Utilising Information Foraging Theory for User Interaction with Image Query Auto-Completion

Amit Kumar Jaiswal(✉), Haiming Liu, and Ingo Frommholz

Institute for Research in Applicable Computing, University of Bedfordshire, Luton, UK
{amitkumar.jaiswal,haiming.liu,ingo.frommholz}@beds.ac.uk

Abstract. Query Auto-completion (QAC) is a prominently used feature in search engines, where user interaction with such explicit feature is facilitated by the possible automatic suggestion of queries based on a prefix typed by the user. Existing QAC models have pursued a little on user interaction and cannot capture a user's information need (IN) context. In this work, we devise a new task of QAC applied on an image for estimating patch (one of the key components of Information Foraging Theory) probabilities for query suggestion. Our work supports query completion by extending a user query prefix (one or two characters) to a complete query utilising a foraging-based probabilistic patch selection model. We present iBERT, to fine-tune the BERT (Bidirectional Encoder Representations from Transformers) model, which leverages combined textual-image queries for a solution to image QAC by computing probabilities of a large set of image patches. The reflected patch probabilities are used for selection while being agnostic to changing information need or contextual mechanisms. Experimental results show that query auto-completion using both natural language queries and images is more effective than using only language-level queries. Also, our fine-tuned iBERT model allows to efficiently rank patches in the image.

Keywords: Query auto completion · Interactive information retrieval · Information Foraging Theory

1 Introduction

Query auto-completion (QAC) is an action of signalling full queries once the user starts typing a prefix of a few characters that eases user query compositions [4]. It is also termed as (dynamic) query suggestion [17], query completion [35] and real-time query expansion [37]. Popular features such as QAC make people more dependent on search engines to find any relevant information. However, such kind of factor lets users express their queries only ambiguously, which are then overly vague to be completely interpreted by search engines. This makes query auto-completion a bottleneck construct in the usability of search engines [5].

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12035, pp. 666–680, 2020.
https://doi.org/10.1007/978-3-030-45439-5_44

Also, users often apply several rounds of search to reformulate their queries further to adhere to their information needs given they find some relevant results. Past work [6,20] demonstrated the use of information scent to model users' information need during web search, and it has been used to understand the factors affecting search and what takes a user to stop the search. Despite the good observation, the exploitation of information scent (from Information Foraging Theory [27]) is under-explored in case of ambiguous queries and have not been extended to take into account an image in query expansion (or suggestion) tasks. For the users' convenience, current search engines generally endue query suggestions for them in order to describe their queries more explicitly. They have been explored extensively in query auto-completion tasks, especially the traditional approach known as Most Popular Completion (MPC) [3] which at the extreme is incapable of anticipating a query it has never seen before. Solutions further improved by recent semantically-driven models [23,24] and neural model [26] approaches which are the current state-of-the art in QAC. However, most of the language embedding models [13] have obtained strong results on multiple benchmarks for understanding the polarity of word compositions. Unsupervised pre-trained natural language embeddings [7,21] successfully model long term dependencies with the purpose of predicting masked terms and assessing if sentences ensue one another, which showed strong results on several natural language processing and information retrieval tasks. Empirically, recent advances in sequence models have been adapted to span a prefix to full text and index [12] but despite the attainment, it has not been generalised to take an image into account. Also, deep neural networks are mature enough and capable of segmenting regions within an image [9,10].

To address the above mentioned gaps, we move one step forward to present a method that extends and modifies the state-of-the-art approaches in query completion and text embedding. We apply our ideas to an image search scenario where we assume patches are regions of images that are relevant to the user's information need. Our work is concerned with providing users of image search engines with a useful query suggestion (via a visually-oriented patch form) during interaction, to further amplify their exploratory search experience. Hence, finding useful patches for query expansion in an image based on textual queries (or descriptions) is the primary focus of our work. Past work [11,30] used both the query and image for typical retrieval and segmentation tasks. In our task formulation, we rely only upon a given arbitrary text prefix rather than having the entire text query which is used to perform search based on the image and supported by a modified deep language model [12] to find the most relevant patch in the image. We break down the task into three sub-tasks: (a) completing the query from user query prefix and an image; (b) finding patch probabilities based on the complete user query, and (c) aligning and segmenting all patches in the image. We summarise our contributions of this paper as follows:

1. To the best of our knowledge, we are the first to present a method for image query auto-completion where a user query prefix is adapted upon an image.

2. We elaborate the analogy of query auto-completion based on Information Foraging Theory and propose an explainable strategy for the observed challenges of query formulation and the varying users' information need.
3. We propose iBERT inspired by [7] to compute probabilities of patches and rank them efficiently in the image.

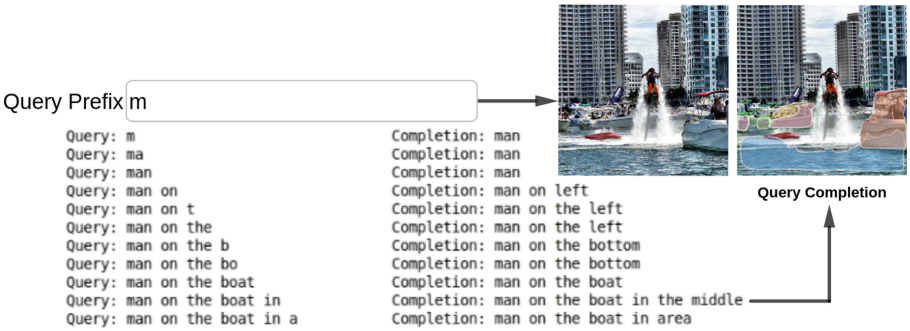

Fig. 1. Query auto-completion using our extended LSTM language model

2 Related Work

This section details a brief overview of query auto-completion, image search suggestion, Information Foraging Theory and BERT pre-trained language embedding model. We will investigate the latter approach experimentally in the following section.

Query Auto-Completion: Query auto-completion is an important aspect for information retrieval systems which allow it to predict what could be the next character (or query item) right after the first key was pressed by a user. The predictions in IR systems are generally driven by the query logs (or query history) which are the factual queries that users have previously entered as they were trying to satisfy their information need [14,37]. [3] introduced a method called NearestCompletion that addresses the situation of "context" which depicts the users' preceding queries in suggestion-based IR systems. The authors' proposed MPC mechanism relies on the entire popularity of the queries conforming to the provided prefix. Recent work reported in [15] studies user reformulation behaviour by leveraging textual features, whereas [31] introduced personalised query auto-completion and found that utilising a user's long-term search logs and locations as well as both context-based textual features and demographic features is more effective. More recent advances in QAC using neural language models are proposed in [26] using recurrent neural networks that effectuate the performance on immediately unseen queries. A generalised and adaptable language model for personalised QAC is introduced in [12]. We extend this adaptable language model to query completion in an image search scenario in the following section.

Query Suggestion in Image Search: Query suggestion and query completion differs in their end goal in which the former search aspect outputs a list of ranked queries against an input query, whereas the latter search aspect outputs queries with the first few characters (or text) similar to the user's input. Recent work [39] introduced a learning-based personalised suggestion framework for query suggestion which uses both visual and textual queries. Their work uses users' click-through data. A new paradigm of attention-based mechanisms for referring expressions in image segmentation [30] is proposed which contains a keyword-aware network and query attention model that demonstrates the relationships with various image regions for a given query. Inspired by the idea of attention models, we modify this mechanism for patch alignments within images via information scent in the following section.

Information Foraging Theory: Information Foraging Theory (IFT) [27] is a theoretical framework for understanding information access behaviour, derived from the ecological science concept of optimal foraging theory which applies to how humans access information. IFT stands on three different models, namely information scent model, information patch model and information diet model, which can illustrate users' search preferences and behaviours [19]: (1) The information within a certain environment scattered in form of *patches* (images, text snippets, documents) consisting of *information features* (colors, words) refers to the *information patch model*; (2) A user can go from one patch to another via a *cue* (e.g., typing a query by following perceptual or heuristic cues [32]), which meets the user's information need. The goal of such cues is to characterise the contents that will be envisaged by trailing the links, which refers to *the information scent model*; (3) Different types of information sources will vary in their information access costs. Users will assess the information sources based on information gain per unit cost or varied profitability, and then the users will narrow or expand diversities of information sources based on their profitability. This user behaviour refers to *the information diet model*.

One of the main IFT concepts are *information patches*. For instance, sections and their associated features in search engine results can be considered patches. From a foraging perspective in image search, the searcher is the predator (or forager [38]), the information patch is any segment or a region within an image (or image itself) in a given information environment. The piece of information a user is looking for is the prey, and the consumed (or gained) information is the information diet. Something on the user interface that informs users about a specific place they should look next is referred to as a *cue* of the information scent.

Language Embeddings: Nowadays, many information retrieval or natural language processing tasks rely on language embeddings, such as word2vec [22], Glove[1], and fastText[2]. They use vector word embeddings for word representation to transform a distinct space of human language into a continuous space,

[1] https://nlp.stanford.edu/projects/glove/.
[2] https://fasttext.cc.

which will be further processed usually through a neural network. In query auto-completion, embeddings have been employed for distributed representation of queries based on a convolutional latent semantic model [23]. Word embeddings have been used to compute query similarity for query auto-completion [29], incorporating the features with the Most Popular Completion model. Very recent work [7] introduced a pre-trained deep language model known as BERT which has shown promising results on several IR and natural language processing tasks. However, it is still not well-explored how to leverage such pre-trained language models for QAC, which poses certain challenges both regarding the task and training. Based on this work, we describe our proposed BERT-based model for computing patch probabilities in the following section.

3 Our Model

Let a set of patches $p_k \in P$, where P is the complete set of recognisable patch classes, be given. The user inputs a query prefix q_p, an incomplete query to retrieve an image I. With the given q_p, we auto-complete the expected query q. We formulate the auto-completion query task as the probability maximisation of a given query adapted on an image as shown in Eq. (1)

$$q_{a^*} = \underset{q}{\operatorname{argmax}}\, P(q|q_p, I) = \underset{\{t_1 t_2 ... t_n\}}{\operatorname{argmax}}\, P(t_1 t_2 ... t_n|q_p, I) \tag{1}$$

where q_{a^*} is the adapted query on an image, $t_i \in S$ is the term in position i in a sequence S.

We consider the task of estimating patch probabilities provided an auto-completion query q_{a^*} as a multi-label problem where each class of patches can independently exist. Let $P_{q_{a^*}}$ be the set of patches attributed to in q_{a^*}. As $\hat{q_{p_k}}$ is the estimate of $P(p_k \in P_{q_{a^*}})$ and $y_k = \mathbb{1}[p_k \in P_{q_{a^*}}]$, the sigmoid cross entropy loss function is minimized by the patch selection model:

$$\mathcal{L}_{f_{selection}} = -\sum_k y_k \log(\hat{q_{p_k}}) + (1 - y_k) \log(1 - \hat{q_{p_k}}) \tag{2}$$

An overview of the proposed end-to-end architecture shown in Fig. 2. The user types his/her query prefix for the given image to autocomplete and we perform image feature extraction using a pre-trained Convolutional Neural Network (CNN). Then, we feed the image features into the extended Long Short-Term Memory (LSTM) language model together with the query prefix which has a context-dependent weight matrix with an adaptation matrix constructed from a context-driven embedding model. These two constructs from image and text as visual features and textual queries are applied to complete a query. The completed query is then passed to iBERT (fine-tuned BERT language embedding model) to compute the patch probabilities, which in are utilised for patch selection. More details are provided in the next section.

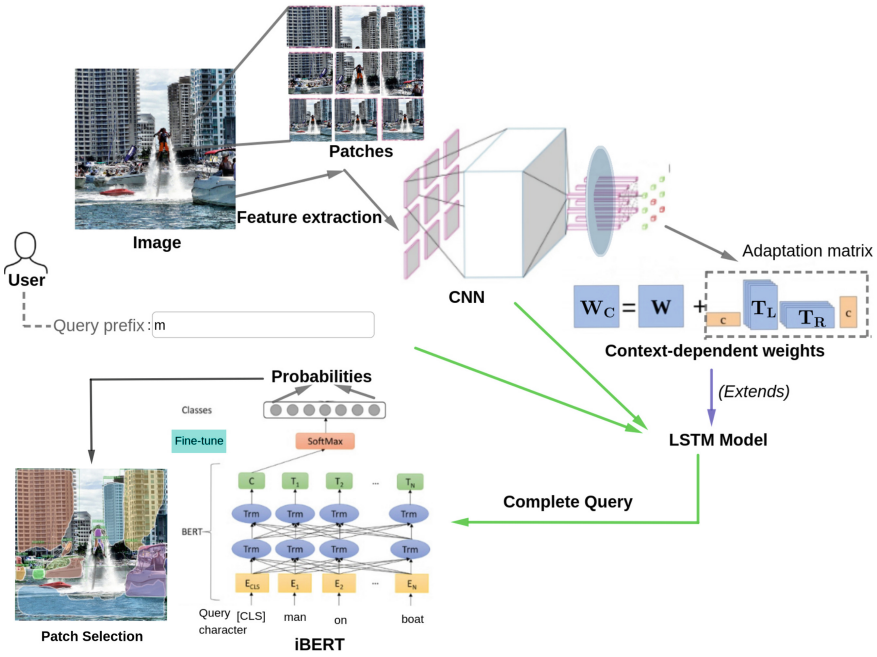

Fig. 2. The end-to-end architecture of Image Query Auto-Completion: User query prefix with the image features generated from a pre-trained CNN are input to an extended LSTM model (by incorporating a context-dependent weight matrix) which predicts a complete query. The resulting query is fed into a fine-tuned BERT pre-trained embedding model which outputs patch probabilities for patch selection.

3.1 Image Query Auto Completion

The challenge of query auto-completion is to predict and generate queries from prefixes that have never been seen in the training set. An initial attempt using neural language models has been introduced in [33]. The benefit of using character-level neural language models is providing more fine-grained predictions but they suffer from the semantic understanding that word-level models provide. For a prefix that has not been seen before (such as an incomplete word), their model enriches the shared information among comparable prefixes to create prediction nonetheless. In our scenario, we are given a prefix to complete a query conditioned on an image. To solve this new QAC problem, we exploit and extend the Long Short-Term Memory (LSTM) language model [12] with combined input and forget gates to auto complete queries. The language model is made up of a single-layer character-level LSTM with layer normalisation [2]. Our extension and modification to this language model is that we replace user embeddings with a low-dimensional representation of images. We adapt this LSTM language model alongside a context-dependent weight matrix $\mathbf{W}$ replaced by $\mathbf{W_C} = \mathbf{W} + \mathbf{M_A}$. We are providing a character embedding $w_c \in \mathbb{R}^e$, a preceding

hidden state $h_{c-1} \in \mathbb{R}^h$, where $\mathbf{M_A}$ is the adaptation matrix constructed by the product ($\times_i$ denotes the i-th-order tensor product) of the context c with two basis tensors, $\mathbf{T_L} \in \mathbb{R}^{u \times (e+h) \times v}$ and $\mathbf{T_R} \in \mathbb{R}^{v \times h \times u}$. Alternatively, the two basis tensors i.e., $\mathbf{T_L}$ and $\mathbf{T_R}$ are re-shaped to $\mathbb{R}^{u \times (v(e+h))}$ and $\mathbb{R}^{vh \times u}$. So the next predicted hidden state and the adaptation matrix can be equated as follows:

$$\begin{aligned} h_c &= \sigma([w_c, h_{c-1}]\mathbf{W_C} + b) \\ \mathbf{M_A} &= (c \times_1 \mathbf{T_L})(\mathbf{T_R} \times_3 c) \end{aligned} \tag{3}$$

We combine the context-driven weight matrix and the immediate preceding hidden state followed by the generated adaptation matrix which able to alter each query completion to be personalised to a particular image representation. We perform feature extraction on an input image using a Convolutional Neural Network (CNN) trained on ImageNet (pre-trained CNN), where we retrain only the last two fully connected layers shown in Fig. 2. The generated image feature vector is then fed into the LSTM language model via the adaptation matrix. We apply beam search decoding [34] in the generated array of predicted characters to select the optimal completion for the user query prefix.

3.2 iBERT - BERT for Patch Probability

We describe our approach to compute the probability of image patches which addresses an important aspect of query auto-completion systems. We assume that during the search process, users are typically interested in some part of the image as well as the image itself if it matches the mental picture of their belief [36]. Our work focuses on a new perspective of query auto-completion on images and the proposed model finds image patches which match the user context based on the query prefix using Eq. (1). BERT (Bidirectional Encoder Representations from Transformers) [7] shows promising results in multiple tasks of natural language processing and information retrieval [25] and is presently the state-of-the-art embedding model. We propose to fine-tune the BERT model as a transfer learning task for patch selection, using images composed of several patches (regions of an image), hence the name iBERT[3]. To the best of our knowledge, BERT has not yet been retraced for the QAC task. We use the BERT embedding model, which has a twelve layer implementation, extending it by adding a dense layer with 10% dropout which then is mapped to the final pooled layer connected the object class, and which outputs patch probabilities as shown in Fig. 2.

3.3 Information Foraging Explanation

Our goal of using Information Foraging Theory [27] from a cognitive viewpoint is to find explanations for the observed behaviour in query auto-completion and to model the information need within query sessions. IFT postulates that

[3] The lowercase "i" represents image patch.

the human information seekers follow an information scent to navigate from one information region to another in an information environment that is instinctively patchy in nature, and from one information patch to another within a region. IFT implies that foragers adapt their behaviour to the structure of the information environment in which they prevail such that the entire system (encompassing the information seeker, the information environment, and the interactions among these two) tries to maximise the ratio of the expected value of the information gained to the total cost of the interaction. Following the IFT analogy, when users start typing a prefix to auto-complete, their perceptual cues (such as mental beliefs [36]) either allow them to type the next character or to access the provided suggestion (under the query field) which acts as a distal cue and visually inspires the user to acquire them instantly to forage or seek. Query auto-completion, from an IFT perspective as query-level user interaction, is initiated by the user typing as little as a single-character query prefix. The user then may follow suggestions in case a completion is generated (which again follows the earlier mentioned strategy). In case the query prefix is unknown to the system (e.g. by being entered for the first time) the information scent associated with a result might be too poor [6] to immediately infer information needs. In this case we are applying beam search to generate the query based on image features. Suggestions are based on information scent values as described in the following subsection. These query suggestions represent the diversity of information scent patterns which elicits a varied distribution of relevant queries in the search field.

Patch Selection. This section describes the foraging-based strategy for patch selection. The technicalities of ranking patches (after patch selection) in the image (from image search results) are illustrated in Sect. 3.2. We utilise IFT to infer the user's information need utilising the Inferring User Need by Information Scent (IUNIS) algorithm [6] which was proposed to weigh each page vector along with the two factors i.e., TF-IDF weight and time, that were used to quantify the associated information scent with the page. In our image search scenario, we have images as search results where an image is considered as a set of patches containing features such as color, shape, texture, etc. In our proposed iBERT model, we use information scent to inspect patches based on image features and select patches which have higher probability estimated by the iBERT model. *Probabilistic Patch Selection Model* (PPSM) is a first attempt to reflect users' information need coherently by means of information scent. PPSM is used for a task that extends finding patches and makes the quantification of semantic uncertainties an important choice in selection. The important requirement for PPSM is a model (iBERT) that identifies patches in an image which are relevant to the user's information need (query). Inspired by the concept of TF-IDF in IR, we represent the categorical distribution of frequency (f_{p_i}) of each patch in an image (from the search results) in a given query session Q_s and the ratio of total number of query session (Q_T) during the entire search process to the number of query sessions (N_q) that contain the given patches (p_i) found in Q_s. We also consider the time spent (T) on the resulting images in a given query session to

estimate the information scent (IS) within a query session as:

$$IS(Q_s) = \sum_{i=1}^{n} f_{p_i} \log(\frac{Q_T}{N_q})T(p_i). \tag{4}$$

The user effort in terms of time is a function of patches which can be diverse and of different image class category. To generalise this for finding the information scent of a patch which then is assessed to select patches with higher information scent and then compared against the patch probability obtained via iBERT to distinguish the result. If we assume that the generated auto-completions induce several suggested queries (representing different information needs) simultaneously, every suggestion is in a competition to be discriminated as evident to the user. In the same way, an image contains multiple related or unrelated patches within it, and users find it difficult to judge which patches are relevant among images, which is due to the high uncertainty of correlated features within an image spread via patches. This motivates us to estimate the information scent of an image patch. There are two ways to compute the information scent of an image; one is to hire individual judges to rate scent on a scale [27] and the second approach is an algorithmic approach [28]. To estimate the information scent of a patch, we consider that PPSM constitutes patches that are probability distributions over images as *observations*. We assume image features as activators to perceptual cues because the user interpretation to image features when matched gives rise to a selection of an item (i.e., patch). The distributions are independent Bernoulli distributions of the features. Each observation is allocated to a patch, but the number of patches is not necessarily fixed i.e., the model is a non-parametric mixture with a product of independent Bernoullis as observation model. Therefore, the log-probability of selecting an image I for patch p_i

$$p(I \mid p_i) = \prod_{q_p} {r_{pf}}^{i_f} \, (1 - r_{pf})^{1-i_f} \tag{5}$$

where $r_p f = f((\pi_i, s_i), (1-\pi_i)s_i)$ is the Bernoulli rate for patch p to emit feature f, i_f is the image containing feature f, and $r_p f$ is a function of prior parameters representing activators (perceptual cues) for the selected patch. There can be a situation when most patches have only one observation (image) and features are very sparse i.e., the possibility of multiple perceptual cues per patch (i.e., $\pi_s \ll 1$) is low. To interpret Bernoulli's prior parameters such as s_i, we find the probability to observe a feature ($f \in i$ meaning $i = 1$) provided that it has been observed for a patch p ($k = 1$) is:

$$p(i = 1|k = 1, n = 1) = \frac{s_1\pi_s + n}{s_1 + n} = \frac{s_1\pi_s + 1}{s_1 + 1} \approx \frac{1}{s_1 + 1} \tag{6}$$

if $\pi_s \ll 1$. The probability of observing a feature in a new image, given that it has been observed before, is a measure of its reliability. We use this probabilistic model to compare the results based on the probabilities of patches obtained from iBERT.

4 Experiments

4.1 Dataset

We use two well-known and diverse datasets: a visual dataset with large-scale knowledge bases that provide a rich collection of language annotations for visual concepts known as *Visual Genome* [18] with over 100k images where most image categories fall within a long tail, and the *ReferIt* dataset [16] which contains $\sim 42k$ image regions with descriptions. These two datasets fit well for our tasks. The Visual Genome dataset includes images, region descriptions, question-answers, objects, relationships, and attributes. The region descriptions confer a substitution for queries as they refer to several objects in various regions of every image. Few region descriptions are referring phrases and few of them are quite alike to descriptions. For example, referring descriptions are "guy sitting on the couch", "white keyboard on the desk" and non-referring descriptions are "couch is brown" and "mouse is in the charger". The huge number of instances from the Visual Genome dataset makes it quite convenient for our task. The ReferIt dataset is a collection of referring expressions engaged to images which quite intently resemble probable user queries of images. We separately train models for query auto-completion and patch selection using both datasets.

4.2 Training

We combine query and image as pairs by utilising the region descriptions from the Visual Genome dataset and referring to expressions from the ReferIt dataset. During training, we taken 85% of the Visual Genome data as the training set consists of 16,000 images and 740,000 corresponding region descriptions in which there are approximately 40–45 text descriptions per image. The training data from the ReferIt dataset consists of 9,000 images and 54,000 referring expression with approximately 4–6 referring expression per image.

For the query auto-completion task, we train our extended LSTM language model where the dimension of image representation is 128, $r = 64$ is the rank of the matching personalised matrix (component from Fig. 2). We use character embeddings with dimension 24, the dimension of the LSTM hidden units is 512, and a maximum length of 50 characters per query with Adam optimizer at a learning rate of 5e-4 for 50,000 iterations as well as a batch size of 32. For the patch selection task, we train our proposed iBERT model using pairs of (region description, patch set) from the Visual Genome dataset, giving rise to a training set of approximately 1.73 million samples. The extra 0.3 million samples are split into test and validation set. We conduct training for the patch selection model that fine-tunes BERT having twelve layers with batch size of 32 for 250,000 iteration using Adam as optimizer at a learning rate of 5e-5 in which the performance increases steeply for the initial 10% of iterations. We use a NVIDIA Tesla T4 GPU which takes a day and half for the complete training activity.

4.3 Performance Measure

We evaluate the quality of our predictions and estimations using the following performance metrics:

Mean Reciprocal Rank: The most standard metrics for QAC tasks is the mean reciprocal rank (MRR), which is the average of the reciprocal ranks of the final queries in the QAC outcomes. The MRR for the query auto-completion system Q_A provided the test dataset D_T is as follows:

$$MRR(Q_A) = \frac{1}{\mid D_T \mid} \sum_{q \in D_T} RR(q, Q_A(q_p))$$

where q_p is a prefix of query q and $Q_A(q_p)$ is the list ranked for candidate completions of q_p from Q_A. RR denotes the reciprocal rank of q if q is present in $Q_A(q_p)$, in other cases reciprocal is 0.

Language Perplexity: Perplexity is a measure to encapsulate uncertainty of the model for a given query prefix. This metric has been explored earlier for an information retrieval task [8] and its correlation with the standard precision-recall measures has been investigated [1]. The average inverse probability is perplexity. A better model has lower perplexity.

$$Perplexity(q_p) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(q_i|q_{i-1})}}$$

where N is the normalised length of the query and $P(q_i|q_{i-1})$ is the probability of the complete query given the immediate preceding query prefix.

We evaluate the patch selection by F1 score.

4.4 Results and Discussion

We report the evaluation result in Table 1. We perform our evaluation in two parts. Firstly, we evaluate the quality of our query completion (query prefix of length one or more character) by mean reciprocal rank and perplexity. Secondly, we evaluate the patch selection task by F1 score. We evaluate the query completion task on Visual Genome and ReferIt datasets which have character vocabulary sizes of 89 and 77. We match index T_q of the true query prefix in the top 10 predicted completions where we estimate the MRR score as $\sum_n \frac{1}{T_q}$ and reinstate the reciprocal rank with 0 in case if query does not appear in the top 10 completions. The perplexity comparison on both collection of test queries utilising corresponding contexts i.e., images and indiscriminate noise. The perplexity on the Visual Genome and ReferIt test queries with both contexts is shown in Table 2. During the evaluation on the Visual Genome and ReferIt test sets (or

queries), we analyse the query prefix with different length for the corresponding context (noise and image). We found that mean reciprocal rank is altered by the query prefix length, as long-tailed queries are comparatively more difficult than queries of average length to match. Hence, we examine quite better performance for all prefix lengths on the ReferIt dataset (from Table 2).

Table 1. Evaluation results of the query completion task. Our MRR score is in bold face.

Model	MRR (Seen+Unseen)
MPC [3]	0.171
Character n-gram (n = 7)	0.287
Mitra10K+MPC+λMART [24]	0.278
Mitra100K+MPC+λMART [24]	0.298
NQLM(S)+WE+MPC [26]	0.345
NQLM(L)+WE+MPC [26]	0.355
NQLM(L)+WE+MPC+λMART [26]	0.354
FactorCell [12]	0.309
E-LSTM LM(Ours)[a]	**0.764**

[a]E-LSTM LM: Extended LSTM Language Model

Table 2. Perplexity of image query auto-completion on both datasets utilising an image and indiscriminate noise. Inclusion of image results in a better (lower) perplexity

Dataset	Context	
	Image	Indiscriminate noise
Visual Genome	2.35	3.81
ReferIt	2.63	3.45

We evaluated our proposed iBERT model for finding patch probabilities which is used to select and rank patches in the image. We achieve an F1 score[4] of **0.7638** over 3,000 patch classes.

5 Conclusion and Future Work

In this work, we propose an extended LSTM language model for a new task of query auto-completion adapted upon an image. The language model enriches both image features and text information in which the surplus of beam search over our model is efficiently able to predict future queries at least on a single character prefix. The significant increase in MRR is due to the inclusion of

[4] F1 score for the baseline methods shown in Table 1 were not available.

visual information within textual queries as explained by IFT model. Also, we present iBERT for patch selection to efficiently rank them in the image and eventually predicts the most suitable image for the auto-completed query, and compare against the result from probabilistic patch selection model. This work is among the first attempt to apply foraging-based strategy to QAC. The self-explanatory power of IFT to understand user interaction at query level leads the foundation of probabilistic patch selection model to devise users' information need. Our future work is to generalise the referring expression with contextual model to distinguish referring and non-referring region descriptions. We intend to aggregate information from textual queries and visual descriptions to scale it for multimodal query auto-completion in a single model.

Acknowledgement. This work is supported by the Quantum Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 721321, and partially supported for computing resources by Google Cloud grant.

References

1. Azzopardi, L., Girolami, M., Van Rijsbergen, K.: Investigating the relationship between language model perplexity and IR precision-recall measures (2003)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bar-Yossef, Z., Kraus, N.: Context-sensitive query auto-completion. In: Proceedings of the 20th International Conference on World Wide Web, pp. 107–116. ACM (2011)
4. Cai, F., De Rijke, M., et al.: A survey of query auto completion in information retrieval. Found. Trends® Inf. Retrieval **10**(4), 273–363 (2016)
5. Cao, H., et al.: Context-aware query suggestion by mining click-through and session data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 875–883. ACM (2008)
6. Chi, E.H., Pirolli, P., Chen, K., Pitkow, J.: Using information scent to model user information needs and actions and the web. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 490–497. ACM (2001)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Hauff, C., Murdock, V., Baeza-Yates, R.: Improved query difficulty prediction for the web. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 439–448. ACM (2008)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
10. Hu, R., Dollár, P., He, K., Darrell, T., Girshick, R.: Learning to segment every thing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4233–4241 (2018)
11. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4555–4564 (2016)

12. Jaech, A., Ostendorf, M.: Personalized language model for query auto-completion. arXiv preprint arXiv:1804.09661 (2018)
13. Jaiswal, A.K., Holdack, G., Frommholz, I., Liu, H.: Quantum-like generalization of complex word embedding: a lightweight approach for textual classification. In: Proceedings of the Conference "Lernen, Wissen, Daten, Analysen", LWDA 2018, Mannheim, Germany, 22–24 August 2018, pp. 159–168 (2018). http://ceur-ws.org/Vol-2191/paper19.pdf
14. Ji, S., Li, G., Li, C., Feng, J.: Efficient interactive fuzzy keyword search. In: Proceedings of the 18th International Conference on World Wide Web, pp. 371–380. ACM (2009)
15. Jiang, J.Y., Ke, Y.Y., Chien, P.Y., Cheng, P.J.: Learning user reformulation behavior for query auto-completion. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 445–454. ACM (2014)
16. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: referring to objects in photographs of natural scenes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 787–798 (2014)
17. Kharitonov, E., Macdonald, C., Serdyukov, P., Ounis, I.: User model-based metrics for offline query suggestion evaluation. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 633–642. ACM (2013)
18. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
19. Liu, H., Mulholland, P., Song, D., Uren, V., Rüger, S.: Applying information foraging theory to understand user interaction with content-based image retrieval. In: Proceedings of the Third Symposium on Information Interaction in Context, pp. 135–144. ACM (2010)
20. Maxwell, D., Azzopardi, L.: Information scent, searching and stopping. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 210–222. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_16
21. McCann, B., Bradbury, J., Xiong, C., Socher, R.: Learned in translation: contextualized word vectors. In: Advances in Neural Information Processing Systems, pp. 6294–6305 (2017)
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
23. Mitra, B.: Exploring session context using distributed representations of queries and reformulations. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3–12. ACM (2015)
24. Mitra, B., Craswell, N.: Query auto-completion for rare prefixes. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1755–1758. ACM (2015)
25. Mitra, B., Rosset, C., Hawking, D., Craswell, N., Diaz, F., Yilmaz, E.: Incorporating query term independence assumption for efficient retrieval and ranking using deep neural networks. arXiv preprint arXiv:1907.03693 (2019)
26. Park, D.H., Chiba, R.: A neural language model for query auto-completion. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1189–1192. ACM (2017)

27. Pirolli, P., Card, S.: Information foraging. Psychol. Rev. **106**(4), 643 (1999)
28. Pirolli, P., Card, S.K., Van Der Wege, M.M.: Visual information foraging in a focus+ context visualization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 506–513. ACM (2001)
29. Shao, T., Chen, H., Chen, W.: Query auto-completion based on word2vec semantic similarity. In: Journal of Physics: Conference Series, vol. 1004, p. 012018. IOP Publishing (2018)
30. Shi, H., Li, H., Meng, F., Wu, Q.: Key-Word-aware network for referring expression image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 38–54. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_3
31. Shokouhi, M.: Learning to personalize query auto-completion. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 103–112. ACM (2013)
32. Sundar, S.S., Knobloch-Westerwick, S., Hastall, M.R.: News cues: information scent and cognitive heuristics. J. Am. Soc. Inform. Sci. Technol. **58**(3), 366–378 (2007)
33. Sutskever, I., Martens, J., Hinton, G.E.: Generating text with recurrent neural networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 1017–1024 (2011)
34. Vijayakumar, A.K., et al.: Diverse beam search: decoding diverse solutions from neural sequence models. arXiv preprint arXiv:1610.02424 (2016)
35. Weber, I., Castillo, C.: The demographics of web search. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 523–530. ACM (2010)
36. White, R.: Beliefs and biases in web search. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3–12. ACM (2013)
37. White, R.W., Marchionini, G.: Examining the effectiveness of real-time query expansion. Inf. Process. Manag. **43**(3), 685–704 (2007)
38. Wittek, P., Liu, Y.H., Darányi, S., Gedeon, T., Lim, I.S.: Risk and ambiguity in information seeking: eye gaze patterns reveal contextual behavior in dealing with uncertainty. Front. Psychol. **7**, 1790 (2016)
39. Wu, C.C., Mei, T., Hsu, W.H., Rui, Y.: Learning to personalize trending image search suggestion. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 727–736. ACM (2014)