Creating Information-maximizing Natural Language Messages Through Image Captioning-Retrieval

Fabian Karl[1,2] and **Mikko Lauri**[1] and **Chris Biemann**[2]

fabian.alexander.karl@gmail.com
lauri@informatik.uni-hamburg.de
biemann@informatik.uni-hamburg.de
[1]Computer Vision, Department of Informatics, MIN, Universität Hamburg
[2]Language Technology, Department of Informatics, MIN, Universität Hamburg

Abstract

In this work, we propose the Image Captioning-Retrieval (ICR) problem that states the objective of language generation as information exchange. To solve the ICR problem, we design and implement an end-to-end neural network architecture that describes the content of images in natural language, and retrieves them solely based on these generated descriptions. The main goal is to be able to generate information-maximizing natural language messages. We experimentally show a strong increase in message information content while losing some grammatical correctness in the generated descriptions in a semi-supervised setting where caption generation is trained towards retrieval quality.

1 Introduction

Human thinking and reasoning are deeply connected to words and language. Turing (1950) famously defined the ability to hold a complex conversation as artificial intelligence. While this notion is debated (Searle, 1980), it is widely accepted that it is language that makes us human. An artificial system capable of producing human language will be received by us as human-like.

Current conversational and language producing systems can broadly be categorized into three classes: rule-based systems, supervised learning systems, and Reinforcement Learning (RL) models. Rule-based systems produce outputs by a set of conditionals and rules of varying complexity. This approach works well for expert systems and the understanding of simple commands. Due to the predictability and traceability, rule-based language systems dominate commercial applications. Supervised learning systems apply supervised optimization strategies to predict appropriate language outputs for given inputs (Vinyals and Le, 2015). A prerequisite is a corpus of conversational training examples containing input sentences and corresponding output sentences. RL-based conversational systems (English and Heeman, 2005; Li et al., 2016) seek to learn a dialog policy that guides how the artificial agent should follow when interacting with a user.

While current state-of-the-art systems are arguably able to produce language that seems human-like, their objective is stated as mere production of well-sounding sentences. However, production of grammatically correct sentences as an end goal falls short of the motivation humans have for language production, namely the exchange of information (Kirby, 2007). In Mathur and Singh (2018) it is noted that especially sequence-to-sequence models cannot solve the language modelling problem, since *"the objective function that is being optimized does not capture the actual objective achieved through human communication, which is typically longer term and based on exchange of information rather than next step prediction"*. The main driver of a conversational system should not be the direct production of sentences in a human-readable language, but the optimal amount of information exchange between agents (Steels, 2015).

In this work, we examine language generation through an alternative objective of maximum information exchange. We propose to train a language production system directly with the motivation of maximizing information content, rather than using language modelling objectives. To achieve this,

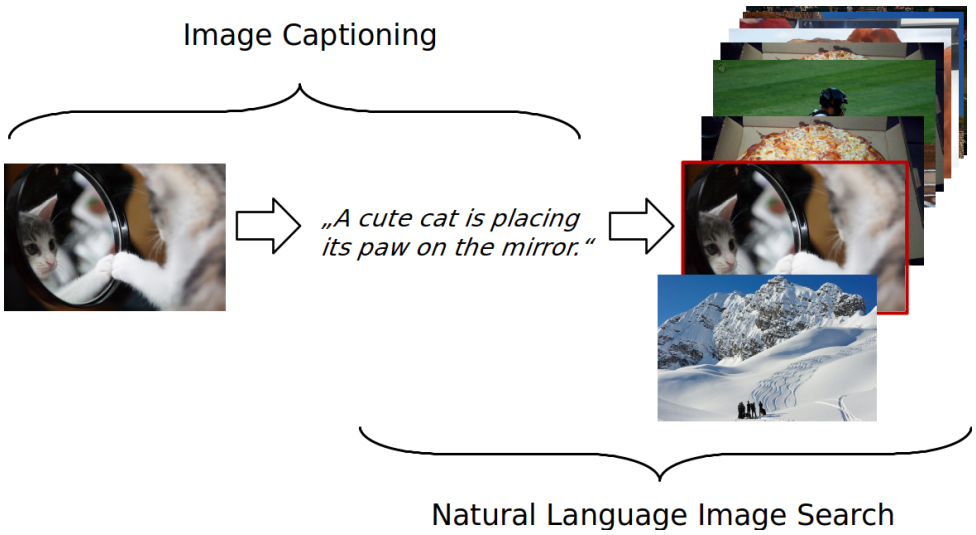

Figure 1: The Image Captioning-Retrieval (ICR) problem simulates a natural language message passed from one agent to another, and is composed of Image Captioning (IC) and Natural Language Image Search (NLIS).

we propose the Image Captioning-Retrieval (ICR) problem. The ICR problem simulates a message passed from one agent to another, and is composed of two parts: Image captioning (IC) and natural language image search (NLIS) as illustrated in Figure 1. IC describes or captions a given image with a sentence in natural language. NLIS takes the caption as an input and retrieves the closest image out of a set of candidate images. By combining IC and NLIS, we can train our language production system directly with the motivation of information exchange. The constraint that the communication takes place in human-understandable language is ensured by producing captions in natural language. For this, we first pre-train the IC system in a supervised fashion using pairs of images and captions, and subsequently continue to train the overall system on the retrieval task. This can be viewed as a semi-supervised setting since captions are improved not through direct supervision on gold captions, but on indirect supervision on discriminating between pictures in the retrieval task.

Our contribution is two-fold. Firstly, we show that solving the ICR problem gives rise to natural language messages, while experimentally showing a strong increase in message information content. Secondly, we qualitatively present that the descriptions generated by our model capture more details of images as compared to plain IC systems.

The remainder of the paper is organized as follows. In Section 2, we review relevant related works in image captioning, natural language image search and neural learning architectures. Section 3 describes our overall approach, detailing the respective subsystems and their combination. The experimental setup is laid out in Section 4, before reporting quantitative evaluation results in Section 5. Qualitative observations are discussed in Section 6, Section 7 draws conclusions and provides directions for further work in natural language learning through conversations.

2 Related Work

State-of-the-art natural language production systems apply supervised learning, in particular the sequence-to-sequence model of Vinyals and Le (2015). This approach was inspired by machine translation (Sutskever et al., 2014), and has since been replicated multiple times. While an in-depth survey of natural language generating systems is beyond the scope of the present paper, we direct the interested reader to the recent survey of Gatt and Krahmer (2018). In our subsequent review, we discuss the two key subtasks of our ICR problem (Fig. 1), IC and NLIS, and the interplay of systems solving these two tasks.

Given an input image, an IC system outputs a description of the image in natural language. In turn, given as input a textual description of an image, an NLIS system finds the image that best matches the input description among a set of candidate images. We review techniques and ideas most closely related to our focus on the information exchange mo-

tivation for language generation. These approaches typically combine an IC network and an NLIS network and train them jointly. For a recent general survey of deep learning techniques applied to IC, we refer the reader to Hossain et al. (2019).

Most related to our work, the idea of scoring image descriptions based on the amount of information carried in the sentence is proposed in Hodosh et al. (2013). Instead of using traditional n-gram based evaluation measures like the BLEU (Papineni et al., 2002) or the CIDEr score (Vedantam et al., 2015), Hodosh et al. (2013) propose to use an NLIS system, pre-trained on human-annotated image-caption-pairs, to score the created image captions. The idea is widely used in other recent works in IC (Devlin et al., 2015; Vinyals et al., 2017; Karpathy and Fei-Fei, 2017; Donahue et al., 2017). The general architecture of these models contains an IC encoder-decoder model that encodes image information into textual form, and an image scoring system that evaluates the created captions using an NLIS system. The IC model is often a combination of a convolutional neural network (CNN) and a long-short-term memory network (LSTM).

Adversarial training is employed by several state-of-the-art works in IC (Dai et al., 2017; Liang et al., 2017; Liu et al., 2018). An NLIS network is applied to discriminate between generated and real samples. In Shetty et al. (2017), the objective is altered from merely reproducing ground truth captions to matching a distribution of human generated captions by applying an approximate Gumbel sampler.

RL is employed in some recent approaches such as the method by Ren et al. (2017b). A reward function is derived by considering visual-semantic embedding similarities: input images and captions both are mapped into a embedding space, and their similarity in this space is measured by an appropriate metric.

In contrast to the reviewed work we explicitly define information exchange as the primary objective for IC and NLIS. Through this we clearly separate us from related studies that use information exchange merely as a performance indicator or a general guidance.

3 Image Captioning Retrieval Network

Our ICR network is an IC network and an NLIS network, combined by a Gumbel softmax layer.

3.1 Image Captioning

The IC model receives an image and returns multiple probability distributions over a vocabulary.

The input for the model is an image $x^{im} \in \mathbb{R}^{h \times w \times c}$, where h, w, c are the height, width and color dimension, together with a sequence of words. The model output is a probability distribution over a fixed vocabulary V. Each word is thus assigned a likelihood of being the next word.

The input image is resized to a fixed size and fed through an image encoder (e.g. CNN) with the parameters θ_ϕ that extract the most important image features in a vector $\phi(x^{im}, \theta_\phi) \in \mathbb{R}^k$, where k is the length of the feature vector.

The respective image annotation is embedded in a dense word embedding, yielding the second model input $x^{se} \in \mathbb{R}^{t \times d}$, where t is the number of words in a sentence and d is the dimensionality of the dense word embedding. The embedded sentence is fed through a sentence encoder (e.g. LSTM) resulting in a $t \times l$ tensor, where l is the length of the feature vector.

Now x^{im} is replicated t times and concatenated with the sentence features. This results in a $t \times (l + k)$ tensor, which is fed through a block of fully connected layers and a final softmax layer, squeezing the model output into t probability distributions with $\mathrm{P}(y_t | x^{se}_{1 \to t-1}, \phi(x^{im}, \theta_\phi))$, where y_t is the probability over the vocabulary V at timestep t, x^{se} is the information from the previous words and $\phi(x^{im}, \theta_\phi)$ is the image vector.

At training time, x^{se} and the target $y \in \mathbb{R}^{t \times d}$, with the same shape as x^{se}, are representations of the same *ground truth* sentence. This training technique is called teacher forcing. x^{se} is shifted one time-step into the future by adding a *start-symbol* at its beginning. This way, word y_t equals x^{se}_{t+1} and the model is trained to predict the next word of the same sentence x^{se}. An *end-symbol* is appended to y, so input and output have the same length and the model is trained on how to end the sentence. The loss is calculated through the cross-entropy of the predicted probability distribution and the *ground truth* distribution. This allows a quick and stable learning process but also leads to the so-called *exposure bias* (Ranzato et al., 2016).

At inference, only the image vector $\phi(x^{im}, \theta_\phi)$ is available. The model starts with $\hat{x}^{se}$, containing only the *start-symbol*, as first input and generates $\mathrm{P}(\hat{y}_t)$. Depending on the selection mode, one word y_t from $\mathrm{P}(\hat{y}_t)$ is selected and appended to the pre-

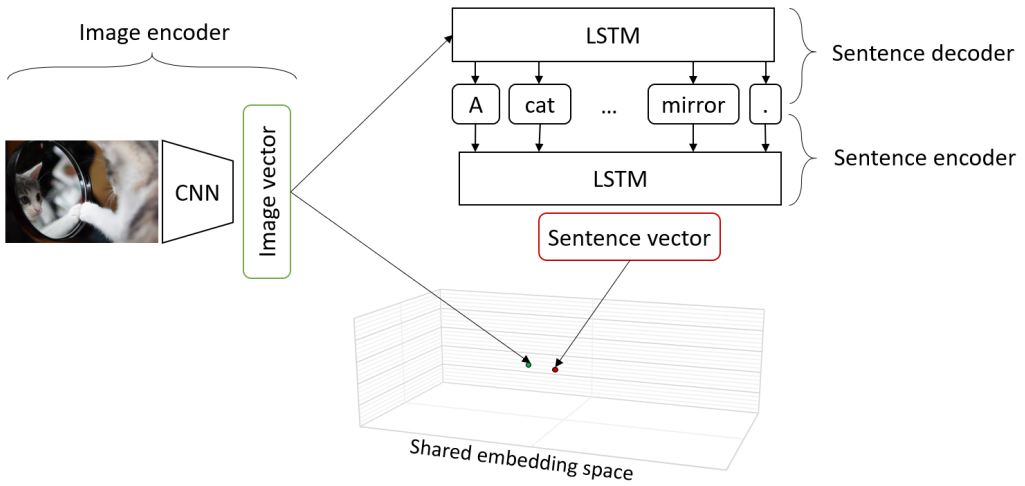

Figure 2: Our ICR model. Annotations and images are both transformed into feature representations, which are mapped into a shared embedding space. The distance in this space defines the similarity of the image-annotation pair.

vious input $\hat{x}^{se}_{1\to t-1}$. It then serves as new input for the next prediction step.

3.2 Natural Language Image Search

Our NLIS model is realized through an image and a sentence encoder that are trained on the triplet ranking loss (Karpathy et al., 2014; Ren et al., 2017a; Karpathy and Fei-Fei, 2017; Wang et al., 2017; Faghri et al., 2018; Liu et al., 2018).

Both encoders are similar to the ones used in our IC model. Images x^{im} are transformed into feature representations $\phi(x^{im},\theta_\phi)\in\mathbb{R}^{d_\phi}$, where ϕ is the image encoder (e.g. CNN), with model parameters θ_ϕ. Correspondingly, sentences x^{se} are embedded and transformed into a feature representation through a model $\psi(x^{se},\theta_\psi)\in\mathbb{R}^{d_\psi}$, where ψ is the sentence encoder (e.g. LSTM) with model parameters θ_ψ.

$$\mathrm{f}_{im}(x^{im},W_{im},\theta_\phi) = \left\|W_{im}^T\phi(x^{im},\theta_\phi)\right\|_2 \quad (1)$$

$$\mathrm{f}_{se}(x^{se},W_{se},\theta_\psi) = \left\|W_{se}^T\psi(x^{se},\theta_\psi)\right\|_2 \quad (2)$$

Both feature representations are mapped into a shared embedding space of size e by linear projection with weight matrices $W_{im}\in\mathbb{R}^{d_\phi\times e}$ and $W_{se}\in\mathbb{R}^{d_\psi\times e}$. The resulting projections are normalized with the L2 norm to lie on the unit hypersphere.

$$\mathrm{s}(im,se) = \mathrm{f}_{im}(x^{im},W_{im},\theta_\phi)\cdot\mathrm{f}_{se}(x^{se},W_{se},\theta_\psi) \quad (3)$$

The similarity between an image-sentence pair is defined as the inner product between the two normalized vectors, resulting in the cosine similarity (Subhashini and Kumar, 2010).

$$\mathscr{L}(\theta,B_{im},B_{se}) = \frac{1}{N}\sum_{n=1}^{N}\mathrm{L}(im_n,se_n,B_{im'},B_{se'}) \quad (4)$$

For a batch of images, $B_{im}=\{x^{im}\}_{n=1}^N$, and corresponding sentences, $B_{se}=\{x^{se}\}_{n=1}^N$, the batch loss is calculated by comparing every image against every sentence and vice versa. In every iteration, one image-sentence pair is selected as *true* pair, marked as (im_n,se_n). The similarity of this pair is compared to the similarities between the image and all other sentences or the sentence and all other images respectively. A batch of sentences, without the correct sentence, is denoted as $B_{se'}$ and a batch of images without the correct image as $B_{im'}$.

All possible parameters to be optimized are defined by $\theta=\{\theta_\phi,\theta_\psi,W_{im},W_{se}\}$. Depending on the experimental setup, however, θ_ϕ and/or W_{im} are not optimized or finetuned.

$$\begin{aligned}\mathrm{L_{SH}}(im,se,\hat{im},\hat{se}) = \\ \sum_{\hat{se}}[\alpha-\mathrm{s}(im,se)+\mathrm{s}(im,\hat{se})]_+ + \\ \sum_{\hat{im}}[\alpha-\mathrm{s}(im,se)+\mathrm{s}(\hat{im},se)]_+\end{aligned} \quad (5)$$

$\mathrm{L_{SH}}$ is defined as the sum of hinges and describes the *classic* triplet ranking loss. Let α be the margin that the similarity of all wrong image-sentence pairs should be smaller than the similarity of the

correct image-sentence pair. $s(im, se)$ describes the similarity of the correct image-sentence pair while $s(im, \hat{se})$ describes the similarity between an incorrect image-sentence pair. In order to avoid negative losses, we use positive values only, as defined by the notation $[x]_+ = \max(0, x)$. The second term is symmetrical to the first term. In the first term, an image is fixed and the similarity with different candidate sentences is calculated and returned. In the second term, a sentence is fixed and all other images are iterated over to calculate the similarities. Faghri et al. (2018) report a steep increase in accuracy on the NLIS task when using triplet ranking loss with the max of hinges, L_{MH}. This refers to selecting the one (negative) sample with the highest loss in every mini-batch. The only difference between L_{MH} and L_{SH} is the selection of the biggest error, $\max_{\hat{se}}[\alpha - s(im, se) + s(im, \hat{se})]_+$, instead of the summation of errors, $\sum_{\hat{se}}[\alpha - s(im, se) + s(im, \hat{se})]_+$.

3.3 Image Captioning-Retrieval

Our ICR network is a combination of the two models described above. In order to overcome the problem of discrete word representations being not differentiable, the Gumbel softmax trick (Jang et al., 2016) is used to transform one-hot probability distributions into pseudo-one-hot-representations.

The original Gumbel-Max trick (Gumbel, 1954) is a simple and efficient way to draw samples from a categorical distribution with class probabilities π. $g \in (0, 1)$ is called the Gumbel distribution and is calculated from u, drawn from a uniform distribution between 0 and 1.

$$g = -\log(-\log(\text{Uniform}(0, 1))) \tag{6}$$

$$z = \text{one hot}\left(\underset{i}{\text{argmax}}[g_i + \log(\pi)]\right) \tag{7}$$

Since argmax is non-differentiable, the continuous softmax function is used as an approximation. τ is the temperature of the softmax. The smaller τ is, the closer the distribution is to a one-hot encoding. y_i is the resulting k-dimensional word distribution.

$$y_i = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(\pi_i) + g_i)/\tau)} \quad \text{for } i, ..., k \tag{8}$$

A second challenge is the sampling of novel sentences. Our ICR model needs a complete input sentence x^{se} to be able to determine the probability for every sub-sentence $x^{se}_{t_1:t_i}$. This can either be achieved by creating complete and novel sentences with our IC model in a pre-step or by directly using the Gumbel softmax trick in this phase. Since the Gumbel softmax activation function introduces randomness into the selection process, unseen word combinations can occur, from which the model will not be able to recover. For this reason we decided to use the first-mentioned approach.

When feeding the novel image annotation through our ICR model, it will be fed through the IC model again and reproduce the output $\hat{y}$. The output is transformed with the Gumbel softmax activation function, which selects one word randomly based on its probability and transforms it into a value close to one. All other words will receive a very low probability, close to 0. Let $\gamma(\hat{y})$ be the Gumbel softmax output.

Together the original image vector $\phi(x^{im}, \theta_\phi)$ and $\gamma(\hat{y})$ are fed into the NLIS network to output a similarity matrix, containing similarities between every image and every sentence. From this similarity matrix, either the sum or the max of hinges loss (Section 3.2) can be calculated and used for training.

4 Experimental Setup

Our experiments are designed to optimize information exchange between the IC and the NLIS system. Information exchange is measured by the image retrieval score, which is reported in the percentage of images ranked within the best 1, 5 or 10 ranked images (r@1, r@5, r@10). The Consensus-based Image Description Evaluation (CIDEr) (Vedantam et al., 2015) score for the generated annotations is presented alongside. CIDEr is an n-gram based evaluation metric especially created for image annotation.

We use MSCOCO dataset with 2017 split (Lin et al., 2014; Chen et al., 2015) for training and validation. The dataset contains 118,287 training and 5,000 validation images, all of them annotated with five ground truth sentences.

In preprocessing, all annotations are cut or padded to contain exactly 16 tokens. Tokens appearing less than 10 times are replaced with the *unknown word* token. Every word is embedded with a pre-trained English fasttext model (Bojanowski et al., 2017). All images were encoded by extracting the last fully connected layer of ResNet50

Table 1: Performance of our IC and NLIS model after stand-alone pre-training on their respective task (*Our), compared to VSA (Karpathy and Fei-Fei, 2017), UVS (Kiros et al., 2014), VSE++ (Faghri et al., 2018), sm-LSTM (Huang et al., 2017), m-RNN (Mao et al., 2014) and LRCN (Donahue et al., 2017) on different measures as reported in the literature. NLIS results refer to 1,000 test images and 5,000 respective descriptions from MSCOCO 2017. r@n shows the percentage of sentences/images ranked under the top n ranks. BLEU4 and CIDEr are received from the C40 test set of the official 2015 COCO Caption Challenge Competition. Results with most similar architectures are listed if available.

	Image captioning			Image retrieval		
System	r@1	r@5	r@10	r@1	r@5	r@10
VSA	38.4	69.9	80.5	27.4	60.2	74.8
*Our	39.9	69.8	80.1	32.0	66.3	80.8
UVS	43.4	75.7	85.8	33.0	67.2	80.6
VSE++	43.6	74.8	84.6	33.7	68.8	82.0
sm-LSTM	53.2	83.1	91.5	40.7	75.8	87.4

	Image captioning	
System	BLEU4	CIDEr
VSA	0.446	0.692
*Our	0.472	0.753
UVS	0.517	0.752
m-RNN	0.578	0.896
LRCN	0.585	0.934

(2,048 nodes), pre-trained on ImageNet (Deng et al., 2009).

IC and NLIS network are separately pre-trained until they yield optimal annotation and ranking results. Multiple hyperparameters (model architecture, number of epochs, learning rate, etc.) of both models were empirically optimized to yield results close to state-of-the-art performance for their respective task. For both models, sentences are encoded by 1,024 LSTM cells. Images are projected onto vectors of the same size with a dense layer. In our NLIS network, encoded sentence features are also projected onto a 1,024 dimensional space by a fully connected layer. In our IC model, sentence and image features are concatenated and fed through two dense layers (1,024 and 2,048 nodes), before the final softmax layer. Between every layer we added dropout layers with 0.4 dropout to prevent the model from overfitting. When training our NLIS model we used *sum* of hinges for one epoch before switching to *max* of hinges loss. This was necessary for a stable training. In all later ICR experiments we used *max* of hinges loss.

Table 1 shows the performance of our models after pre-training compared to related studies that used similar techniques with similar network architectures. The performance of our NLIS model builds the baseline for further training with our complete ICR model. In order to combine IC and NLIS model in our final model, we implemented both models in the same framework. Simply reusing models from related work was not possible due to the incompatibility of different neural network frameworks.

In our main training loop, 20,000 images are randomly selected per epoch and fed through our ICR network. A loss is calculated for the generated annotation and for the retrieved image. Mini-batch size is set to 128 for all experiments. The model was trained for 40 epochs with Adam as optimizer. The learning rate is set to 0.0002 for the first 20 epochs and then decreased to 0.00002 for the rest of the training process.

Optimizing all weights in the ICR network leads to an unstable training process and often resulted in sudden drops in performance. Freezing the weights of the image projection layer from the beginning of training (IP=F) or at a certain epoch (IP=17) stabilized the training process. Freezing the weights of the sentence encoder (SE=F) had a similar stabilizing effect on the training. Training with only self-generated sentences right from the start leads to an instant decrease in performance since the model has no time to adjust to flawed input sentences. To counter this issue, novel self-generated annotations are slowly added to existing ground truth sentences. This is implemented by randomly selecting an annotation from a list of both ground truth annotations and generated ones. In the beginning, this list contains only ground truth samples. At every epoch, novel annotations are added. When the list reaches a defined size (INF=5, 10, 15), a random sentence is dropped from the list. This way, novel sentences are slowly infused into the training process.

To increase ranking performance, true image-sentence pairs were added to the output from the IC network. In this case, one mini-batch contains 64 image-sentence pairs generated by our IC network and 64 true image-sentence pairs directly from the dataset (TP=T). Otherwise the whole mini-batch contains only self-generated samples (TP=F). Both methods result in a 128×128 similarity matrix for one mini-batch. After the training phase, 1,000 validation images are captioned and retrieved to

Table 2: Ranking retrieval results for different experimental settings on 1000 validation images from MSCOCO 2017. *TP=True Pairs, SE=Sentence encoders trainable, IP=Image Projection layer trainable or trained until which epoch, INF=Infusion list size, NLIS sum=Sum over all image scores, C=CIDEr*

				Sentence Retrieval			Image retrieval				
TP	SE	IP	INF	r@1	r@5	r@10	r@1	r@5	r@10	NLIS sum	C
F	T	F	10	34.7	72.1	86.9	33.2	69.7	83.7	186.6	0.061
F	F	F	10	40.8	77.8	89.9	38.8	76.3	88.9	204.0	0.101
T	T	17	10	44.4	79.9	90.5	40.8	79.1	89.7	209.6	0.049
T	F	T	10	47.6	84.2	93.6	43.1	81.8	92.5	217.4	0.094
T	F	T	15	46.2	86.4	93.6	46.0	83.0	93.0	222.0	0.083
Pre-training Baseline				39.9	69.8	80.1	32.0	66.3	80.8	179.1	0.753

determine the performance of our model.

5 Results

Table 2 shows various experimental settings and their resulting ranking and CIDEr score. In the last row, the baseline ranking and annotation performance is reported. It represents our best performance of the two models when trained on their respective tasks alone.

The table shows that the usage of true image pairs (TP) generally increases the ranking performance of the network. The best experimental results were observed when freezing the sentence encoder weights for the ICR training (SE=F) but not the image projection layer (IP=T). An infusion list size of 10 (INF=10) yields optimal sentence retrieval scores while an infusion list size of 15 (INF=15) results in a 3 percent-point increase in the r@1 for the NLIS score and the best overall retrieval score (NLIS sum). Training runs with no infusion list (not mentioned in Table 2) were abandoned early in the experimental phase, for they resulted in unstable training and worse ranking scores than our baseline.

Compared to the ranking performance of our baseline (Table 1), we observe improvements for all reported experiments. Under the same evaluation set (1000 validation images), our best model improves image r@1 results by 14.0 percentage points resulting in 46.0% correctly retrieved images through our self-generated image descriptions. 80.0% of all described images were retrieved within the 10 top ranks. Not only could we increase our retrieval performance immensely compared to our baseline, but we also outperform all related studies using similar image encoders. This indicates that our self-generated sentences contain more image information than the *ground-truth* annotations, created by human annotators. CIDEr scores, however, decrease from our baseline performance of 0.72 to around 0.10.

The increase in retrieval scores and the decrease in CIDEr can be observed in Figure 3 as well. It shows a selection of images and different annotations. The first annotation is the annotation generated by our IC system, after pre-training (PT). GT shows one of the *ground truth* captions for comparison. The last sentence is the generated description from our best performing (ICR) model. Word repetitions, missing of stop-words and the selection of more specific and precise words (e.g. locomotive instead of train) are at the same time responsible for higher retrieval scores and lower CIDEr score. Since n-gram based evaluation metrics use direct comparison between prediction and ground-truth sentence, using often occurring words (e.g. stop-words) and general terminology (e.g. train) normally yields better results. Ironically, these words often carry the least amount of information.

6 Discussion

A comparison between the images in Figure 3 and their descriptions after the pre-training phase and after the ICR training phase shows that the increase in information exchange is not only visible in the ranking scores, but also leads to arguably better generated descriptions.

The sentences created after the pre-training are almost exclusively grammatically correct and describe the image content more or less accurately. Generated descriptions show less grammatical structure after the IC system was trained to maximize the ranking performance, but the content of the sentence describes the image in much more detail and correctness.

The generated sentences after ICR training often

PT: *a train is parked on the tracks in* a city .
GT: *A train traveling past tall white buildings on train tracks.*
ICR: *an old locomotive train caboose on an railroad train tracks wires wires wires wires overpass overpass*

PT: *a man is holding a frisbee in his hand .*
GT: *A woman wearing glasses holding a tennis racket.*
ICR: *this man holding his tennis racket wearing his neck shirt shirt shirt sleeve sleeve wrist wrist*

PT: *a man is playing tennis on a court .*
GT: *Two men shaking hands while standing on a tennis court.*
ICR: *two men man standing on an tennis court on an tennis court net courts courts courts*

PT: *a man riding on the back of an elephant .*
GT: *Man riding an elephant up a hill near a field.*
ICR: *two man riding on elephant elephant walking through some river river saddle a river stream river*

Figure 3: Next to every image, the description generated by the pre-trained IC model (PT), one of the ground truth descriptions (GT) and the descriptions, generated after training the ICR model.

contain repeating words, and they do not contain the *end-symbol* anymore. Both of these effects are likely due to the pre-training of the system. During the pre-training phase, only correct sentences were used as input for the model. In the ICR training phase, new sentences are generated and used for training. Additionally, since the Gumbel softmax trick is a statistical sampling method, the word with the highest probability is not always picked, as it has been before with greedy picking. This means the system encounters new situations that it has to deal with. Since it was trained with teacher-forcing, it has developed little robustness against these novel situations. Interestingly, the ICR system tries to fully use the maximum length of 16 tokens, possibly conveying the importance of image elements with word repetition.

It is important to mention that the grounding between words and entities in the images stays intact during the training. This means, the network keeps using the same words for certain scenes or objects, learned in the pre-training phase. This is highly relevant for a system trying to learn language without explicit targets. It means that the system keeps connections between image entities and words, even when trained on a different task. This allows us to focus on a more implicit goal like information exchange.

Regarding the first image in Figure 3, one can see, that the description after the ICR training includes *"an old locomotive"* instead of only *"a train"*. The description also contains *"wires overpass"*, describing the electrical wires over the train, even though this information was not present in any of the 5 human annotated sentences. This shows that the model is no longer explicitly trained on the *true* sentences, but has a much more implicit objective. In order to optimize the ranking performance, additionally, highly distinct image information is reflected in the wording. The fourth image in Figure 3 shows similar increases in content and detail description. The information that the elephant is *"walking through some river"* is crucial to distinctly rank this image higher than other elephant images.

In the third picture, the new description is less general. The pre-trained system is producing a generic sentence, more or less fitting to any tennis scene. The description, generated after the ICR training is more accurate in its context. The same is true for images 2 and 3. In general, the image content is described in more detail and in more accuracy. The sentences are less grammatical than before, however.

These findings are satisfying and show that our objective trains our system to transfer information while still creating human-readable sentences. The fact that the created sentences are still grounded show that our language system, once pre-trained, keeps its relations between objects and words intact. Our main goal of increasing the amount of exchanged information is clearly reached. Our secondary goal of insuring the human-readability of the generated language is partly satisfied and could be addressed with future work.

7 Conclusion

We clearly show how training an IC network with a more implicit objective like the ranking results

from our NLIS network can improve the amount of information captured in the generated sentences. The newly generated sentences are not grammatically perfect but understandable by humans. More importantly, after training our ICR model, generated descriptions capture more distinct details of images and describe more aspects of the images. The ranking performance was increased by a large margin, surpassing previous image search approaches.

This work has strengthened our belief that language generation and comprehension learning can benefit from implicit objectives in a joint learning setup as opposed to learn them from explicit supervision separately. Language offers a mapping from a high dimensional to discrete space. It offers the exchange of complex information in an equally complex but agreed-upon system. If information exchange is a major goal, more effort should be placed in implicitly modeling, with objectives like information exchange in order to solve tasks, requiring content that can only be transferred by language. The proposed language game in this work builds one of the most basic language games: describing and finding an image.

More sophisticated games, like solving riddles, answering questions, walking through a maze or executing commands can all be implemented based on language instructions. These games all have to be designed in a way that succeeding is a direct implication of information exchange. If this approach is used, while language grounding and correct grammar are enforced and guaranteed for, we will have a chance of optimizing language generation and comprehension directly on target tasks, which should result in more targeted and better-suited systems as opposed to training on auxiliary objectives.

In future work, conversation generation can also be targeted. The challenge there is that conversation should only be as informative as required in a given situation to not distract or cause an unnecessarily high cognitive load.

References

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

X. Chen, H. Fang, T.Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *Computing Research Repository*, abs/1504.00325.

B. Dai, S. Fidler, R. Urtasun, and D. Lin. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN. In *International Conference on Computer Vision*, pages 2989–2998, Venice, Italy.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami Beach, MI, USA.

J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. 2015. Language Models for Image Captioning: The Quirks and What Works. In *Conference on Empirical Methods in Natural Language Processing*, pages 100–105, Lisbon, Portugal.

J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. 2017. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691.

M. English and P. A. Heeman. 2005. Learning Mixed Initiative Dialog Strategies By Using Reinforcement Learning On Both Conversants. In *Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 1011–1018, Vancouver, BC, Canada.

F. Faghri, D. J. Fleet, R. Kiros, and S. Fidler. 2018. VSE++: Improved Visual-Semantic Embeddings. In *Proceeding of the British Machine Vision Conference*, Newcastle upon Tyne, UK.

A. Gatt and E. Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

E. J. Gumbel. 1954. *Statistical theory of extreme values and some practical applications; a series of lectures*. U.S. Government Printing Office, Washington, D.C., USA.

M. Hodosh, P. Young, and J. Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

MD. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6):118:1–118:36.

Y. Huang, W. Wang, and L. Wang. 2017. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7254–7262, Honolulu, HI, USA.

E. Jang, S. Gu, and B. Poole. 2016. Categorical Reparameterization by Gumbel-Softmax. *Computing Research Repository*, abs/1611.01144.

A. Karpathy and L. Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.

A. Karpathy, A. Joulin, and L. F. Fei-Fei. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Advances in Neural Information Processing Systems*, pages 1889–1897, Montréal, QC, Canada.

S. Kirby. 2007. The evolution of language. *Oxford Handbook of Evolutionary Psychology*, pages 669–681.

R. Kiros, R. Salakhutdinov, and R. S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *Computing Research Repository*, abs/1411.2539.

J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, TX, USA.

X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing. 2017. Recurrent Topic-Transition GAN for Visual Paragraph Generation. In *International Conference on Computer Vision*, pages 3382–3391, Venice, Italy.

T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. Microsoft COCO: Common Objects in Context. *Computing Research Repository*, abs/1405.0312.

X. Liu, H. Li, J. Shao, D. Chen, and X. Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the 15th European Conference on Computer Vision - ECCV 2018*, pages 353–369, Munich, Germany.

J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. 2014. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *Computing Research Repository*, abs/1412.6632.

V. Mathur and A. Singh. 2018. The rapidly changing landscape of conversational agents. *Computing Research Repository*, abs/1803.08419.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

M. A. Ranzato, S. Chopra, M. Auli, and W. Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, San Juan, Puerto Rico.

Z. Ren, H. Jin, Z. Lin, C. Fang, and A. Yuille. 2017a. Multiple instance visual-semantic embedding. In *Proceedings of the British Machine Vision Conference*, London, UK.

Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li. 2017b. Deep Reinforcement Learning-Based Image Captioning with Embedding Reward. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1151–1159, Honolulu, HI, USA.

J. R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417424.

R. Shetty, M. Rohrbach, L. A. Hendricks, M. F., and B. Schiele. 2017. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In *International Conference on Computer Vision*, pages 4155–4164, Venice, Italy.

L. Steels. 2015. *The Talking Heads experiment: Origins of words and meanings*. Language Science Press, Berlin, Germany.

R. Subhashini and V. J. S. Kumar. 2010. Evaluating the performance of similarity measures used in document clustering and information retrieval. In *International Conference on Integrated Intelligent Computing*, pages 27–31, Bangalore, India.

I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, Montréal, QC, Canada.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 49:433–460.

R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, Boston, MA, USA.

O. Vinyals and Q. V. Le. 2015. A Neural Conversational Model. *Computing Research Repository*, abs/1506.05869.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663.

L. Wang, Y. Li, and S. Lazebnik. 2017. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *Computing Research Repository*, abs/1704.03470.