 17.9 | 32.8 | 25.0 | 54.2 | 104.0 | 18.5 |
| FC + HSG                       | 33.2 | 25.5 | 53.9 | 96.1  | 18.3 | 33.9 | 25.9 | 54.8 | 107.5 | 18.4 |
| Up-Down (Anderson et al. 2018) | 36.0 | 27.0 | 56.3 | 113.1 | 20.4 | 36.3 | 27.5 | 56.8 | 120.7 | 21.4 |
| Up-Down + HSG                   | 35.6 | 27.3 | 56.7 | 113.9 | 20.6 | 37.4 | 28.0 | 57.7 | 124.0 | 21.5 |
| SGAE (Yang et al. 2018)        | 35.8 | 28.0 | 56.7 | 114.0 | 20.9 | 37.8 | 28.2 | 58.2 | 125.9 | 22.1 |
| SGAE + HSG                      | 35.9 | 28.1 | 56.9 | 115.2 | 21.0 | 38.5 | 28.4 | 58.5 | 127.7 | 21.8 |

5 Experimental Evaluation

In this section, we verify the effectiveness of HSG using both standard automatic metrics and human evaluation.

---

3We use a smaller batch size than the original implementation, leading to the performance drop (16 vs 100)
4https://github.com/ruotianluo/self-critical.pytorch
5https://github.com/yangxuntu/SGAE
5.1 Dataset
We use the MSCOCO 2015 dataset (Chen et al. 2015) for image captioning. In particular, we use the Karpathy configuration that includes 110K images for training and 5K images each for validation and test. Each image has 5 human caption annotations. Similar to Anderson et al. (2018), we first convert all sentences to lower case, tokenized on white spaces, and filter words that occur less than 5 times.

5.2 Results on Image Captioning

Comparison with the Base Decoders. In Table 1, we present the standard automatic evaluation for FC, Up-Down and SGAE decoders trained using either Maximum Likelihood alone or using REINFORCE with CIDEr rewards and compare them with the state-of-the-art image captioner. Metrics included are BLEU-4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE-L (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016). The left part reports the results trained with maximum likelihood and the right part shows the results using REINFORCE. We also include an ensemble model, where each member is trained using REINFORCE with a different random seed. We observe that when training with maximum likelihood, our model outperforms both of the baseline decoders for all of the metrics, except ROUGE-L when using FC, and BLEU-4 when using Up-Down, demonstrating the effectiveness of introducing our hidden state guidance. More importantly, we observe a significant improvement on the CIDEr scores over all of the baseline models using REINFORCE (i.e. 107.5 v.s. 104.0 using FC Model, 124.0 v.s. 120.7 using FC Model and 127.7 v.s. 125.9 using SGAE Model). We attribute the improvements to both HSG and the word-level intermediate rewards.

Teacher Autoencoder Performance In this section, we report automatic evaluation results for FC, Up-Down and SGAE teacher autoencoders. As shown in Table 3, all of the

<table>
<thead>
<tr>
<th>Model</th>
<th>B-4</th>
<th>M</th>
<th>R-L</th>
<th>C</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>REINFORCE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REINFORCE + HSG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up-Down (B-4)</td>
<td>37.5</td>
<td>26.9</td>
<td>57.0</td>
<td>111.0</td>
<td>20.3</td>
</tr>
<tr>
<td>Up-Down (M)</td>
<td>31.4</td>
<td>28.7</td>
<td>56.1</td>
<td>105.0</td>
<td>22.0</td>
</tr>
<tr>
<td>Up-Down (R-L)</td>
<td>36.5</td>
<td>26.6</td>
<td>57.9</td>
<td>114.2</td>
<td>19.8</td>
</tr>
<tr>
<td>Up-Down (C)</td>
<td>36.3</td>
<td>27.5</td>
<td>56.8</td>
<td>120.7</td>
<td>21.4</td>
</tr>
</tbody>
</table>

Table 2: Evaluation on Up-Down decoder’s performance on Karpathy test set with various evaluation metrics as rewards when training using REINFORCE. B-4, M, R-L, C and S are short hands for BLEU-4, METEOR, ROUGE-L, CIDEr and SPICE.

Ablation Study on Hidden State Weight Table 4 shows results when we vary the hidden state loss weight \( \lambda \) from 0.0 to 2.0 using Up-Down decoders. Our model consistently outperforms the baseline model (\( \lambda = 0 \)) on all evaluation metrics when using CIDEr rewards. We also observe that the CIDEr scores are fairly robust to the exact setting of \( \lambda \).

Human Evaluation We conducted an Amazon Mechanical Turk (AMT) evaluation by asking human judges to directly compare captions from HSG and the baseline models. Following (Wu and Mooney 2018; Park et al. 2018; Venugopalan et al. 2017), we use a ranking-based approach with majority voting. In particular, we randomly chose 1,000 images and generated captions with and without HSG trained with CIDEr rewards using Up-Down with a beam size of 5. The two captions for each image were randomly
ordered and we asked workers to rank them in terms of descriptiveness, allowing for ties. We test each image with two captions with 3 different judges and the final rankings are determined as follows. For each human judgement, we assign +1 to the better caption, 0 if they are tied and −1 to the worse one. Then, we compute the average scores for the two captions to decide their final rankings. In Table 5, we report the percentage of our captions that are better than, equivalent to, or worse than the captions without HSG. We observe that our captions are better more than 50% of the time and worse less than 30%, indicating that HSG is effective at improving captioning.

| Table 5: Percentage of our captions that are better than, equivalent to, or worse than those without HSG. |
|---|---|---|---|
| Up-Down | Better | Equivalent | Worse |
| | 54.3% | 16.1% | 29.6% |

Table 6 reports the performance of Up-Down decoder using maximum likelihood and REINFORCE with CIDEr rewards. We observe obvious improvements over Professor Forcing in both the maximum likelihood and REINFORCE cases. Unlike Professor Forcing, our teacher has access to more information (i.e., the initial hidden state that encodes the set of human captions) than the student, and therefore possesses a richer hidden state representation to guide the student.

| Table 6: Comparison with Professor Forcing using Up-Down decoders on Karpathy test set. |
|---|---|---|---|---|
| Maximum Likelihood | B | M | R-L | C | S |
| Professor HSG | 35.8 | 27.0 | 56.1 | 113.3 | 20.3 |
| REINFORCE (CIDEr) | 35.6 | 27.3 | 56.7 | 113.9 | 20.6 |

6 Conclusion and Future Work

We have presented a novel image captioning framework that uses an image-conditioned caption autoencoder. This teacher autoencoder, which has the same architecture as the original RNN-based decoder, is trained using an easier task that learns to generate image captions with gold standard human captions as inputs in addition to visual features. This autoencoder is used to guide the hidden state representation learned by the student caption decoder. We integrated this hidden state guidance into both maximum likelihood and REINFORCE using three state-of-the-art image captioners. We observe that especially in the REINFORCE case, the word-level hidden state guidance assigns an intermediate reward that emphasizes the most relevant words. Extensive experimental results demonstrate the effectiveness of our approach. In the future, we would like to explore metrics that can measure and minimize the semantic difference between the teacher and student hidden states better than L2, this could potentially further improve the learned hidden-state representation.

