Don't Mention the Shoe! A Learning to Rank Approach to Content Selection for Image Description Generation

Josiah Wang
Department of Computer Science
University of Sheffield
United Kingdom
j.k.wang@sheffield.ac.uk

Robert Gaizauskas
Department of Computer Science
University of Sheffield
United Kingdom
r.gaizauskas@sheffield.ac.uk

Abstract

We tackle the sub-task of content selection as part of the broader challenge of automatically generating image descriptions. More specifically, we explore how decisions can be made to select what object instances should be mentioned in an image description, given an image and labelled bounding boxes. We propose casting the content selection problem as a learning to rank problem, where object instances that are most likely to be mentioned by humans when describing an image are ranked higher than those that are less likely to be mentioned. Several features are explored: those derived from bounding box localisations, from concept labels, and from image regions. Object instances are then selected based on the ranked list, where we investigate several methods for choosing a stopping criterion as the 'cut-off' point for objects in the ranked list. Our best-performing method achieves state-of-the-art performance on the ImageCLEF2015 sentence generation challenge.

1 Introduction

In recent years, there has been significant interest in developing systems capable of generating literal, sentential descriptions of images (*a boy playing with a frisbee in the park*). The task poses an interesting and difficult challenge for natural language generation, and is important for improved text and image retrieval. The image description task could potentially advance research and provide insights into multimodal natural language generation, e.g. building language models of how humans naturally describe the visual world.

A standard paradigm for approaching this task is to first detect instances of pre-defined concepts in the image to be described, and then to reason about the detected concepts to generate image descriptions. Thus, such approaches may involve various components of a standard Natural Language Generation pipeline (Reiter and Dale, 2000), such as document planning (including content determination), microplanning (lexicalisation/referring expression generation) and realisation.

In this paper, we concentrate on a specific subproblem in such an image description generation pipeline. More specifically, we explore the *content selection* problem proposed by Wang and Gaizauskas (2015). In this setting, object instances are assumed to have already been localised in an image. Thus, given gold standard labelled bounding boxes of object instances in an image, the task is to select the appropriate bounding box instances to be mentioned in the eventual image description that is to be generated (see Figure 1 for an example). To our knowledge, there has been minimal work specifically tackling the content selection problem. However, the task is important to image description generation as not all entities depicted in an image will be mentioned by humans. For example, a fork lying on a table probably will not be mentioned in a picture of a family having dinner in the kitchen. Determining which entity will be described thus poses an interesting research question, and may provide insights into how humans decide what is important enough to be described in an image description.

Thus, the main objective of this paper is to propose methods for learning to predict the object entities depicted in an image that will be mentioned in a human-authored description of the image. Our main contribution is to develop a ranking-based content selection system that exploits stronger tex-

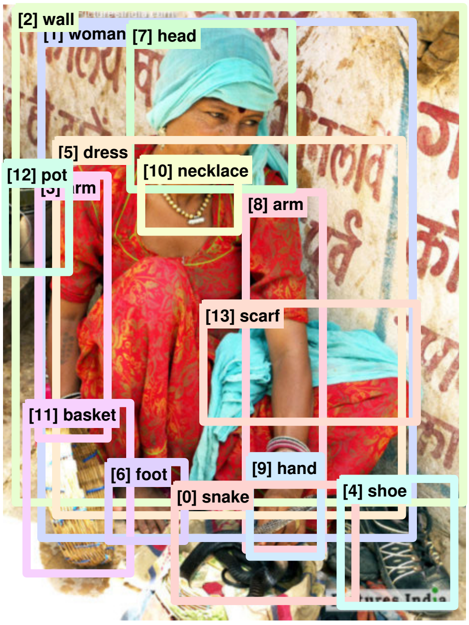

Figure 1: Given labelled bounding boxes as input, we tackle the *content selection* task, i.e. deciding which bounding box instances should be selected to be mentioned in the corresponding image description. This is an important task as humans do not mention everything that is depicted in an image. We propose casting the content selection problem as a ranking task, that is to order the bounding box instances by how likely they are to be mentioned in a human-authored image description.

tual and image features from data for the content selection problem, than those used in the baselines proposed in Wang and Gaizauskas (2015). We propose casting the content selection problem as a learning to rank problem. More specifically, given a set of labelled bounding boxes in an image, bounding boxes instances are ranked by how likely they are to be mentioned in a corresponding human description. However, as we are interested in both precision and recall, we do not require all labelled bounding boxes to be ranked; for example object instances that are unlikely to be mentioned in the description need not be ranked. Thus, we also propose various 'stopping criterion' to automatically select only relevant instances based on the rankings. Our hypothesis is that humans inherently prioritise important entities to be selected based on background knowledge and other cues, and we will thus be able to exploit this to tackle the content selection problem.

1.1 Overview

We discuss related work on the content selection problem in Section 2. In Section 3, we present our proposed approach to treat content selection as a learning to rank problem, discussing the formulation of the task (Section 3.1), features derived from bounding box localisations, concept labels and visual appearances (Section 3.2), and the various ranking algorithms explored (Section 3.3). In Section 3.4, we also propose some automatic stopping criteria to select important objects to be described from the ranking list. Experimental results are presented in Section 4, with regards to concatenating all features (Section 4.2) as well as treating individual features independently (Section 4.3). We also provide a summary of our feature ablation study in Section 4.4, and present conclusions in Section 5.

2 Related work

Image description generation. Various approaches have been proposed in the literature for the task of generation image descriptions, for example (Yao et al., 2010; Kulkarni et al., 2011; Yang et al., 2011; Mitchell et al., 2012; Karpathy and Fei-Fei, 2015; Donahue et al., 2015; Vinyals et al., 2015), among others. Most previous work concentrates on solving the problem 'end-to-end', that is to generate a description given an image as input. Such systems are also evaluated in an extrinsic manner, that is by comparing output image descriptions to multiply-annotated gold standard descriptions of the same image using global measures such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), Meteor (Denkowski and Lavie, 2014) or CIDEr (Vedantam et al., 2015). Whilst such evaluation methodologies are useful to evaluate image description generation systems as a whole (how similar is the generated description to human-authored descriptions?), they make it hard to identify which components of the generation process contribute to any performance gains or losses. Wang and Gaizauskas (2015) propose evaluating image description generation systems in a fine-grained manner, i.e. evaluating each component of the image description generation pipeline independently. To demonstrate this, they proposed the task of content selection as a precursor to generating image descriptions and performed fine-grained evaluation on this specific task.

Content selection. There has been some work on selecting objects that are important or interesting in an image. Elazary and Itti (2008) propose learning to predict object *interestingness* by the order in which objects are labelled by annotators in LabelMe. Spain and Perona (2010) propose learning to predict object importance, by asking multiple annotators (25 per image) to name 10 objects they see in each image. The annotations are then aggregated: important objects are those that are mentioned by many annotators.

Most related to our work is Berg et al. (2012), who explore factors (compositional, semantic, and contextual) that can be used to predict what is being described in an image. For prediction, they focus on a binary prediction problem – is this object described? yes or no? – and treat bounding boxes as independent of each other. In our case, we treat other bounding boxes as context, as a frequently occurring object may not be mentioned when co-occurring with some other object. Dodge et al. (2012) tackle an inverse problem: learning to predict segments of Flickr captions (noun phrases) that are 'visual', i.e. predicting whether a noun phrase in the caption is depicted in the image.

There has also been some work on measuring image memorability (what makes an image memorable to humans?), for example, Isola et al. (2011), among others. However, most work deals with memorability at image-level, rather than object level. Dubey et al. (2015) tackle image memorability at object level, that is, what objects are memorable (worth remembering) to a person in an image. This acts as a precursor to the content selection problem of choosing what to describe in an image description.

Ortiz et al. (2015) treat image description generation as a Statistical Machine Translation (SMT) task, and concentrate on describing abstract, clipart scenes. Part of their pipeline involves a content selection module where rankings of object *pairs* are optimised as an integer linear programming (ILP) problem, allowing object pairs that frequently co-occur and are close to each other to be ranked higher than those that are not. Our approach is not constrained to pairwise features, and automatically learns to optimise rankings across all *instances* directly from a training set, using arbitrary feature vectors.

Directly related to our work is Wang and Gaizauskas (2015), who propose some baselines for content selection assuming 'clean' visual input is provided in the form of bounding boxes labelled with concepts. The baselines are based on various textual and visual cues. We aim to move beyond these baselines and attempt to improve the performance of content selection on the same dataset used in their paper.

Learning to rank. Learning to rank is a problem common in the field of Information Retrieval. Many approaches have been proposed to learn to rank instances in a document in order of their relevance to a query. The approaches can generally be divided into three main groups:

- **Pointwise** ranking: Each instance in a document are treated independently of each other.
- **Pairwise** ranking: The relative rank of pairs of instances are optimized in the objective function.
- **Listwise** ranking: The rankings are optimised directly on the evaluation metric (e.g. normalied discounted cumulative gain (NDCG)).

We refer readers to Li (2011) for a summary of different techniques for learning to rank.

3 Learning to rank object instances

In this paper, we use the dataset from the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation challenge (Villegas et al., 2015; Gilbert et al., 2015). More specifically, we tackle the 'clean track' of the sentence generation task. In this track, participants are provided with images with bounding box instances labelled with a WordNet sysnet (from 251 possible synset categories). Each image also contains 5-51 corresponding descriptions per image. Each description has been annotated with the correspondence between a bounding box instance and a textual term in the description (e.g. "man" in description refers to bounding box instance 1 in the image). There are 500 development images and 450 test images. At test time, participants are provided labelled bounding boxes as input, and are asked to produce systems capable of selecting the bounding boxes that are mentioned in the human-authored descriptions.

3.1 Problem definition

Let $B^i = \{b^i_1, b^i_2, ..., b^i_k\}$ be the set of labelled bounding boxes for an image $i \in I$, where $b^i_j = (l^i_j, c^i_j)$, and l^i_j is the bounding box localisation (position and size), and $c^i_j \in C$ is the concept label for the bounding box j, and $|C| = 251$ is the number of pre-defined categories. Given the set of input bounding boxes B^i for each image i, the eventual task is to predict the set of bounding box instances that are most likely to be mentioned in the gold standard descriptions. Casting this as a ranking task, we aim to predict the relevance of the bounding boxes, i.e. most likely to be mentioned in the gold standard, and then rank the bounding box instances by their relevance.

As a learning to rank problem, our objective is to learn, from some training data, to predict the relevance of an unseen bounding box instance for a test image, given other bounding box instances of the same image as well as features x^i_j derived from each bounding box instance b^i_j.

3.2 Features

We explore different features, derived from (i) the bounding box localisation, l^i_j; (ii) the concept label, c^i_j; or (iii) the visual appearance of the region in image i bounded by l^i_j. The features we explore are:

- **bboxsize**: the area of the object bounding box relative to the image.
- **bboxdist**: distance of the centre of the object bounding box from the image centre. For this paper, we negate the distance to accommodate classifiers that assume positive linear relations.
- **textiv**: a 251 dimensional one-hot vector with 1 for the matching concept label and 0 for the others.
- **textemb**: a 300 dimensional synset embedding derived from word2vec pretrained on the Google News Dataset (Mikolov et al., 2013). As each concept label is a WordNet synset, we further fine-tuned the embeddings to obtain *synset embeddings* in the original word2vec embedding space with AutoExtend (Rothe and Schütze, 2015), where an autoencoder is learnt based on WordNet terms, lexemes and hypernym relations.
- **imgemb**: a 4,096 dimensional image embedding for the object region enclosed by the bounding box. For this paper we used the penultimate layer (FC7) of the 16-layer variant of VGGNet (VGG-16) (Simonyan and Zisserman, 2014). Intuitively, this feature represents the visual appearance of the region enclosed by the bounding box.

In early experiments, we experimented with using the absolute bounding box positions (x and y coordinates) as a features. However, these features yielded poor performance, and were thus discarded in subsequent experiments.

We also explore combining the features to examine the contribution of each feature, to determine which features play a role in the content selection task.

3.3 Ranking algorithms

For ranking, we consider several commonly used algorithms in the literature for Learning to Rank. We select one example from each of the group of approaches (pointwise, pairwise, listwise):

- **rforest**: Random forests (Breiman, 2001), an algorithm using *pointwise* ranking. We use the implementation of random forests in RankLib[1] in this paper.
- **svmrank**: Ranking SVM (Joachims, 2002), an algorithm using *pairwise* ranking. We use the SVM^{rank} implementation (Joachims, 2006) of Ranking SVM in this paper. A linear kernel is used for this paper. [2]
- **cascent**: Coordinate ascent (Metzler and Croft, 2007), an algorithm using *listwise* ranking. In our paper, we optimise the rankings using NDCG@10 as a metric. Again, we use the implementation of coordinate ascent in RankLib.

For these algorithms, we compute the relevance score for each bounding box instance as the proportion of human-authored, gold standard descriptions that mention the concept. The task is to learn to predict the relevance score given the features in Section 3.2, and subsequently rank the bounding box instances for each image by this score. As

[1] http://www.lemurproject.org/ranklib.php

[2] We have experimented with an RBF kernel, but found the results comparable to a linear kernel.

such, this task is treated as a continuous regression problem.[3]

Our intuition is that pairwise and listwise ranking algorithms would suit our task better than pointwise algorithms, as pairwise/listwise ranking implicitly considers all other object instances as context rather than treating each instance independently as in pointwise ranking. For example, a *table* might be important and frequently mentioned, but might not be mentioned when co-occurring with *kitchen*.

3.4 Stopping criteria

While the ranking process will result in a ranked list of *all* input object instances per images, there is a need to provide a cut-off point in the rankings for the eventual task of content selection.

From our initial experiments, we found that the number of selected object instances greatly affects the F-scores (see Section 4.1 for evaluation measure). Selecting fewer good object instances per image will raise precision at the expense of lower recall, while selecting more objects will increase recall at the expense of lower precision. Wang and Gaizauskas (2015) propose a fixed threshold for the maximum number of object instances to be selected, and found that selecting 3 to 4 object instances yields an optimal balance between precision and recall (the mean number of unique bounding box instances per description is 2.89 in the development dataset). However, it may be more beneficial to have a variable threshold across images depending on the number of input object instances. For example, the bigram-based feature proposed in Wang and Gaizauskas (2015) has an internal stopping criterion, resulting in higher overall precision when compared to other fixed length features.

Motivated by the high precision scores of the aforementioned system, in this paper we propose two variable stopping criteria:

- **absolute**: Retaining only object instances with a predicted relevance score above a certain threshold.
- **relative**: Setting the cut-off point at the largest *difference* in relevance scores.

In the former case (**absolute**), we first normalise the predicted score across bounding boxes per image, where the highest-ranked bounding box is assigned a score of 1 and the lowest-ranked a score of 0. We retain only bounding box instances where the normalised predicted score is above a threshold (0.5 in our experiments).

The motivation for the latter case (**relative**) stems from our observation that the relevance scores in the development set reduces dramatically once the most important object instances are selected. For example, the most relevant object instances may have a relevance score of 0.9 and 0.8 followed by 0.2. Thus, a suitable cut-off point would be between 0.8 and 0.2. Cutting off at the point that immediately precedes the biggest difference in scores (after 0.8 in the example above) we refer to as **relative1** in our experiments. We also found that cutting off the ranked list after the point that *follows* the largest difference in score (after 0.2 in the example above) produces a marginally higher F-score (increased recall at the expense of precision). We therefore also report the results for this as a variant, which we refer to as **relative2**.

4 Experimental results

4.1 Evaluation measure

Following the convention of the ImageCLEF2015 Sentence Generation challenge, we evaluate content selection using the fine-grained evaluation metric proposed in Wang and Gaizauskas (2015) and Gilbert et al. (2015). More specifically, we measure the F-score (including Precision and Recall) when comparing the object instances selected by our system to the object instances mentioned in the gold standard human-authored image descriptions. The human upper-bound is estimated by evaluating one description against the other descriptions of the image and repeating the process for all descriptions.

We compare our results to the winning participants of past ImageCLEF challenges. **RUC 2015** (Li et al., 2015) achieved the best performance in the 2015 edition (Villegas et al., 2015; Gilbert et al., 2015) with high precision, but used an external image description dataset to train their joint CNN-LSTM image captioning system, and performed content selection in a retrospective manner. **DUTh 2016** (Barlas et al., 2016) achieved the best performance (high recall) in the 2016 edition (Villegas et al., 2016; Gilbert et al., 2016),

[3] We also experimented with ordinal regression, where regression scores are partitioned into a set of integers {0,1,2,3,4} based on the relevance score (with 4 being the most relevant). We found performance to be lower, in general. Thus, we only report results for continuous regression.

	Stopping Criterion	P	R	F
	RUC 2015	0.68 ± 0.30	0.48 ± 0.24	0.53 ± 0.23
	DUTh 2016	0.45 ± 0.17	0.79 ± 0.20	0.55 ± 0.15
	W&G 2015	0.59 ± 0.19	0.58 ± 0.22	0.56 ± 0.18
cascent	**k = 3**	0.59 ± 0.22	0.56 ± 0.23	0.55 ± 0.20
	k = 4	0.50 ± 0.20	0.63 ± 0.22	0.54 ± 0.17
	absolute	0.42 ± 0.22	0.72 ± 0.22	0.49 ± 0.17
	relative1	0.72 ± 0.33	0.57 ± 0.29	0.53 ± 0.22
	relative2	0.56 ± 0.25	0.66 ± 0.26	0.54 ± 0.20
svmrank	**k = 3**	0.60 ± 0.20	0.59 ± 0.22	0.57 ± 0.18
	k = 4	0.53 ± 0.18	0.68 ± 0.21	0.58 ± 0.16
	absolute	0.43 ± 0.20	0.80 ± 0.19	0.52 ± 0.15
	relative1	0.67 ± 0.31	0.61 ± 0.29	0.53 ± 0.19
	relative2	0.55 ± 0.25	0.70 ± 0.25	0.55 ± 0.18
rforest	**k = 3**	0.69 ± 0.18	0.68 ± 0.21	0.66 ± 0.16
	k = 4	0.60 ± 0.17	0.76 ± 0.19	0.65 ± 0.14
	absolute	0.84 ± 0.19	0.64 ± 0.21	$\underline{\mathbf{0.70 \pm 0.16}}$
	relative1	0.89 ± 0.18	0.57 ± 0.23	0.66 ± 0.18
	relative2	0.71 ± 0.18	0.69 ± 0.21	0.68 ± 0.17
	Human	0.77 ± 0.11	0.77 ± 0.11	0.74 ± 0.12

Table 1: Results of combining all features: Mean Precision, Recall and F-score (with standard deviations) for different algorithms and stopping criteria, compared to the winning ImageCLEF participants (**RUC 2015** and **DUTh 2016**), the best reported results of Wang and Gaizauskas (2015) (**W&G 2015**) and a human upper-bound.

using a binary SVM classifier with bounding box localisation and visual features. We also compare our performance to the best reported results in Wang and Gaizauskas (2015) (**W&G 2015**), namely by combining bigram and bounding box size priors with a stopping criterion of $k = 3$.

4.2 Combining features

We first report the results of concatenating all features (Section 3.2) as a single vector, and compare the performance of the various ranking algorithms (Section 3.3) and stopping criteria (Section 3.4). The intuition is that the ranking algorithm will perform automatic feature selection to select the most discriminative features useful for predicting the relevance score.

Table 1 shows the results of using a combination of all features. The pointwise ranking based Random Forests classifier performs best overall, achieving an F-score of 0.70, close to the human upper-bound of 0.74. This significantly exceeds the previous state-of-the-art result on the same training and test data of $F = 0.56$, as reported in Wang and Gaizauskas (2015). The coordinate ascent ranker and Ranking SVM achieved comparable scores, the latter perhaps having a slight edge.

	Stopping Criterion	P	R	F
cascent	**k = 3**	0.63 ± 0.21	0.62 ± 0.21	0.60 ± 0.17
	k = 4	0.55 ± 0.19	0.69 ± 0.21	0.59 ± 0.16
	absolute	0.54 ± 0.22	0.71 ± 0.20	0.58 ± 0.15
	relative1	0.84 ± 0.25	0.57 ± 0.24	0.61 ± 0.18
	relative2	0.63 ± 0.22	0.66 ± 0.23	0.61 ± 0.17
svmrank	**k = 3**	0.65 ± 0.19	0.64 ± 0.22	0.62 ± 0.17
	k = 4	0.57 ± 0.18	0.72 ± 0.21	0.61 ± 0.15
	absolute	0.81 ± 0.24	0.55 ± 0.23	0.62 ± 0.18
	relative1	0.85 ± 0.24	0.51 ± 0.23	0.59 ± 0.17
	relative2	0.69 ± 0.21	0.65 ± 0.22	0.64 ± 0.18
rforest	**k = 3**	0.69 ± 0.18	0.68 ± 0.20	0.66 ± 0.16
	k = 4	0.60 ± 0.17	0.75 ± 0.19	0.64 ± 0.14
	absolute	0.83 ± 0.19	0.66 ± 0.21	$\underline{\mathbf{0.71 \pm 0.16}}$
	relative1	0.88 ± 0.18	0.59 ± 0.23	0.67 ± 0.18
	relative2	0.70 ± 0.17	0.70 ± 0.21	0.68 ± 0.15

Table 2: Results of combining features derived from bounding box localisation and concept labels (excluding image region features). In contrast to Table 1, excluding image region features improves the performance of both **cascent** and **svmrank**.

The performance of the various stopping criteria seems to be dependent on the ranking algorithm. The **absolute** stopping criterion seems to be sensitive to the type of ranking algorithm. As expected, **relative1** achieved higher precision than **relative2**, whereas **relative2** achieved better recall with the additional object instance being selected.

In an earlier experiment, we have explored combining only features derived from bounding box localisation and concept labels, excluding image region features (**imgemb**). Interestingly, we found better performance by excluding image region features for **cascent** and **svmrank**, but not much difference for **rforest** (compare Table 1 and Table 2). This is very likely because the high dimensional image features (4,096D) dominated the ranking decisions for these rankers, compared to **rforest** which seemed less affected by the imbalance. The performance of **cascent** and **svmrank** in Table 1 is similar to that of using only image region features (c.f. Table 5, to be discussed later), further confirming our suspicion.

4.3 Individual features

We now explore each feature individually to investigate the contributions of each. Table 3 shows the results for the features derived from bounding box localisation (**bboxsize** and **bboxdist**). The same scores are obtained from both **cascent** and **svmrank**, possibly because both these features are single dimensional vectors. **rforest** requires higher dimensionality to operate, and as such is unable

	Stopping Criterion	P	R	F
bboxsize	**k = 3**	0.53 ± 0.20	0.55 ± 0.26	0.53 ± 0.21
	k = 4	0.50 ± 0.16	0.66 ± 0.24	0.55 ± 0.17
	absolute	0.56 ± 0.28	0.44 ± 0.28	0.46 ± 0.25
	relative1	0.56 ± 0.34	0.36 ± 0.29	0.40 ± 0.27
	relative2	0.54 ± 0.22	0.51 ± 0.28	0.49 ± 0.22
bboxdist	**k = 3**	0.39 ± 0.22	0.40 ± 0.27	0.38 ± 0.23
	k = 4	0.36 ± 0.18	0.48 ± 0.28	0.39 ± 0.21
	absolute	0.32 ± 0.19	0.71 ± 0.20	0.41 ± 0.16
	relative1	0.40 ± 0.30	0.64 ± 0.32	0.40 ± 0.21
	relative2	0.34 ± 0.21	0.69 ± 0.31	0.39 ± 0.19

Table 3: Mean Precision, Recall and F-score for features derived from bounding box localisation. Both **cascent** and **svmrank** return the same scores (shown). **rforest** is unable to handle single dimensional vectors. The results for **k=3** and **k=4** are comparable to Wang and Gaizauskas (2015).

to handle these one-dimensional features. The results are consistent with what was reported by Wang and Gaizauskas (2015) – that whilst both **bboxdist** and **bboxsize** show that content selection is dependent on these features, **bboxsize** is a better predictor for an object being selected compared to **bboxdist**.[4]

Table 4 shows the results for features derived from concept labels (**textiv** and **textemb**). For these three rankers, **textemb** seems to outperform **textiv**. The only exception is for **cascent** when the stopping criterion is **absolute**, where **textiv** seemed to give better precision than **textemb**. Comparing Table 3 and Table 4, we can see that features derived from concept labels are stronger predictors for content selection.

Table 5 shows the results of using only image region features (**imgemb**). Here, **cascent** does not perform as well as **svmrank** and **rforest**, due to the high dimensionality of the CNN embeddings. The performance of image region features seem to be on par with features derived from concept labels (Table 4), and better than bounding box features (Table 3). Noteworthy is how image region features yield higher recall than other features in general, at the expense of lower precision.

4.4 Feature ablation

We also performed a feature ablation study to gain insights into which features are important to content selection and the interaction between the features. This is done by testing different combinations of features to investigate which features contribute better to the overall performance and thus play a bigger role for content selection.

	Stopping Criterion	P	R	F
cascent textiv	**k = 3**	0.61 ± 0.22	0.59 ± 0.22	0.58 ± 0.19
	k = 4	0.53 ± 0.20	0.66 ± 0.22	0.57 ± 0.17
	absolute	0.54 ± 0.30	0.76 ± 0.20	0.55 ± 0.19
	relative1	0.58 ± 0.36	0.69 ± 0.28	0.48 ± 0.18
	relative2	0.48 ± 0.29	0.78 ± 0.23	0.51 ± 0.20
cascent textemb	**k = 3**	0.60 ± 0.21	0.59 ± 0.21	0.57 ± 0.18
	k = 4	0.52 ± 0.19	0.65 ± 0.21	0.56 ± 0.17
	absolute	0.36 ± 0.19	0.79 ± 0.19	0.46 ± 0.15
	relative1	0.59 ± 0.37	0.71 ± 0.27	0.50 ± 0.21
	relative2	0.45 ± 0.26	0.76 ± 0.25	0.49 ± 0.20
svmrank textiv	**k = 3**	0.60 ± 0.22	0.58 ± 0.22	0.57 ± 0.19
	k = 4	0.53 ± 0.19	0.68 ± 0.21	0.57 ± 0.16
	absolute	0.70 ± 0.32	0.60 ± 0.27	0.54 ± 0.16
	relative1	0.71 ± 0.33	0.59 ± 0.27	0.53 ± 0.17
	relative2	0.57 ± 0.26	0.69 ± 0.25	0.55 ± 0.18
svmrank textemb	**k = 3**	0.60 ± 0.21	0.58 ± 0.22	0.57 ± 0.18
	k = 4	0.51 ± 0.20	0.64 ± 0.21	0.55 ± 0.17
	absolute	0.77 ± 0.28	0.56 ± 0.23	0.59 ± 0.18
	relative1	0.82 ± 0.26	0.52 ± 0.23	0.58 ± 0.18
	relative2	0.63 ± 0.22	0.62 ± 0.23	0.60 ± 0.18
rforest textiv	**k = 3**	0.64 ± 0.21	0.63 ± 0.22	0.61 ± 0.18
	k = 4	0.56 ± 0.19	0.70 ± 0.21	0.60 ± 0.16
	absolute	0.79 ± 0.23	0.62 ± 0.22	0.66 ± 0.19
	relative1	0.84 ± 0.23	0.57 ± 0.23	0.64 ± 0.20
	relative2	0.66 ± 0.19	0.67 ± 0.21	0.64 ± 0.17
rforest textemb	**k = 3**	0.65 ± 0.20	0.64 ± 0.22	0.62 ± 0.18
	k = 4	0.57 ± 0.19	0.71 ± 0.21	0.61 ± 0.16
	absolute	0.78 ± 0.23	0.64 ± 0.21	0.67 ± 0.18
	relative1	0.84 ± 0.22	0.58 ± 0.23	0.65 ± 0.19
	relative2	0.67 ± 0.19	0.68 ± 0.21	0.65 ± 0.17

Table 4: Mean Precision, Recall and F-score for features derived from concept labels (one-hot indicator vectors and text embeddings).

Because of space constraints, we only provide a summary of interesting observations. Table 6 shows the F-scores for the **rforest** ranker with the **absolute** stopping criterion. We found that the features based on concept labels are dominant and influential in our experiments compared to those based on bounding box localisation or visual appearances. Combining **textiv** and **textemb** alone already yielded an F-score of 0.67. This demonstrates that semantic concept labels are the best predictors for content selection. Adding **bboxsize** to **imgemb** improves the F-scores marginally, suggesting that the object size does play some role on top of visual appearances in selecting important objects. We also found that for **rforest** rankers, **textemb** plays a larger role in predicting content selection compared to **textiv**, as evidenced by a greater drop in F-scores when omitting **textemb** compared to **textiv**.

[4] This was demonstrated in the errata provided by Wang and Gaizauskas (2015) after the paper was published.

	Stopping Criterion	P	R	F
cascent imgemb	**k = 3**	0.50 ± 0.23	0.47 ± 0.24	0.47 ± 0.21
	k = 4	0.45 ± 0.19	0.55 ± 0.24	0.48 ± 0.19
	absolute	0.29 ± 0.14	0.80 ± 0.22	0.40 ± 0.14
	relative1	0.39 ± 0.30	0.73 ± 0.32	0.39 ± 0.18
	relative2	0.34 ± 0.22	0.79 ± 0.29	0.40 ± 0.17
svmrank imgemb	**k = 3**	0.60 ± 0.20	0.59 ± 0.22	0.57 ± 0.18
	k = 4	0.53 ± 0.18	0.67 ± 0.21	0.57 ± 0.16
	absolute	0.43 ± 0.20	0.80 ± 0.19	0.52 ± 0.15
	relative1	0.66 ± 0.31	0.61 ± 0.29	0.53 ± 0.20
	relative2	0.54 ± 0.25	0.69 ± 0.26	0.54 ± 0.19
rforest imgemb	**k = 3**	0.60 ± 0.20	0.59 ± 0.22	0.58 ± 0.18
	k = 4	0.53 ± 0.18	0.67 ± 0.22	0.57 ± 0.16
	absolute	0.47 ± 0.19	0.76 ± 0.20	0.55 ± 0.15
	relative1	0.64 ± 0.29	0.62 ± 0.28	0.55 ± 0.19
	relative2	0.52 ± 0.22	0.69 ± 0.26	0.55 ± 0.17

Table 5: Mean Precision, Recall and F-score for features derived from image region features (image embeddings).

4.5 Discussion

We observed that the pointwise-based random forests ranker performs better than the pairwise and listwise-based rankers. This is surprising as we expected either pairwise- or listwise-based rankers to perform better than pointwise-based rankers, which treat each instance in a document as independent without considering other instances within the same document. It still remains unclear whether this is due to the random forests classifier itself being strong or that context plays a lesser role in content selection for this particular dataset. Further work is required to ascertain this.

5 Conclusion

We explored the content selection problem of deciding what needs to be mentioned in the description of an image, given labelled bounding boxes as input. We proposed casting the problem as a learning to rank task, where object instances that are more likely to be mentioned in human-authored descriptions are ranked higher than those less likely to be mentioned. Several features are explored: those derived from bounding box localisations, concept labels and visual appearances for each object instance. We also proposed methods to automatically estimate a cut-off point in each ranked list, to select only object instances that are likely to be mentioned in the image description.

Feature					**rforest**
bboxdist	**bboxsize**	**textiv**	**textemb**	**imgemb**	F
✓					-
	✓				-
		✓			0.66 ± 0.19
			✓		0.67 ± 0.18
				✓	0.55 ± 0.15
✓	✓				-
✓		✓			0.66 ± 0.17
✓			✓		0.69 ± 0.16
✓				✓	0.55 ± 0.15
	✓	✓			0.67 ± 0.18
	✓		✓		0.70 ± 0.16
	✓			✓	0.57 ± 0.16
		✓	✓		0.67 ± 0.18
		✓		✓	0.62 ± 0.16
			✓	✓	0.70 ± 0.16
✓	✓	✓			0.67 ± 0.17
✓	✓		✓		0.70 ± 0.16
✓	✓			✓	0.57 ± 0.15
✓		✓	✓		0.69 ± 0.16
✓		✓		✓	0.63 ± 0.16
✓			✓	✓	0.69 ± 0.16
	✓	✓	✓		0.70 ± 0.16
	✓	✓		✓	0.64 ± 0.16
	✓		✓	✓	0.70 ± 0.17
		✓	✓	✓	0.69 ± 0.16
✓	✓	✓	✓		0.71 ± 0.16
✓	✓	✓		✓	0.64 ± 0.17
✓	✓		✓	✓	0.70 ± 0.17
✓		✓	✓	✓	0.69 ± 0.16
	✓	✓	✓	✓	0.70 ± 0.16
✓	✓	✓	✓	✓	0.70 ± 0.16

Table 6: Results of the feature ablation test: mean F-scores for **rforest** with the **absolute** stopping criterion, for various combinations of features. Some results are omitted because **rforest** does not work well with single or two dimensional features.

Our method showed excellent results, achieving the state-of-the-art F-score of 0.70 on the ImageCLEF2015 content selection dataset, substantially out-performing the highest figures previously reported on this test set. We also found that for the proposed features, those that are derived from the concept labels are better predictors for the content selection task than those derived from bounding box localisations or visual appearance of regions.

The proposed learning to rank approach is general enough and may also be relevant to content selection tasks in other areas of natural language generation. Future work could include exploring even stronger features. There is also scope to automatically gather a larger noisy dataset to enable more robust learning and reduce reliance on annotating training data. We hope that these additions will further improve the content selection capabilities of the proposed system.

6 Acknowledgements

The authors acknowledge funding from the ERA-NET CHIST-ERA Visual Sense (ViSen) project – UK EPSRC grant reference: EP/K019082/1.

References

Georgios Barlas, Maria Ntonti, and Avi Arampatzis. 2016. DUTh at the ImageCLEF 2016 Image Annotation Task: Content Selection. In *CLEF2016 Working Notes*, CEUR Workshop Proceedings, Évora, Portugal, September. CEUR-WS.org.

Alexander C. Berg, Tamara L. Berg, Hall Daumé III, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, and Kota Yamaguchi. 2012. Understanding and predicting importance in images. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32, October.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C. Berg, and Tamara L. Berg. 2012. Detecting visual text. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. 2015. What makes an object memorable? In *Proceedings of the IEEE International Conference on Computer Vision*.

Lior Elazary and Laurent Itti. 2008. Interesting objects are visually salient. *Journal of Vision*, 8(3:3):1–15, Mar.

Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Emmanuel Dellandrea, Robert Gaizauskas, Mauricio Villegas, and Krystian Mikolajczyk. 2015. Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In *CLEF2015 Working Notes*, CEUR Workshop Proceedings, Toulouse, France, September 8-11. CEUR-WS.org.

Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Arnau Ramisa, Emmanuel Dellandrea, Robert Gaizauskas, Mauricio Villegas, and Krystian Mikolajczyk. 2016. Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task. In *CLEF2016 Working Notes*, CEUR Workshop Proceedings, Évora, Portugal, September 5-8. CEUR-WS.org.

Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable? In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, pages 145–152.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Xirong Li, Qin Jin, Shuai Liao, Junwei Liang, Xixi He, Yu-Jia Huo, Weiyu Lan, Bin Xiao, Yanxiong Lu, and Jieping Xu. 2015. RUC-Tencent at ImageCLEF 2015: Concept Detection, Localization and Sentence Generation. In *CLEF2015 Working Notes*, CEUR Workshop Proceedings, Toulouse, France, September 8-11. CEUR-WS.org.

Hang Li. 2011. A short introduction to learning to rank. *IEICE Transactions*, 94-D(10):1854–1862.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

Donald Metzler and W. Bruce Croft. 2007. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daume III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France, April. Association for Computational Linguistics.

Luis Gilberto Mateos Ortiz, Clemens Wolff, and Mirella Lapata. 2015. Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1505–1515, Denver, Colorado, May–June. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL)*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Merrielle Spain and Pietro Perona. 2010. Measuring and predicting object importance. *International Journal of Computer Vision*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, June.

Mauricio Villegas, Henning Müller, Andrew Gilbert, Luca Piras, Josiah Wang, Krystian Mikolajczyk, Alba García Seco de Herrera, Stefano Bromuri, M. Ashraful Amin, Mahmood Kazi Mohammed, Burak Acar, Suzan Uskudarli, Neda B. Marvasti, José F. Aldana, and María del Mar Roldán García. 2015. General Overview of ImageCLEF at the CLEF 2015 Labs. In Josiane Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth J. F. Jones, Eric San Juan, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9283 of *Lecture Notes in Computer Science*, pages 444–461. Springer International Publishing.

Mauricio Villegas, Henning Müller, Alba García Seco de Herrera, Roger Schaer, Stefano Bromuri, Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Arnau Ramisa, Emmanuel Dellandrea, Robert Gaizauskas, Krystian Mikolajczyk, Joan Puigcerver, Alejandro H. Toselli, Joan-Andreu Sánchez, and Enrique Vidal. 2016. General Overview of ImageCLEF at the CLEF 2016 Labs. In Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9822 of *Lecture Notes in Computer Science*. Springer International Publishing.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

Josiah Wang and Robert Gaizauskas. 2015. Generating image descriptions with gold standard visual inputs: Motivation, evaluation and baselines. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 117–126, Brighton, UK, September. Association for Computational Linguistics.

Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 444–454. Association for Computational Linguistics.

Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song Chun Zhu. 2010. I2T: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508.