(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 March 2021 (18.03.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/050776 A1

(51) International Patent Classification:
G06K 9/20 (2006.01) *G06F 16/33* (2019.01)
G06K 9/48 (2006.01) *G06F 40/284* (2020.01)
G06F 16/53 (2019.01) *G06N 3/02* (2006.01)

(21) International Application Number:
PCT/US2020/050258

(22) International Filing Date:
10 September 2020 (10.09.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/899,307 12 September 2019 (12.09.2019) US
17/014,984 08 September 2020 (08.09.2020) US

(71) Applicant: **NEC LABORATORIES AMERICA, INC.** [US/US]; Suite 200, 4 Independence Way, Princeton, New Jersey 08540 (US).

(72) Inventors: **LAI, Farley**; 1009 Deer Creek Drive, Plainsboro, New Jersey 08536 (US). **KADAV, Asim**; 108 Moffett Boulevard, Mountain View, California 94043 (US). **XIE, Ning**; 1432 Forest Lane, Fairborn, Ohio 45324 (US).

(74) Agent: **KOLODKA, Joseph**; NEC Laboratories America, Inc., Suite 200, 4 Independence Way, Princeton, New Jersey 08540 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: CONTEXTUAL GROUNDING OF NATURAL LANGUAGE PHRASES IN IMAGES

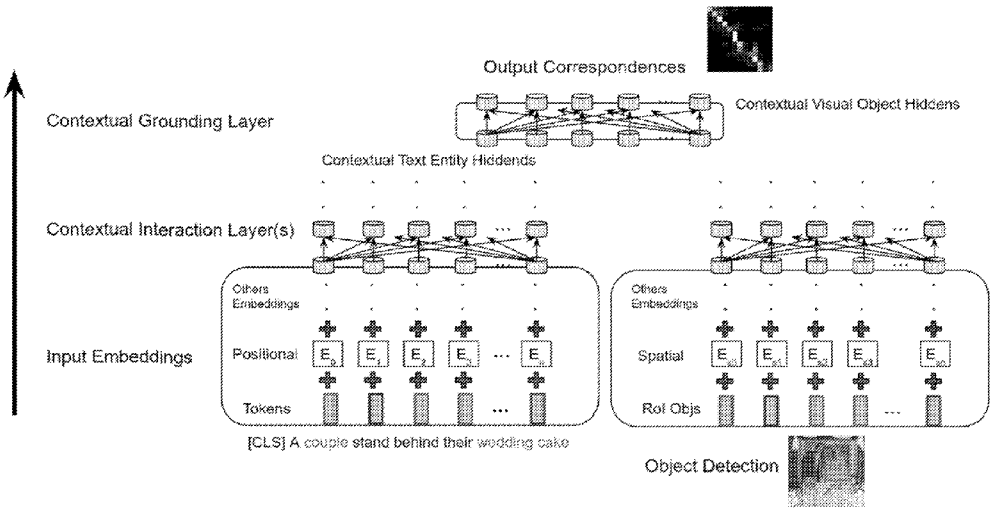

FIG. 3

(57) Abstract: Aspects of the present disclosure describe systems, methods and structures providing contextual grounding - a higher-order interaction technique to capture corresponding context between text entities and visual objects.

WO 2021/050776 A1

[Continued on next page]

Published:
— *with international search report (Art. 21(3))*

CONTEXTUAL GROUNDING OF NATURAL LANGUAGE PHRASES IN IMAGES

TECHNICAL FIELD

[0001] This disclosure relates generally to language text and images . More particularly, it describes techniques for corresponding language text with visual objects included in images.

BACKGROUND

[0002] Language grounding is a fundamental task to address visual reasoning challenges that require understanding the correspondence between text entities and objects in images. One straightforward, real-world application of language grounding - is a natural language retrieval system that takes as input a textual query and returns as output a visual object in a given image referred to by the language entity in the query. Notwithstanding such great need and utility, automated systems, methods, and structures that perform language grounding present significant technical challenges not yet met in the art.

SUMMARY

[0003] An advance in the art is made according to aspects of the present disclosure directed to systems, methods, and structures that provide contextual grounding of natural language entities in images.

[0004] In sharp contrast to the prior art, systems, methods, and structures according to aspects of the present disclosure introduce a novel architecture that advantageously captures a context of corresponding text entities and image regions thereby improving grounding accuracy.

[0005] In further contrast to the prior art, systems, methods, and structures according to aspects of the present disclosure introduces a contextual grounding approach

that captures the context in corresponding text and images respectively without any specific embedding or object feature extraction.

[0006] Operationally, our architecture disclosed herein accepts pre-trained text token embeddings and image object features from an object detector as input. method d. Additional encoding that captures positional and spatial information can be added to enhance the feature quality. Separate text and image branches facilitate respective architectural refinements for different modalities. The text branch is pre-trained on a large-scale masked language modeling task while the image branch is trained from scratch.

[0007] Our model learns contextual representations of the text tokens and image objects through layers of high-order interaction, respectively. A final grounding head ranks a correspondence between the textual and visual representations through cross-modal interaction.

[0008] Finally, in our evaluation, we show that our model achieves the state-of-the-art grounding accuracy of 71.36% over the Flickr30K Entities dataset. No additional pre-training is necessary to deliver competitive results compared with related work that often requires task-agnostic and task-specific pre-training on cross-modal datasets.

BRIEF DESCRIPTION OF THE DRAWING

[0009] A more complete understanding of the present disclosure may be realized by reference to the accompanying drawing in which:

[0010] **FIG. 1** is a schematic diagram illustrating an example image from Flickr30K entities annotated with bounding boxes corresponding to entities in the caption;

[0011] **FIG. 2** is a schematic diagram illustrating a natural language object retrieval system diagram according to aspects of the present disclosure; and

[0012] **FIG. 3** is a schematic diagram illustrating contextual grounding architecture and workflow according to aspects of the present disclosure.

[0013] The illustrative embodiments are described more fully by the Figures and detailed description. Embodiments according to this disclosure may, however, be embodied in various forms and are not limited to specific or illustrative embodiments described in the drawing and detailed description.

DESCRIPTION

[0014] The following merely illustrates the principles of the disclosure. It will thus be appreciated that those skilled in the art will be able to devise various arrangements which, although not explicitly described or shown herein, embody the principles of the disclosure and are included within its spirit and scope.

[0015] Furthermore, all examples and conditional language recited herein are intended to be only for pedagogical purposes to aid the reader in understanding the principles of the disclosure and the concepts contributed by the inventor(s) to furthering the art and are to be construed as being without limitation to such specifically recited examples and conditions.

[0016] Moreover, all statements herein reciting principles, aspects, and embodiments of the disclosure, as well as specific examples thereof, are intended to encompass both structural and functional equivalents thereof. Additionally, it is intended that such equivalents include both currently known equivalents as well as equivalents developed in the future, i.e., any elements developed that perform the same function, regardless of structure.

[0017] Thus, for example, it will be appreciated by those skilled in the art that any block diagrams herein represent conceptual views of illustrative circuitry embodying the principles of the disclosure.

[0018] Unless otherwise explicitly specified herein, the FIGs comprising the drawing are not drawn to scale.

[0019] By way of some additional background, we note that cross-model reasoning is challenging for grounding entities and objects in different modalities such as text and images. Representative tasks include visual question answering (VQA) and image captioning that leverages grounded features between text and images to make predictions.

[0020] While recent advances in these tasks achieve impressive results, the quality of the correspondence between textual entities and visual objects in both modalities is neither convincing nor interpretable. This is likely because the grounding from one modality to the other is trained implicitly and any intermediate results are not often evaluated as explicitly as in object detection.

[0021] To address this issue, the *Flickr30K Entities* dataset with precise annotations of the correspondence between language phrases and image regions to ease the evaluation of visual grounding was created.

[0022] **FIG. 1** is a schematic diagram illustrating an example image from Flickr30K entities annotated with bounding boxes corresponding to entities in the caption. In that figure, two men are referred to as separate entities. To uniquely ground the two men in the image, a grounding algorithm must take respective context and attributes into consideration for learning the correspondence.

[0023] Historically over the years, many deep learning based approaches were proposed to tackle this localization challenge. The basic idea behind such approaches is to derive representative features for each entity as well as object, and then score their correspondence. In the modality of caption input, individual token representations usually start with the word embeddings followed by a recurrent neural network (RNN), usually Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU), to capture the

contextual meaning of the text entity in a sentence. On the other hand, the visual objects in image regions of interest (RoI) are extracted through object detection.

[0024] Each detected object typically captures limited context through the receptive fields of 2D convolutions. Advanced techniques such as the *feature pyramid network* (FPN) enhance the representations by combining features at different semantic levels w.r.t. the object size. Even so, those conventional approaches are limited to extracting relevant long range context in both text and images effectively. In view of this limitation, non-local attention techniques have been proposed to address the long range dependencies in natural language processing (NLP) and computer vision (CV) tasks.

[0025] Inspired by this advancement, we introduce the contextual grounding approach to improving the representations through extensive intra- and inter-modal interaction to infer the contextual correspondence between text entities and visual objects.

[0026] **Related Work.** On the methodology of feature interaction, the *Transformer* architecture for machine translation demonstrates a systematic approach to efficiently computing the interaction between language elements. Around the same time, *non-local networks* generalize the transformer to the CV domain, supporting feature interaction at different levels of granularity from feature maps to pooled objects.

[0027] Recently, the image transformer adapts the original transformer architecture to the image generation domain by encoding spatial information in pixel positions while we deal with image input at the RoI level for grounding. Additionally, others have proposed BERT (Bidirectional Encoder Representations from Transformers) as a pre-trained transformer encoder on large-scale masked language modeling, facilitating training downstream tasks to achieve state-of-the-art (SOTA) results.

[0028] As we shall show and describe, our work extends BERT to the cross-modal grounding task by jointly learning contextual representations of language entities and visual objects. Coincidentally, another line of work named *VisualBERT* also

integrates BERT to deal with grounding in a single transformer architecture. However, their model requires both task-agnostic and task-specific pre-training on cross-modal datasets to achieve competitive results. Ours, on the contrary, achieves SOTA results without additional pre-training and allows respective architectural concerns for different modalities.

[0029] Contextual Grounding

[0030] The main approach of the prior art uses RNN/LSTM to extract high level phrase representations and then apply different attention mechanisms to rank the correspondence to visual regions. While the hidden representations of the entity phrases take the language context into consideration, the image context around visual objects is in contrast limited to object detection through 2D receptive fields. Nonetheless, there is no positional ordering as in text for objects in an image to go through the RNN to capture potentially far apart contextual dependencies.

[0031] In view of the recent advances in NLP, the transformer architecture addresses the long range dependency through pure attention techniques. Without RNN being incorporated, the transformer enables text tokens to efficiently interact with each other pairwise regardless of the range. The ordering information is injected through additional positional encoding. Enlightened by this breakthrough, the corresponding contextual representations of image RoIs may be derived through intra-modal interaction with encoded spatial information.

[0032] **FIG. 2** is a schematic diagram illustrating a natural language object retrieval system diagram according to aspects of the present disclosure. With reference to that figure, it may be observed that the contextual ground module is depicted as a functional block.

[0033] Access to such a system is achieved through – for example – a computer browser that shows an input field for a user to type a query w.r.t. an image and renders

retrieval results in an image. Accordingly, input to the system is a pair of textual query(ies) and image(s).

[0034] The query is parsed into tokens and applied (fed) into an object detector to locate salient regions as visual object candidates for subsequent grounding. The contextual grounding module accepts both entity embeddings and visual object representations as input and scores their correspondences in probabilities. Finally, the object corresponding to the query language entity with the highest probability score is retrieved and visualized in a bounding box to the user.

[0035] **FIG. 3** is a schematic diagram illustrating contextual grounding architecture and workflow according to aspects of the present disclosure.

[0036] According to aspects of the present disclosure, the grounding objective guides the attention to the corresponding context in both the text and image with improved accuracy. Consequently, we describe contextual grounding architecture as shown in **FIG 3**.

[0037] As we shall describe in greater detail, inside the contextual grounding module shown above, each input entity embedding vectors and visual objects go through multiple contextual interaction layers to attend to each other in the same modality such that the resulting representations involve features from the context. To further improve the performance, additional encoding features may be added such as positional encoding to add order information to textual entities in the query, and the spatial encoding to add the locations information of visual objects in the image. Lastly, the contextual grounding layer ranks the contextual entity as well as visual object representations pairwise, and output the resulting scores.

[0038] As shown in that figure, the model is composed of two transformer encoder branches for both text and image inputs to generate their respective contextual representations for the grounding head to decide the correspondence. The text branch is

pre-trained from the BERT base model which trains a different positional embedding from the original transformer. On the other hand, the image branch takes RoI features as input objects from an object detector.

[0039] Correspondingly, we train a two layer multi-layer perceptron (MLP) to generate the spatial embedding given the absolute spatial information of the RoI location and size normalized to the entire image. Both branches add the positional and spatial embedding to the tokens and RoIs respectively as input to the first interaction layer. At each layer, each hidden representation performs self-attention to each other to generate a new hidden representation as layer output. The self-attention may be multi-headed to enhance the representativeness. At the end of each branch, the final hidden state is fed into the grounding head to perform the cross-modal attention with text entity hidden states as queries and image object hidden representations as the keys. The attention responses serve as the matching correspondences. If the correspondence does not match the ground truth, the mean binary cross entropy loss per entity is back propagated to guide the interaction across the branches. We evaluate the grounding recall on the Flickr30K Entities dataset and compare the results with SOTA work in the next section.

[0040] **Evaluation**

[0041] Our contextual grounding approach uses the transformer encoder to capture the context in both text entities and image objects. While the text branch is pre-trained from BERT, the image branch is trained from scratch. In view of the complexity of the transformer, previous work has shown that performance varies with different numbers of interaction layers and attention heads. Also, the intra-modal object interaction does not necessarily consider the relationship in space unless some positional or spatial encoding is applied. In our evaluation, we vary both the number of layers and heads, along with adding the spatial encoding to explore the performance variations summarized in **Table 1**.

[0042] We achieve the SOTA results in all top 1, 5 and 10 recalls based on the same object detector as used by previous SOTA BAN. The breakdown of per entity type

recalls is given in **Table 2**. As may be observed therein, six out of the eight entity type recalls benefit from our contextual grounding. Interestingly, the recall of the instrument type suffers. This may be due to the relatively small number of instrument instances in the dataset preventing the model from learning the context well.

[0043] On the other hand, compared with the text branch consisting of 12 layers and 12 heads with hidden size of 768 dimensions, the best performance is achieved with the image branch having 1 layer, 2 attention heads and hidden size of 2048 dimensions. Moreover, adding the spatial embedding consistently improves the accuracy by 0.5% or so. This is likely because image objects, unlike word embedding requiring the context to produce representative hidden states for its meaning, may already capture some neighborhood information through receptive fields.

[0044] Finally, in **Table 3**, we compare the results with the recent work in progress, VisualBERT, which also achieves improved grounding results based on a single transformer architecture that learns the representations by fusing text and image inputs in the beginning. Marginally, ours performs better in the top 1 recall.

[0045] Note that advantageously our approach according to aspects of the present disclosure - unlike VisualBERT which requires task-agnostic and task-specific pre-training on COCO captioning and the target dataset - needs no similar pre-training to deliver competitive results. In addition, our architecture is also flexible to adapt to different input modalities, respectively.

Model	Detector	R@1	R@5	R@10	Upper Bound
# 1	Fast RCNN	50.89	71.09	75.73	85.12
# 2	YOLOv2	53.97	-	-	-
# 3	Query-Adaptive RCNN	65.21	-	-	-
# 4	Bottom-Up [1]	69.69	84.22	86.35	87.45
Ours L1-H2-abs	Bottom-Up [1]	71.36	84.76	86.49	87.45

Ours L1-H1-abs	Bottom-Up [1]	71.21	84.84	86.51	87.45
Ours L1-H1	Bottom-Up [1]	70.75	84.75	86.39	87.45
Ours L3-H2-abs	Bottom-Up [1]	70.82	84.59	86.49	87.45
Ours L3-H2	Bottom-Up [1]	70.39	84.68	86.35	87.45
Ours L6-H4-abs	Bottom-Up [1]	69.71	84.10	86.33	87.45

Table 1.

Model	People	Clothing	Body Parts	Animals	Vehicles	Instruments	Scene	Other
# 1	64.73	46.88	17.21	65.83	68.75	37.65	51.39	31.77
# 2	68.71	46.83	19.50	70.07	73.75	39.50	60.38	32.45
# 3	78.17	61.99	35.25	74.41	76.16	56.69	68.07	47.42
# 4	79.90	74.95	47.23	81.85	76.92	43.00	68.69	51.33
Ours L1-H2-abs	81.95	76.5	46.27	82.05	79.0	35.8	70.23	53.53
# of instances	5656	2306	523	518	400	162	1619	3374

Table 2.

Model	R@1 Dev Test	R@5 Dev Test	R@10 Dev Test	Upper Bound Dev Test
VisualBERT w/o COCO Pre-training	68.07 -	83.98 -	86.24 -	86.9787.45
VisualBERT	70.40 71.33	84.49 84.98	86.31 86.51	
Ours L1-H2-abs	69.8 71.36	84.22 84.76	86.21 86.49	86.9787.45

Table 3.

[0046] Note that advantageously our approach according to aspects of the present disclosure - unlike VisualBERT which requires task-agnostic and task-specific pre-training on COCO captioning and the target dataset - needs no similar pre-training to deliver competitive results. In addition, our architecture is also flexible to adapt to different input modalities, respectively.

[0047] To summarize, those skilled in the art will appreciate that systems, methods, and structures according to aspects of the present disclosure advantageously improves the performance of grounding module(s) by matching relevant textual entities with corresponding visual objects. As will be further understood and appreciated – with respect to the present disclosure - there are two branches accepting the textual entity embeddings and visual object representations respectively which are later ranked by the correspondences following the steps below.

[0048] First, the two branches assume the input of the textual query and image is preprocessed and converted to some embeddings and object representations. In particular, the input query is tokenized in words or smaller tokens to extract language entity embeddings as the text branch input. Advantageously, additional information such as the positional encoding may be used to enrich the order information of the sequence of tokens. The encoding can be derived and trained from absolute 1D or relative positions of each other and the encoding may be applied to the input elements and/or attention across subsequent contextual interaction layers. The input visual objects are extracted by some object detector that provides object features as the image branch input wherein additional information such as spatial encoding may be used to distinguish the spatial relationships between different visual objects; the encoding can be derived and trained from absolute 2D relative locations to each other; and the encoding may be applied to the intpu elements and/or attention across the subsequent contextual interaction layers;

[0049] Second, each branch is then followed by one or more contextual interaction layers where the input elements from the same modality attend to each other to capture relevant context as the layer output representations.

[0050] Third, all pairs of the last layer language entity embeddings and visual object representations are scored to rank their correspondences as the grounding output in probabilities.

[0051] At this point, while we have presented this disclosure using some specific examples, those skilled in the art will recognize that our teachings are not so limited. Accordingly, this disclosure should only be limited by the scope of the claims attached hereto.

Claims:

1. A method for text-image retrieval including text and image branches, said method comprising:

receiving as input a text query and an image;

parsing the input text query into tokens and converting them to entity embedding vectors;

locating visual object candidates in the input image.

scoring correspondences between the entity embeddings and visual object candidates;

providing, visualized in a bounding box, the object corresponding to the query text entity with the highest probability score, to a user of the system.

wherein no specific embedding or object feature extraction is used in the method.

2. The method of system of claim 1 further comprising:

pre-training the text branch utilizing a BERT, Bidirectional Encoder Representations from Transformers, base model.

3. The method of claim 2 further comprising:

receiving, by the image branch, region of interest (RoI) features as input objects from an object detector.

4. The method of claim 3 further comprising:

training, a two-layer multi-layer perceptron (MLP) to generate spatial embedding given absolute spatial information of the RoI location and size normalized to the entire image.

5. The method of claim 4 further comprising:

adding, by both branches, positional and spatial embedding to tokens and RoIs respectively as input to a first interaction layer of the MLP.

6. The method of claim 5 further comprising:

performing, at each layer of the MLP, self-attenuation by each hidden representation to each other to generate a new hidden representation as layer output.

7. The method of claim 6 further comprising:

providing, at the end of each branch, a final hidden state to a ground head to provide cross-modal attention responses with text entity hidden states as queries and image object hidden representations as keys.

8. The method of claim 7 wherein matching correspondences are determined from the attention responses.

9. The method of claim 8 further comprising

back propagating a mean binary cross entropy loss per entity if the correspondence(s0 does not match a ground truth.

FIG. 1

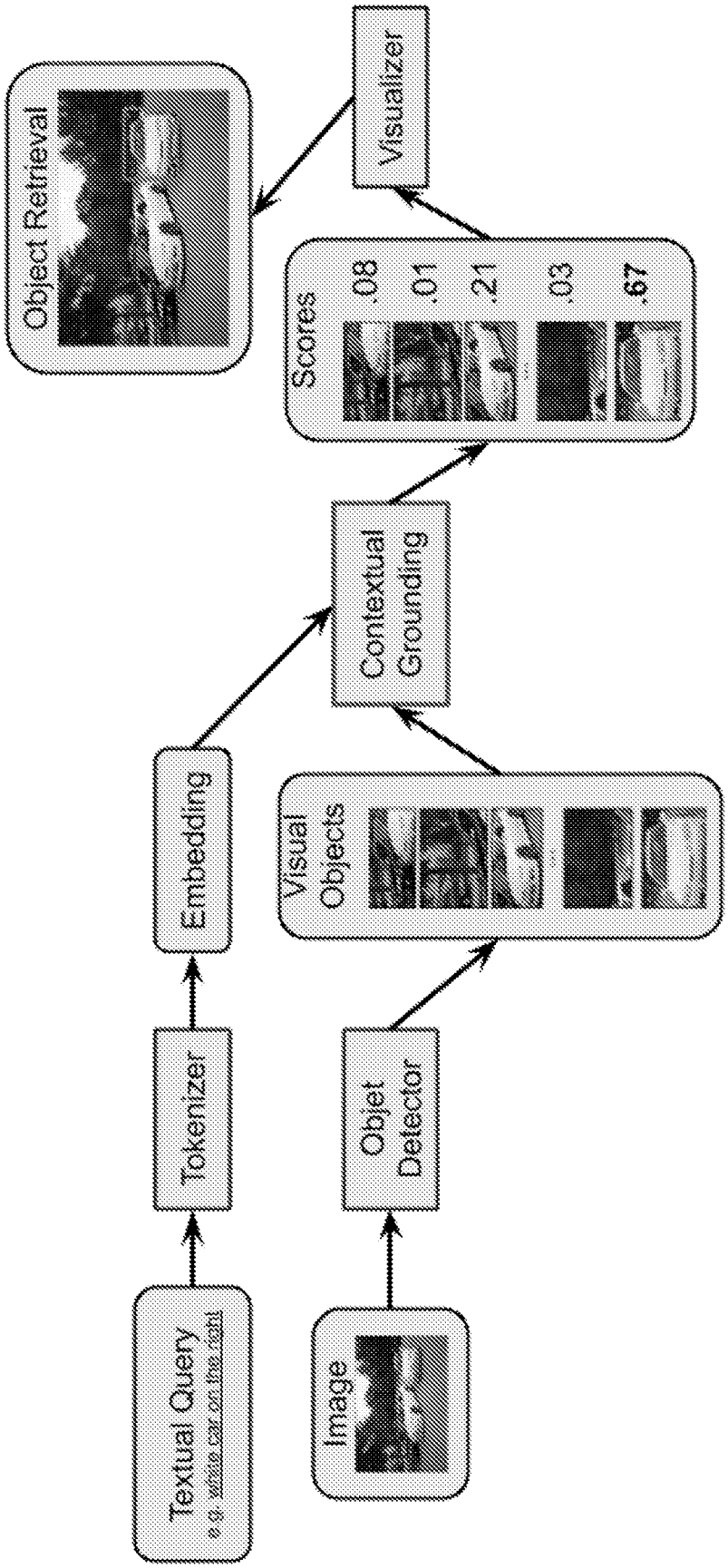
Object Retrieval
Visualizer
Scores
.08
.01
.21
.03
.67
Contextual Grounding
Embedding
Visual Objects
Tokenizer
Objet Detector
Textual Query
e.g. what's the car on the right
Image

FIG. 2

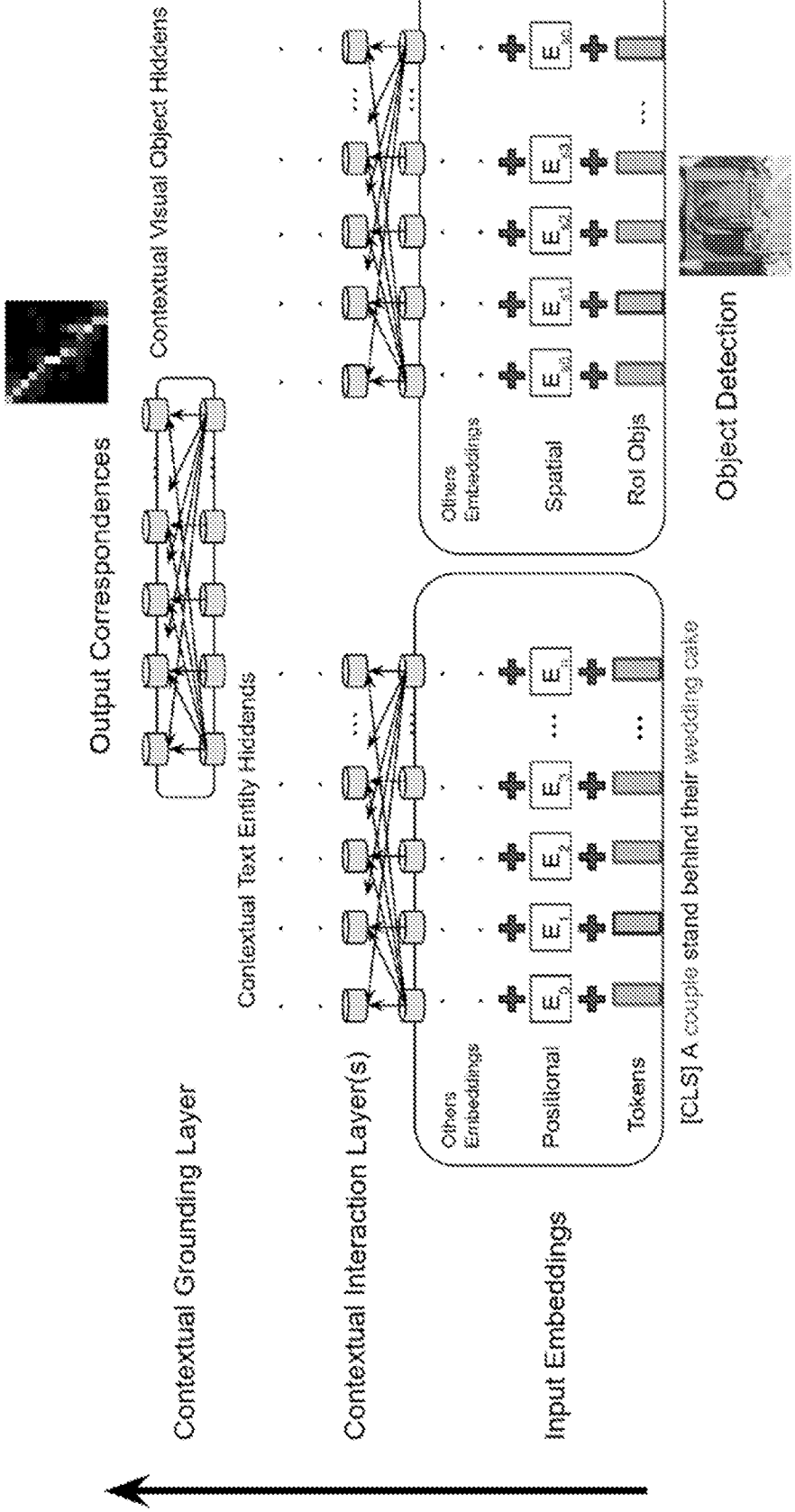
Output Correspondences
Contextual Visual Object Hiddens
Contextual Text Entity Hiddends
Contextual Grounding Layer
Contextual Interaction Layer(s)
Others Embeddings
Positional
Tokens
Input Embeddings
[CLS] A couple stand behind their wedding cake
Others Embeddings
Spatial
Roi Objs
Object Detection

FIG. 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2020/050258

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/20(2006.01)i; **G06K 9/48**(2006.01)i; **G06F 16/53**(2019.01)i; **G06F 16/33**(2019.01)i; **G06F 40/284**(2020.01)i; **G06N 3/02**(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K 9/20(2006.01); G06F 17/27(2006.01); G06F 17/28(2006.01); G06F 17/30(2006.01); G06K 9/00(2006.01); G06K 9/46(2006.01); G06N 7/00(2006.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: text, image, retrieval, candidate, score

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2019-0130206 A1 (SALESFORCE.COM, INC.) 02 May 2019 (2019-05-02) paragraphs [0031]-[0055], claims 8-15 and figures 1-11	1-3
A		4-9
Y	US 2019-0266236 A1 (INTEL CORPORATION) 29 August 2019 (2019-08-29) paragraphs [0021]-[0038] and figures 1-8	1-3
Y	US 2017-0262475 A1 (A9.COM, INC.) 14 September 2017 (2017-09-14) paragraph [0076] and figures 1(a)-10	3
A	KR 10-2019-0049886 A (GOOGLE LLC) 09 May 2019 (2019-05-09) claim 1 and figures 1-12	1-9
A	US 2019-0102646 A1 (XNOR.AI INC.) 04 April 2019 (2019-04-04) claim 1 and figures 1-10B	1-9

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 December 2020	**18 December 2020**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office **189 Cheongsa-ro, Seo-gu, Daejeon** **35208, Republic of Korea**	**KANG MIN JEONG**
Facsimile No. **+82-42-481-8578**	Telephone No. **+82-42-481-8131**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US2020/050258

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2019-0130206 A1	02 May 2019	US 10592767 B2	17 March 2020
US 2019-0266236 A1	29 August 2019	None	
US 2017-0262475 A1	14 September 2017	US 10409856 B2	10 September 2019
		US 9697234 B1	04 July 2017
KR 10-2019-0049886 A	09 May 2019	CN 110178132 A	27 August 2019
		EP 3516537 A1	31 July 2019
		JP 2019-536135 A	12 December 2019
		JP 6625789 B2	25 December 2019
		KR 10-2050334 B1	29 November 2019
		US 10146768 B2	04 December 2018
		US 2018-0210874 A1	26 July 2018
		WO 2018-140099 A1	02 August 2018
US 2019-0102646 A1	04 April 2019	US 10579897 B2	03 March 2020