Transfer Reward Learning for Policy Gradient-Based Text Generation

James O' Neill and **Danushka Bollegala**
Department of Computer Science, University of Liverpool, UK
{james.o − neill, danushka.bollegala}@liverpool.ac.uk

Abstract

Task-specific scores are often used to optimize for and evaluate the performance of conditional text generation systems. However, such scores are non-differentiable and cannot be used in the standard supervised learning paradigm. Hence, policy gradient methods are used since the gradient can be computed without requiring a differentiable objective. However, we argue that current n-gram overlap based measures that are used as rewards can be improved by using model-based rewards transferred from tasks that directly compare the similarity of sentence pairs. These reward models either output a score of sentence-level syntactic and semantic similarity between entire predicted and target sentences as the expected return, or for intermediate phrases as segmented accumulative rewards.

We demonstrate that using a *Transferable Reward Learner* leads to improved results on semantical evaluation measures in policy-gradient models for image captioning tasks. Our InferSent actor-critic model improves over a BLEU trained actor-critic model on MSCOCO when evaluated on a Word Mover's Distance similarity measure by 6.97 points, also improving on a Sliding Window Cosine Similarity measure by 10.48 points. Similar performance improvements are also obtained on the smaller Flickr-30k dataset, demonstrating the general applicability of the proposed transfer learning method.

Introduction

Neural network based encoder-decoder architectures are increasingly being used for conditional text generation given the recent advances in Convolutional Neural Networks(CNNs) (Szegedy et al. 2017) and Recurrent Neural Networks (RNN) (Hochreiter and Schmidhuber 1997) that use internal gating mechanisms to preserve long-term dependencies (Hochreiter and Schmidhuber 1997), showing impressive results for density estimation (i.e language modelling) and text generation. These encoder-decoder networks are usually trained end-to-end using Maximum Likelihood (ML) training with full supervision (i.e the model learns explicitly from expert demonstrations). This is also referred to as *teacher forcing* (Sutton et al. 2000).

Typically for text generation tasks, such as image captioning, CNNs (LeCun, Bengio, and others 1995) encode an image, which is passed to the beginning of the decoder RNN language model via a linear map from the image encoding. The decoder policy $\hat{\pi}$ then performs a set of actions given an *expert* policy π^* for T time steps (e.g human captions).

However, ML can act as a poor surrogate loss for a task-specific score we are interested in and evaluate on (e.g BLEU). Moreover, these scores are non-differentiable and hence cannot be used in the standard supervised learning paradigm.

Deep Reinforcement Learning (DRL) can be used to optimize for task scores as rewards (Bahdanau et al. 2016; Zhang et al. 2017), for the purposes of higher quality text generation e.g n-gram overlap based metrics such as ROUGE-L and BLEU.

Actor-critic networks in particular have shown State of The Art (SoTA) results for image captioning (Bahdanau et al. 2016; Zhang et al. 2017). These models use a policy network that produces actions which are evaluated by a value network that output scores given both actions and targets as input for each state. The critic value prediction is then used to train the actor network, assuming the critic values are exact (pre-training the critic is often necessary). However, parameter-free measures such as BLEU and ROUGE-L do not correlate well with human judgements when compared to using word and sentence-based embedding semantic similarity evaluation measures between predicted and target sequences, moreover the frequency distribution of generated captions is significantly different to human captions (Cui et al. 2018).

We argue that n-gram overlap based measures in DRL models can be improved using model-based reward estimators that are transferred from sentence similarity models that directly learn from the manually annotated similarity of paired sentences.

In the context of DRL, we view this type of transfer learning as generating the environment of a target task given a source model that implicitly learns relationships on a pairwise source task (i.e sentence similarity). Additionally, the reward is continuous everywhere and out-of-vocabulary terms do not hinder the TRL model's ability to estimate return since sentence similarity can still be inferred even when ⟨unk⟩ tags are used at test time to replace tokens we have not seen at training time. We are further motivated by the fact that transfer learning for text generation has already been successfully demonstrated using pretrained ImageNet CNN encoders (Krizhevsky, Sutskever, and Hinton 2012). However, transfer learning with respect to the decoder of DRL-based encoder-decoder models has been unexplored until this point.

We also note that for the common use of DRL in games and robotics, transfer learning is often made difficult since the environments and dynamics are often distinctly different from one another (e.g games usually do not have the same states, actions, transition probabilities and rewards i.e Markov Decision Process (MDP)). In contrast, the MDP for natural language is defined by the vocabulary used for a given corpus. Thus, given a sufficient amount of text the MDP for all corpora converge. Hence, transfer learning becomes easier which is not typical for robotics and games.

Contributions This paper proposes to transfer pairwise models that have been trained to learn a similarity score between various universal phrase and sentence representations. These models are trained on a set of sentence-pairwise learning problems such as sentence similarity and natural language inference (NLI).

Herein, we refer to this as *Transferable Reward Learning* (TRL), a method that incorporates model-based reward shaping to improve task-specific scores in relation to semantic similarity as a measure of language generation quality. We baseline both unsupervised and supervised TRL models against ML training and previous actor-critic models with model-free rewards such as BLEU and ROUGE-L. To our knowledge, this is the first work that focuses on learning to transfer model-based reward in sequence prediction.

Related Work

DRL-based Conditional Text Generation

Zhang et al. have previously used actor-critic sequence training for image captioning using a token-level advantage and value function, achieving SoTA on MSCOCO at that time. In contrast, TRL can evaluate return both on token and sentence level learned from human judgment similarities between sentences.

Ranzato et al. proposed Mixed Incremental Cross-Entropy Reinforce (MIXER) which uses REINFORCE (Williams 1992) for text generation with a mixed expert policy (originally inspired by DAgger (Ross and Bagnell 2014)), whereby incremental learning is used with REINFORCE and cross-entropy (CE). During training, the policy gradually deviates from π^* provided using CE, to using its own past predictions.

Rennie et al. propose Self-Critical Sequence Training (SCTS) which extends REINFORCE by using the test-time model predictions to normalize the reward. This avoids the use a baseline to normalize the rewards and reduce variance, while mitigating exposure bias.

Reward Augmented ML (RAML) (Norouzi et al. 2016) combine conditional log-likelihood and reward objectives while showing that highest reward is found when the conditional is proportional to the normalized exponentiated rewards, referred to as the payoff distribution.

Ren et al. have defined the reward as the embedding similarity between both images and sentences that are projected into the same embeddding space, instead of similarity of embeddings corresponding to predicted and target tokens alone. To the best of our knowledge, it is the only other method that uses a continuous reward signal from an embedded space. We can consider this to be an embedding measure that is model-free. In the context of this work, we also consider a model-free sentence embedding similarity measure (see section) in our AC model.

Sentence Representations

Given that our main contribution is the adaptation of TRL pairwise models to model rewards (i.e sentence-level similarity between predicted and target caption), we briefly introduce the relevant SoTA that we consider in our experiments. Kiros et al. presented Skipthought vectors which are formed using either a bidirectional or unidirectional Long Short Term Memory (LSTM) encoder-decoder that learns to predict adjacent sentences from encodings of the current sentence, which is a fully unsupervised approach.

In contrast to Skipthought, Conneau et al. (2017) propose InferSent which is a supervised method to learn sentence representations from natural language inference data (Bowman et al. 2015). InfeSent outperforms unsupervised sentence representations such as Skipthought on various sentence-pair tasks. We too include this approach in our experiments as the second TRL model to estimate accumulative rewards. O' Neill and Bollegala (2018) proposed contextualized embeddings by learning to reconstruct a weighted combination of multiple pretrained word-embeddings as an auxiliary task, acting as a regularizer for the main task.

Learning Reward Functions

Apprenticeship learning (Abbeel and Ng 2004) has also focused on learning the reward function from expert demonstrations, albeit in the context of robotics. Christiano et al. proposed to learn to play Atari using deep RL from human preferences. This is analgous to learning from similarity scores that are used as labels for pairwise learning between sentence representations. We share a similar motivation in that learning from demonstrations (i.e reference captions) can be difficult, particularly when there is many permutable demonstrations that lead to a similar goal (i.e many semantically and grammatically correct ways to describe what is in an image). By learning a pairwise-model learned from human preference scores (e.g sentence similarity scores) between trajectories, we can model the reward.

Methodology

Image Caption Setup

For an image I there is a corresponding caption sequence Y that contains tokens $Y = (y_1, .., y_T)$ and $y \in \mathcal{V}$ where $\mathcal{V}$ is the vocabulary. f_ω encodes an image I into a hidden state z that is then passed to a Recurrent Neural Network (RNN) decoder f_ψ that generates a predicted sequence $\hat{Y}$ which is then evaluated with a task-specific score $R(Y, \hat{Y})$. In the DRL setting, we can consider the problem as a finite

Markov Decision Process (MDP) where each word $w \in \mathcal{V}$ is considered a state $s \in S$ and a prediction $\hat{y}_t$ is considered as an action $\pi_\theta(a_t|s_t)$ in action space $a \in \mathcal{A}$ with probability $p_{\pi_\theta}(a_t|s_t)$. The environment then issues a discounted return $g = \sum_{t \in T} \gamma_t r_t$ where the discount factor $\gamma \in [0, 1]$, after receiving the set of actions and the objective is then to maximize the total expected return G. We use an actor-critic model (Barto, Sutton, and Anderson 1983) as the basis of our experiments with a ResNet-152 encoder (He et al. 2016) and an LSTM network.

Policy Gradient Training

We define a policy network π_θ as an encoder-decoder architecture that encodes an image $I_s \in \mathbb{R}^{m \times n}$ as $h_s \in \mathbb{R}^n$ through the Resnet-152 (He et al. 2016) CNN-based encoder and a linear projection $W_s \in \mathbb{R}^{d \times n}$ shown in Equation 1. This is then concatenated with the embedding $x_t \in \mathbb{R}^m$ corresponding to the input word $w_t \in \mathbb{Z}$, which forms state $s_t = x_t \oplus h_s$ where $\oplus$ denotes concatenation. This LSTM decoder takes (s_t, h_{t-1}) as input, as shown in Equation 2, omitting $t = 0$ where h_0 is used instead. Therefore the policy network parameters include the ResNet-152 parameters ω, the linear projection W_s, the LSTM parameters ψ and decoder projection layer W_t, hence $\theta := \{\omega, W_s, \psi, W_t\}$. The predictions for a given sequence length of $T = |Y|$, predictions are defined as $\hat{Y} = \{a_1, ..a_t, ..a_T\}$ where the action space $a_t \in \mathcal{A}$ is defined by the vocabulary $w \in \mathcal{V}$ and the targets $Y = \{w_1, ., w_t, ...w_T\}$.

$$h_s = W_s \text{ResNet-152}(I_s) \tag{1}$$

$$h_t = \text{LSTM}(s_t, h_{t-1}) \tag{2}$$

$$p_{\pi_\theta}(s_t) = \phi(W_t \cdot h_t) \tag{3}$$

$$\pi_\theta(a_t|s_t) = p_{\pi_\theta}(a_t|s_t) \tag{4}$$

Value Function Approximation For Value Function Approximation (VFA), the gradient of the expected accumulative discounted reward is typically estimated as $\mathbb{E}[\sum_{t=0}^{T} \gamma_t r_t | a_{t+1}, ..a_T]$, for reward r_t at time t for an l-step return.

We then use a critic network to estimate the state-value function that takes the policy π, actions $a_{t:t+l} \in \hat{Y}$, parameterized rewards $r^{\vartheta}_{t+1:t+l}$ and compute the expected return $V^\pi(s_t)$ from state s_t for l steps. This is given as the expectation over the sum of discounted rewards expressed as Equation 5 where rewards r^ϑ are issued by our proposed TRL model-based reward with frozen parameters ϑ and γ not used as discounted rewards are not applicable in the TRL.

$$V^\pi(s_t) = \mathbb{E}\Big[\sum_{t=0}^{l} r^{\vartheta}_{t+1} | a_{t+1}, ..a_l\Big] \tag{5}$$

Advantage Function Approximation Above, we considered using $V^\pi(s_t)$ to estimate g. However, training the value network from scratch can result in high-variance in the gradient and therefore poor convergence. Similarly to Zhang et al., we use the Advantage Function Approximator (APA) A^π to reduce the variance in gradient updates. This is achieved using temporal-difference λ (TD-λ) learning as shown in Equation 6 where the Q-function is $Q^\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}:T, a_{t+1}:T}\Big[\sum_{i=0}^{T} r_{t+l}\Big]$, for N step expected return G^i_t and λ is the trace decay weight $0 \leq \lambda \leq 1$ (larger λ assigns more credit assigned to distant rewards).

$$\begin{aligned} A^\pi(s_t, a_{t+1}) &= Q^\pi(s_t, a_{t+1}) - V^\pi(s_t) \\ &= (1 - \lambda) \sum_{i=1}^{N} G^i_t - V^\pi(s_t) \end{aligned} \tag{6}$$

The gradient of the policy network can then be rewritten as Equation 7. Here, the trace decay parameter $0 < \lambda < 1$, in our experiments $\lambda = 1$ which corresponds to Monte-Carlo and means that large traces are also assigned to distant states and actions. In the context of image captioning, it is typical that the episodes are short ($T < 30$) and hence it is feasible.

$$\begin{aligned} G = \mathbb{E}\Big[&\sum_{t=0}^{T-1} \big((1 - \lambda) \sum_{i=0}^{N} G^m_t \\ &- V^\pi(s_t) \nabla_\theta \log \pi_\theta(a_{t+1}|s_t) \Big] \end{aligned} \tag{7}$$

This is achieved by computing the gradient of the log likelihood multiplied by the advantage function $A^\pi(s_t, a_{t+1})$ shown in Equation 8. Here, $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$ reduces the variance of the of the gradient by increasing the probability of actions when $A^\pi(s_t, a_{t+1}) > 0$ and decrease otherwise.

$$G := \mathbb{E}[\sum_{t=0}^{T-1} A^\pi(s_t, a_{t+1}) \nabla_\theta \log \pi_\theta(a_{t+1}|s_t)] \tag{8}$$

Transfer Reward Learner

We now consider two sentence encoders, as mentioned in section , as the TRL. We note that, although there has been considerable breakthroughs in recent years for models that could be used for sentence similarity tasks (Devlin et al. 2018), these models are too computationally costly to consider for issuing rewards and typically have more parameters than the whole actor-critic network combined.

Both TRL models that evaluate state-action pairs are denoted as $R_\vartheta(s, a)$ where the ϑ parameters are not-updated as rewards are kept static throughout training. The advantage of this is that we are not restricted to choosing $\lambda = 1$, which is used for the sentence-level n-gram overlap measures such as BLEU, ROUGE and CIDEr. The critic can evaluate partially generated sequences and sentence pairs of different length, since they have been trained to learn similarity between sentences of non-equal length. We emphasize at this point that the TRL is not updated for value function estimation in our experiments, this is only carried out for approximating the advantage and value functions.

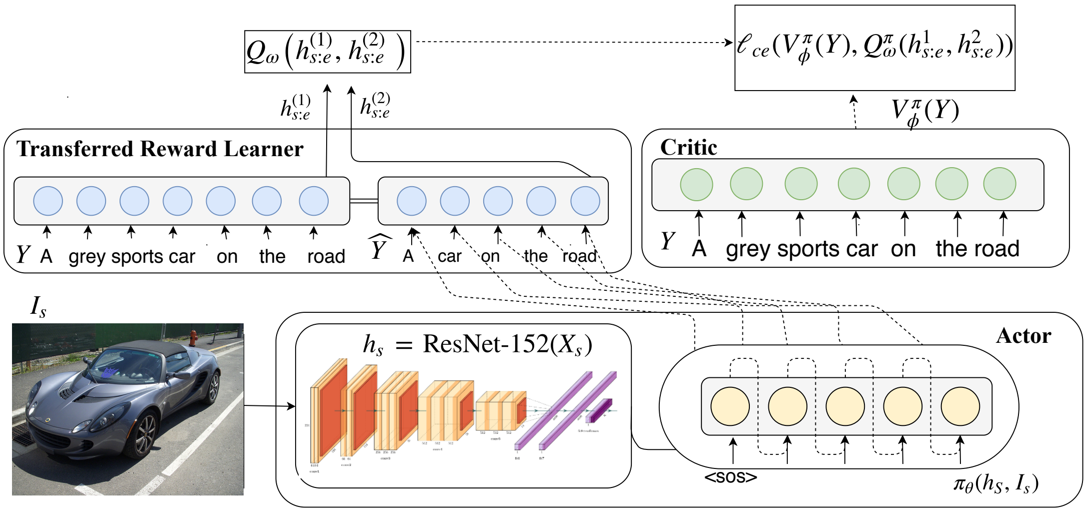

$Q_\omega\left(h_{s:e}^{(1)}, h_{s:e}^{(2)}\right)$
$\ell_{ce}(V_\phi^\pi(Y), Q_\omega^\pi(h_{s:e}^1, h_{s:e}^2))$
$h_{s:e}^{(1)}$
$h_{s:e}^{(2)}$
$V_\phi^\pi(Y)$
Transferred Reward Learner
Y A grey sports car on the road
$\widehat{Y}$ A car on the road
Critic
Y A grey sports car on the road
I_s
h_s = ResNet-152(X_s)
Actor
<SOS>
$\pi_\theta(h_S, I_s)$

Figure 1: Actor $\pi_\theta(\cdot)$ produces actions $\hat{Y}$ given an encoded image h_s and caption X_t that are passed to TRL$(Y, \hat{Y})$ and encoded into (h_1, h_2) respectively and scored with action-value $A^\pi = V_\Theta^\pi(s) - Q^\pi(s,a)$

InferSent Rewards For reward shaping, we use a pretrained sentence similarity model such as InferSent, tuned on SemEval 2017 Semantic Textual Similarity (STS) dataset[1] consisting of English monolingual pairs that are labelled with a score from 0 (semantic independence) to 5 (semantic equivalence). The scores are scaled from the continuous [0-5] range to [0-1] using a sigmoid $\sigma(\cdot)$ to convert to a probability.

Conneau et al. (2017) use the scoring function in Equation 9 between two encoded sentence pairs (h_1, h_2) where $h_1, h_2 \in \mathbb{R}^d$, corresponding to two sentences $(\mathcal{S}_1, \mathcal{S}_2)$. We also use this scoring function for the pretrained InferSent model.

$$\phi\big([h_1, h_2, |h_1 - h_2|, h_1 \cdot h_2] \cdot W + b\big) \tag{9}$$

The model is a Bidirectional-GRU (or BiLSTM) with max (or mean) pooling, as in Equation 10 where g represents the pooling function and e_i is the embedding corresponding to word x_i.

$$h = g_{\text{max-pool}}\big([\overrightarrow{\text{GRU}}(e_1, .., e_T), \overleftarrow{\text{GRU}}(e_1, .., e_T)]\big) \tag{10}$$

We also use the self-attentive variation in Equation 11, where the max-pooling operation g is replaced with self-attention that produces a weighted average where $g_{\text{avg}}(\cdot)$ sum the weights to 1 $\forall t \in T$. Hence, attention focuses on the hidden states of important tokens prior to using the scoring function.

$$h = \sum_{t=1}^{T} g_{\text{avg}}(\tanh(W h_t + b)) h_t \tag{11}$$

Skipthought Rewards We also consider the Skipthought vectors as unsupervised sentence representations, which have shown competitive performance on sentence-level pairwise tasks (Kiros et al. 2015). This allows us to compare against the supervised InferSent model. Similarly to the InferSent model, we use the same scoring function as in Equation 9.

Critic Loss Prior work in this area used an ℓ_2 loss between $\pi_\theta(a|s)$ policy values and the critic scores $Q^\pi(s,a)$ (Zhang et al. 2017). In our experiments, we found a KL-divergence loss outperformed an ℓ_2 loss.

We minimize the KL divergence between the $[0,1]$ normalized $\tilde{Q}^\pi(s_t, a_{t+1})$ and $\tilde{V}_\Theta^\pi(s_t)$ which corresponds to minimizing the cross-entropy loss when ignoring $\mathcal{H}\big(\tilde{Q}^\pi(s_t, a_{t+1})\big)$ that does not rely on Θ. The logit penalizes values that are much higher than the baseline more than the ℓ_2 loss but show a larger gap between small value improvements over the baseline.

$$\begin{aligned}\ell_{ce} = \mathcal{D}_{\text{KL}}\Big(\tilde{Q}^\pi(s_t, a_{t+1}) || \tilde{V}_\Theta^\pi(s_t)\Big) \\ + \mathcal{H}\big(\tilde{Q}^\pi(s_t, a_{t+1})\big)\end{aligned} \tag{12}$$

This has the effect of stabilizing the critic network over consecutive iterations, as the critics gradient updates are not as large, stabilizing the training of the critic and subsequently ensuring the difference in the policy network $\pi_\theta(a_t|s_t)$ is less drastic between iterations.

Experimental Setup

Dataset Details

We use the Microsoft *Common Objects in Context* (MSCOCO) 2014 image captioning dataset proposed by Lin et al., which is a de-facto benchmark for image captioning.

[1] http://alt.qcri.org/semeval2017/task1/

Methods	B1	B2	B3	B4	PPL
NeuralTalk Karpathy and Fei-Fei	0.57	0.37	0.24	0.16	-
Minds Eye Chen and Lawrence Zitnick	-	-	-	0.13	19.10
NIC Vinyals et al.	0.66	-	-	-	-
LRCN Donahue et al.	0.59	0.39	0.25	0.17	-
m-Rnn-AlexNet Mao et al.	0.54	0.36	0.23	0.15	35.11
m-Rnn-VggNet Mao et al.	0.60	0.41	0.28	0.19	20.72
Hard-Attention Xu et al.	0.67	0.44	0.30	0.20	-
Liu et al.					
Implicit-Attention	-	-	0.29	0.19	-
Explicit-Attention	-	-	0.29	0.19	-
Strong Sup	-	-	30.2 21.0	0.19	-
Wu et al.					
Att-GT+LSTMz	0.78	0.57	0.42	0.30	14.88
Att-SVM+LSTM	0.68	0.49	0.33	0.23	16.01
Att-GlobalCNN+LSTM	0.70	0.50	0.35	0.27	16.00
Att-RegionCNN+LSTM	0.73	0.55	0.40	0.28	15.96

Table 1: SoTA Methods for Flickr30k

The dataset includes 164,062 images (82,783 training images, 40,504 validation images, and 40,775 test images) with 5 manually labelled captions per image of 80 object categories and 91 *stuff* categories. Each image is paired with as least five manually annotated captions.

We also use the smaller Flickr-30k (Young et al. 2014) dataset, which contains 30k images with 150k corresponding captions, which also includes a constructed denotation graph that can be used to define denotational similarity, giving more generic descriptions through lexical and syntactic operations.

Training Details

As mentioned before, we use the ResNet-152 (He et al. 2016) classifier trained on ImageNet as our encoder. The reported experimental results are that of a 2-hidden layer LSTM decoder (Hochreiter and Schmidhuber 1997) network, with embedding input size and hidden layer size of $|e| = |h| = 512$. For both MSCOCO and Flickr30k a mini-batch size of $|x_{sub}| = 80$ and use adaptive momentum (`adam`) (Kingma and Ba 2014) for stochastic gradient descent (SGD) optimization for the LSTM decoder while, as mentioned, the image encoder is not updated for our experiments.

Training both actor and critic networks from scratch is difficult, or more generally for related policy gradient algorithms since it is often the case the reward signal leads to high variance in the gradient updates, particularly in the early phase of training where the parameters θ and ϕ are initialized randomly.

Therefore, in all our experiments we pre-train the actor and critic networks (similarly to Ren et al.) for 5 and 7 epochs respectively using by minimizing the cross entropy loss $-\sum_{t=1}^{T} \log p_\theta(a_t|s_t)$ for both the actor network and the critic network (as mentioned we also consider CSP loss). After the actor is pre-trained, the critic network is passed sampled actions from the fixed pre-trained actor and updated accordingly. After this initial phase, we then begin training both actor and critic together.

Embedding Similarity Evaluation

Word Mover's Distance Sentence Similarity We also include WMD (Kusner et al. 2015) for measuring semantic similarity between ℓ_2 normalized embeddings associated with predicted and target words. Word-level embedding similarities offer a faster alternative to model-based sentence-level evaluation, hence we include it for our experiments.

To align WMD with word overlap metrics, we also include the penalization terms such as the brevity penalties used in BLEU (Papineni et al. 2002), as shown in Equation 13. Here, γ is the similarity measure, the length ratio lr $= |Y|/|\hat{Y}|$ (Shi, Knight, and Yuret 2016) and the brevity penalty bp $= \min\big(\exp(1 - 1/lr), 1\big)$ which penalizes shorter length generated sentences.

$$s = \sigma\Bigg(\text{BP} \cdot \gamma_{\text{wmd}}\big(E_{\hat{Y}}, E_Y\big)\Bigg) \tag{13}$$

Sliding Kernel Cosine Similarity We also considered decayed k-pairwise cosine similarity where k is a sliding window span that compares embeddings corresponding to n-gram groupings with a decay factor $\gamma \in [0, 1]$ that depends on the distance such that $\gamma_{(i,j)} = d(Y_i, Y_j)/k \ \forall i, j \in T$. In the below case we use the kernel shown in Equation 14 where i is the index corresponding to $y \in Y$ and j for $\hat{Y}$ respectively.

$$\gamma = \exp(-||i - j||) \tag{14}$$

This allows for any mis-alignments between sentences, as some may be shorter than others. There are T/k window spans, therefore we multiply the k/T by the brevity penalty.

$$s_{kcos} = \sigma\Big(\frac{k}{T}\text{BP}\sum_{i=1}^{T}\sum_{j=i-k}^{t+k}\gamma_{(i,j)} cos\big(E_{Y_i}, E_{\hat{Y}_j}\big)\Big)$$
$$s.t, \quad t \leq i \leq T - k \tag{15}$$

Results

Flickr30k Results

Table 1 shows the SoTA for image captioning on the Flickr30k dataset, not specific to policy-gradient methods as not all relevant papers include Flickr30k in experiments. Models proposed by Wu et al. incorporate external knowledge (SPARQL queries over DBpedia knowledge base) for image captioning, hence the increase in BLEU and Perplexity (PPL).

Table 2 compares ML training with previously published actor-critic approaches that use BLEU and ROUGE as the reward signal (Zhang et al. 2017). We a beam search of a beam size $B = 5$ at test time. The beam search retains B most probable prediction at each timestep and considers the possible next token w_{t+1}^{b} extensions for a beam b and repeats until timestep $T, \forall b \in B$.

We find that when using only WMD as the reward signal, which is model-free, we find an improvement on semantic

	Flickr-30k	ROUGE-L		BLEU2		BLEU3		BLEU4		WMD		COS	
		Val.	Test	Val.	Test	Val.	Test	Val.	Test	Val.	Test	Val.	Test
Baseline	**ML**	33.09	31.46	67.52	65.20	42.12	40.68	27.81	26.87	65.35	63.62	60.14	59.22
	BLEU	35.68	32.93	70.03	70.39	48.65	48.45	30.88	30.79	73.99	72.70	66.22	66.10
	ROUGE-L	36.47	33.67	68.98	68.55	47.55	47.14	31.68	31.56	71.94	70.08	65.55	62.42
Our	**WMD**	31.61	30.76	70.06	68.05	46.24	44.23	29.78	28.94	76.59	73.40	69.83	68.73
	InferSent	30.08	29.29	69.48	68.76	44.76	43.96	28.79	28.70	***78.28***	***77.01***	***71.23***	***69.16***
	Skipthought	26.22	25.36	70.01	67.43	43.41	42.20	29.23	27.81	***76.50***	***75.18***	***72.63***	***68.60***

Table 2: Flickr30k Results for ML, Actor-Critic and our proposed TRL AC Models (using a beam width of 5)

	MSCOCO	ROUGE-L		BLEU1		BLEU2		BLEU3		BLEU4		CIDEr		METEOR		WMD		COS	
		Val.	Test	Val.	Test	Val.	Test	Val.	Test	Val.	Test	Val.	Test	Val.	Test	Val.	Test	Val.	Test
SoTA	**MIXER** (Ranzato et al. 2015)	-	53.8	-	-	-	-	-	-	-	30.9	-	101.9	-	25.5	-	-	-	-
	MIXER-BCMR (Ranzato et al. 2015)	-	53.2	-	72.9	-	55.9	-	41.5	-	30.6	-	92.4	-	24.5	-	-	-	-
	PG-BCMR (Ranzato et al. 2015)	-	55.0	-	75.4	-	59.1	-	44.5	-	33.2	-	101.3	-	25.7	-	-	-	-
	SPIDEr (Liu et al. 2016)	-	54.4	-	74.3	-	57.9	-	43.1	-	31.7	-	100.0	-	25.1	-	-	-	-
	RAML @ $\tau = 0.9$ (Ma et al. 2017)		-	-	-	-	-	-	-	-	27.6	-	-	-	-	-	-		
	VSE@$\lambda = 0.4$ (Ren et al. 2017)	-	52.5	-	71.3	-	53.9	-	40.3	-	30.4	-	93.7	-	24.7	-	-	-	-
	TD-AC (Zhang et al. 2017)	-	***55.4***	-	***77.8***	-	***61.2***	-	***45.9***	-	***33.7***	-	***116.2***	-	***26.7***	-	-	-	-
	SCST (Rennie et al. 2017)	-	54.3	-	-	-	-	-	-	-	31.9	-	106.3	-	25.5	-	-	-	-
	SCST (Wu et al. 2018)	-	54.3	-	-	-	-	-	-	-	31.9	-	106.3	-	25.5	-	-	-	-
Baselines	**ML**	51.39	50.28	72.73	69.09	49.70	49.33	31.89	31.45	24.09	23.67	84.93	84.03	23.68	23.45	72.89	71.46	71.01	70.55
	BLEU	52.75	52.01	75.91	74.17	61.34	61.72	47.91	46.58	35.09	34.57	94.46	93.41	25.55	25.27	74.38	73.09	73.39	71.86
	ROUGE-L	56.28	55.25	72.98	72.55	51.55	50.29	37.44	35.38	32.44	31.09	95.53	95.51	25.61	25.54	73.53	72.65	73.88	72.10
Proposed	**WMD**	51.61	52.05	73.01	72.81	52.33	52.70	39.17	38.24	32.74	30.09	99.03	98.46	27.12	27.09	79.12	78.42	80.07	78.72
	InferSent	***55.46***	***54.25***	***75.58***	***75.02***	***60.40***	***57.16***	***46.24***	***41.68***	***31.93***	***31.24***	***106.12***	***105.68***	27.31	27.18	***82.86***	***80.26***	***83.71***	***82.58***
	Skipthought	53.02	52.71	74.49	73.61	54.54	51.08	32.28	31.02	29.78	28.59	105.56	105.08	***27.42***	***27.20***	81.95	80.21	81.25	80.63

Table 3: MSCOCO Results for ML, Actor-Critic and our proposed TRL AC Models (B=5)

similarity measures (i.e WMD and COS). Here, COS refers to the Sliding Kernel Cosine Similarity described in the previous section. Interestingly, we also find WMD improves over ML for word-overlap eventhough the measure does not optimize for a dirac distribution, like ML training.
Both TRLs (InferSent and Skipthought) make significant improvements on WMD and COS. Hence, we infer that these TRLs that learn sentence similarity produce semantically similar sentences at the expense of a decrease in word-overlap (expected since the model is not restricted to predicting the exact ground truth tokens). This relaxes the strictness of word-overlap and allows for diversity in the generated captions. Moreover, WMD does not penalize sentence length and thus promotes diversity in caption length. However, as mentioned, we do include brevity penalty in WMD and COS for the purposes of easier comparison to word overlap metrics.

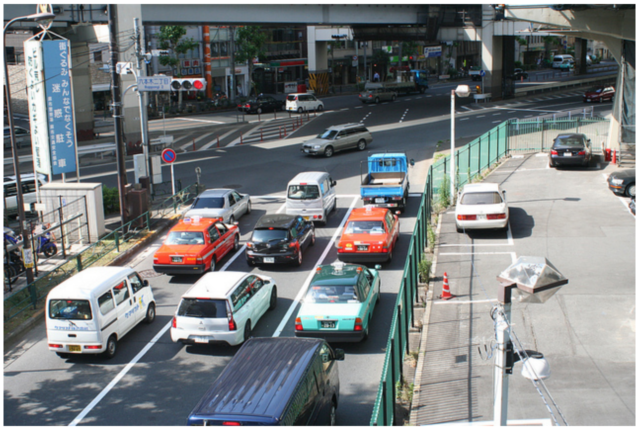
MSCOCO Results

The top of Table 3 shows SoTA results for policy-gradient methods based on the best average score on BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Denkowski and Lavie 2014), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) evaluation metrics. VSE is the aforementioned Visual Semantic Embedding model that uses TD(λ) at $\lambda = 0.4$. For SCST, these results are from the test portion of the Karpathy splits (Karpathy and Fei-Fei 2015) using CIDEr for optimization. Policy gradient methods have reached near top of the MSCOCO competition leaderboard without using ensemble models.

The lower end shows the results of our proposed models and baselines evaluated on both n-gram overlap based measures and word-level (Cosine) and sentence-level (WMD) embedding based evaluation measures. We find that the largest gap in performance between our proposed TRLs and n-gram overlap metrics (BLEU and ROUGE-L) reward signals are found on the embedding-based evaluation measures.
For all TRLs (WMD, InferSent, Skipthought) performance consistently improves over ML, BLEU and ROUGE-L when evaluated on WMD and Cosine. This suggests that even though we may not strictly predict the correct word as measured by word-overlap measures, the semantic similarity of sentences is preserved as measured by WMD and Sliding Kernel Cosine Similarity. Furthermore, this results in more diverse text generation as the policy network is not penalized for constructing candidate sentences that do not have high word overlap with the reference captions.

TRLs outperform word overlap policy rewards such as BLEU and ROUGE-L on our embedding similarity based metrics. Of the three, we find the InferSent TRL to outperform the other two, with the unsupervised Skipthought TRL being competitive for all metrics. We also see results are competitive to the SoTA. We find similar findings for TRL models evaluated on CIDEr and METEOR.

Figure 2 shows an example of the ground truth captions (Human), ML trained generated caption, a baseline AC



Human	*An image of a cars driving on the highway* *A section of traffic coming to a stop at an intersection.* *A bunch of cars sit at the intersection of a street.* *This is a picture of traffic on a very busy street.* *A busy intersection filled with cars in asia.*
ML	*an image of a sitting car in traffic*
AC-BLEU	*A group of cars at an intersection.*
AC-WMD	*A group of cars at lights near a traffic intersection.*
AC-Skipthought	*A group of cars near a busy intersection road.*
AC-InferSent	*A picture of cars stopping near the traffic intersection.*

Figure 2: Qualitative Results on MSCOCO Val. Set

trained with BLEU scores and our three proposed alternatives that improve for semantic similarity. We demonstrate the difference between text generated for an image of a traffic jam near an intersection. The example also illustrates that the ground truth itself is imperfect, both syntactically ('..of a cars..') and semantically ('..cars sit at the intersection..'). The TRL will assign lower return in these cases, whereas word-overlap measures do not explicitly penalize how bad the semantic or syntactic differences are between predicted and ground truth sentences.

Conclusion

We proposed to use pretrained models that are specifically trained on sentence similarity tasks that can be used to issue rewards and to define, optimize and evaluate language quality for neural-based text generation. We find performance on semantic similarity metrics improve over a policy gradient model, namely the actor-critic model, that uses unbiased word overlap metrics as rewards.

The InferSent actor-critic model improves over a BLEU trained actor-critic model on MSCOCO when evaluated on a Word Mover's Distance similarity measure by 6.97 points and 10.48 points on sentence-level cosine embedding metric. Large performance gains are also found for Flickr-30k dataset, demonstrating the general applicability of the proposed transfer learning method. We conclude that model-based task should be considered for reinforcement learning based approaches to conditional text generation.

References

[Abbeel and Ng 2004] Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 1. ACM.

[Bahdanau et al. 2016] Bahdanau, D.; Brakel, P.; Xu, K.; Goyal, A.; Lowe, R.; Pineau, J.; Courville, A.; and Bengio, Y. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*.

[Barto, Sutton, and Anderson 1983] Barto, A. G.; Sutton, R. S.; and Anderson, C. W. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics* 1(5):834–846.

[Bowman et al. 2015] Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

[Chen and Lawrence Zitnick 2015] Chen, X., and Lawrence Zitnick, C. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2422–2431.

[Christiano et al. 2017] Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

[Cui et al. 2018] Cui, Y.; Yang, G.; Veit, A.; Huang, X.; and Belongie, S. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5804–5812.

[Denkowski and Lavie 2014] Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

[Devlin et al. 2018] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Donahue et al. 2015] Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625–2634.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[Hochreiter and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

[Karpathy and Fei-Fei 2015] Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137.

[Kingma and Ba 2014] Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

[Kiros et al. 2015] Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, 3294–3302.

[Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

[Kusner et al. 2015] Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, 957–966.

[LeCun, Bengio, and others 1995] LeCun, Y.; Bengio, Y.; et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361(10):1995.

[Lin et al. 2014] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

[Lin 2004] Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

[Liu et al. 2016] Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2016. Optimization of image description metrics using policy gradient methods. *CoRR* abs/1612.00370.

[Liu et al. 2017] Liu, C.; Mao, J.; Sha, F.; and Yuille, A. 2017. Attention correctness in neural image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[Ma et al. 2017] Ma, X.; Yin, P.; Liu, J.; Neubig, G.; and Hovy, E. 2017. Softmax q-distribution estimation for structured prediction: A theoretical interpretation for raml. *arXiv preprint arXiv:1705.07136*.

[Mao et al. 2014] Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; and Yuille, A. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.

[Norouzi et al. 2016] Norouzi, M.; Bengio, S.; Jaitly, N.; Schuster, M.; Wu, Y.; Schuurmans, D.; et al. 2016. Reward augmented maximum likelihood for neural structured prediction. In *Advances In Neural Information Processing Systems*, 1723–1731.

[Papineni et al. 2002] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

[Ranzato et al. 2015] Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

[Ren et al. 2017] Ren, Z.; Wang, X.; Zhang, N.; Lv, X.; and Li, L.-J. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 290–298.

[Rennie et al. 2017] Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7008–7024.

[Ross and Bagnell 2014] Ross, S., and Bagnell, J. A. 2014. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*.

[Shi, Knight, and Yuret 2016] Shi, X.; Knight, K.; and Yuret, D. 2016. Why neural translations are the right length. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2278–2282.

[Sutton et al. 2000] Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, 1057–1063.

[Szegedy et al. 2017] Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[Vedantam, Lawrence Zitnick, and Parikh 2015] Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

[Vinyals et al. 2015] Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

[Williams 1992] Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.

[Wu et al. 2018] Wu, Q.; Shen, C.; Wang, P.; Dick, A.; and van den Hengel, A. 2018. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence* 40(6):1367–1381.

[Xu et al. 2015] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

[Young et al. 2014] Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.

[Zhang et al. 2017] Zhang, L.; Sung, F.; Liu, F.; Xiang, T.; Gong, S.; Yang, Y.; and Hospedales, T. M. 2017. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*.