Separating Skills and Concepts for Novel Visual Question Answering

Spencer Whitehead[1*], Hui Wu[2], Heng Ji[1], Rogerio Feris[2], Kate Saenko[2,3]
[1]UIUC   [2]MIT-IBM Watson AI Lab, IBM Research   [3]Boston University
{srw5,hengji}@illinois.edu, {wuhu,rsferis}@us.ibm.com, saenko@bu.edu

Abstract

Generalization to out-of-distribution data has been a problem for Visual Question Answering (VQA) models. To measure generalization to novel questions, we propose to separate them into "skills" and "concepts". "Skills" are visual tasks, such as counting or attribute recognition, and are applied to "concepts" mentioned in the question, such as objects and people. VQA methods should be able to compose skills and concepts in novel ways, regardless of whether the specific composition has been seen in training, yet we demonstrate that existing models have much to improve upon towards handling new compositions. We present a novel method for learning to compose skills and concepts that separates these two factors implicitly within a model by learning grounded concept representations and disentangling the encoding of skills from that of concepts. We enforce these properties with a novel contrastive learning procedure that does not rely on external annotations and can be learned from unlabeled image-question pairs. Experiments demonstrate the effectiveness of our approach for improving compositional and grounding performance.[1]

1. Introduction

When humans answer questions, such as in Visual Question Answering (VQA), we first interpret the question, dissecting its content into parts (like concepts, relations, actions, question types), and then we select and execute the *skill* (or plan/program) necessary to produce an answer based on this information and the relevant knowledge base (*e.g.,* the image) [25, 34, 54, 57]. The skills needed to produce an answer are general and can be applied to (composed with) many types of question-specific content. For example, if one can answer questions about "*colors*" for a variety of objects as well as recognize and answer questions about "*cars*", then questions like "*What color is the car?*" should be straightforward to answer even if this specific composition has yet to be seen (Fig. 1). This ability of seam-
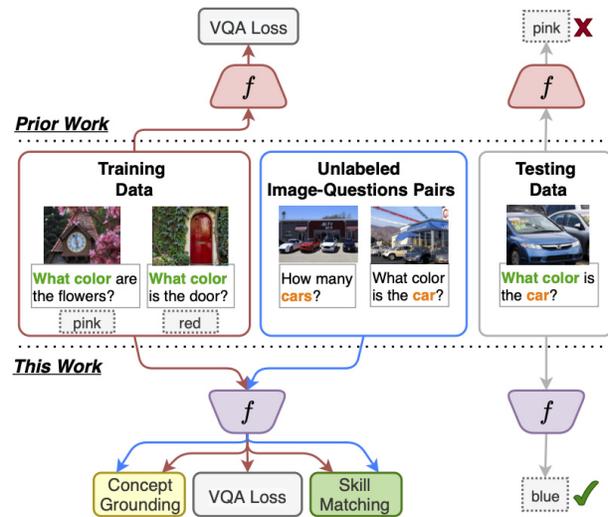

Figure 1. We propose a new view of compositionality in VQA that explores the ability to answer questions about unseen compositions of skills (*e.g.,* color) and concepts (*e.g.,* car). We present a method that learns to separate skills and concepts that can utilize both labeled and unlabeled image-question pairs in order to generalize to novel questions with new skill-concept compositions and new concepts.

lessly adapting and composing conceptual representations with skills is crucial to demonstrating true understanding of VQA and learning to generalize from less labeled data.

Compositionality is recognized as one of the essential properties of human cognition [33], but more research is still needed on incorporating compositionality into models and developing data-efficient, generalizable systems. While much progress has been made to achieve better performance on standard VQA test benchmarks [4, 16, 32, 56], most state-of-the-art models are still designed without any notion of built-in compositionality and tend to entangle skills and concepts in their learned representations. Some previous work has studied the lack of generalization ability of VQA models, and evaluated models using test splits with different answer distributions from the training data. However, this measurement only indirectly addresses the central issue

---

*Work was partly done as an intern at the MIT-IBM Watson AI Lab.
[1]Code: https://github.com/SpencerWhitehead/novelvqa

(lack of compositionality), which manifests itself as poor generalization and over-reliance on language priors [2, 42].

To address these issues, our first contribution is a new view of VQA compositionality, called *skill-concept composition*, and a new evaluation setting that directly targets how VQA models can generalize to novel compositions of skills and concepts. This view is motivated by our observation that, to answer a natural question on real images requires the understanding of two distinct elements: 1) the visual concept referred to by the question; and 2) what information we need to extract from the referred concept. We elucidate this in Sec. 3 and evaluate a number of VQA architectures using this setting and demonstrate that the existing models have much to improve upon to answer novel questions.

We propose a novel approach to improve generalization that utilizes contrastive learning to separate skills and concepts within the internal representations of a model, while jointly learning to answer questions. We use grounding as a proxy to separate concepts so that the model learns to identify a concept in both the question and image, regardless of the specific context. Akin to weakly supervised grounding [3, 17], we train the model to recover a concept mentioned in a given image-question pair by contrasting the multi-modal representation of the masked concept word to the multi-modal representations of words in other questions. We utilize a new way to curate positive and negative examples for the contrastive loss so that the model learns to predict the concept based on relevant visual information rather than using superficial contextual cues. Additionally, our approach learns to separate skills from concepts by contrasting question representations that have the same or different skills. These properties are learned jointly alongside the VQA objective, on top of state-of-the-art models, and are generalizable to new architectures.

Some advantages of our approach are: 1) We learn grounding in a self-supervised manner using the VQA data *alone*, without external annotations. This is in contrast to previous approaches with similar goals that incur large expenses due to annotation requirements [44, 55]. 2) Our method does not rely on answer labels to learn skill-concept separation, so we are able to use *unlabeled* image-question pairs to learn these properties. Consequently, we are able to acquire new concepts and learn to answer questions about them without having labeled data with these concepts, which is pivotal for generalizing to a new domain or novel instances. Moreover, we focus on data-efficient methods and do not use prodigious amounts of data external to VQA, like pre-training approaches [39, 50, 8], which is expensive to obtain and can require prior knowledge of the domain and/or concepts in order to perform well [47, 19].

Our main contributions in this paper are: 1) We present a novel view and evaluation setting for compositionality in VQA, called *skill-concept composition*, which enables a more direct and interpretable evaluation of VQA models on real-image question answering. 2) We propose a novel contrastive learning approach, which combines the supervised VQA objective with self-supervised learning, to achieve skill-concept disentanglement at no additional annotation cost. 3) Our approach shows significant improvements over existing models on novel skill-concept compositions as well as generalization to unlabeled image-question pairs containing unseen concepts.

2. Related Work

VQA and Evaluations. Much progress has been made towards VQA [4, 32, 38, 50]. Large-scale pre-trained transformers [53], inspired by BERT [11], have recently become prevalent [8, 24, 36, 37, 38, 39, 49, 50, 56]. Progress is often measured by the widely adopted human-annotated VQA v2 [16] benchmark as well as other synthetic benchmarks [26, 29, 30]. Various datasets for reducing biases have been presented [2, 16], including VQA-CP [2] that creates distinct distributions of question prefixes and answers between the train and validation splits of VQA v2. Another dataset [30] evaluates on a conglomeration of data from VQA v1 [6] and synthetic data generated from question templates and image annotations, which they break down by vision task. The basic premise of considering questions by task is similar to our "skills", but our setting examines compositional generalization of skills and concepts, where we evaluate on unseen compositions of skills and concepts. Existing work for evaluating generalization [2, 30] does not explore generalization to such novel compositions.

Compositionality and VQA. CLEVR [29] and GQA [26] are two VQA datasets that have received interest in recent years. Both offer *compositional questions*, which means that the questions involve various relational chains (*e.g.,* "*What color is the apple to the left of the bowl on the table?*"). While GQA does not focus on novel compositions, CLEVR does investigate novel compositions of attributes and objects, where, for example, models see cubes of certain colors and cylinders of other colors during training, then the cubes and cylinders have their colors swapped in testing. This is analogous to our setting, although we propose to investigate skill-concept compositions and we experiment with natural questions about real images. Other efforts create compositional models to handle the relational reasoning chains [22, 23, 25, 27, 46]. Our approach implicitly learns compositional capabilities within state-of-the-art multi-modal transformer architectures, unlike these methods that do so explicitly. We also provide results on how well current compositional models (*i.e.,* neural module networks) can generalize to novel questions.

Grounding Visual Concepts. Visual grounding is typically studied on image-caption pairs. Previous work on visual grounding often learns grounding in a weakly supervised
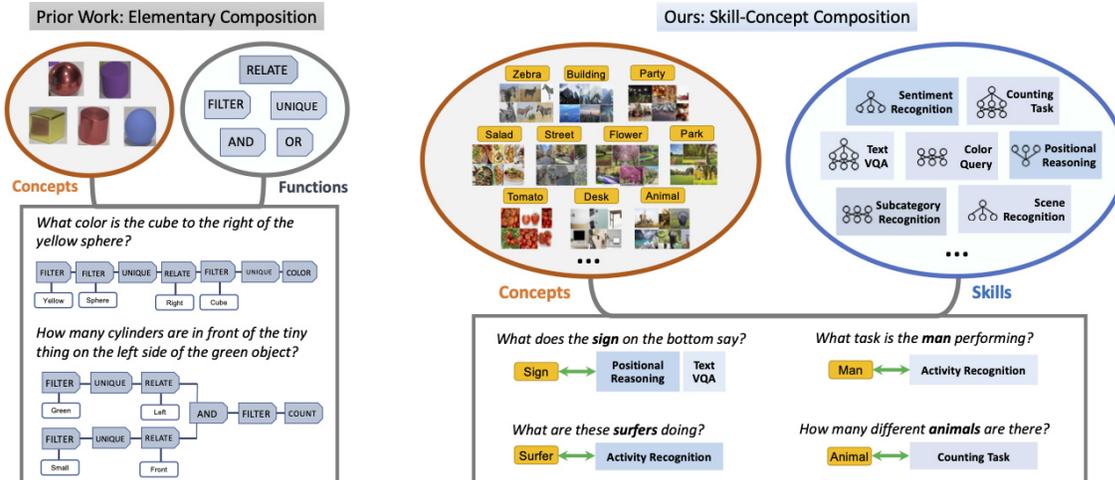
Figure 2. Illustration of skill-concept composition, a new view and evaluation setting for compositionality in VQA.

manner [3, 10, 14, 17], and sometimes leverages contrastive learning techniques [3, 17]. However, grounding concepts that appear in VQA questions has been less studied in previous work, which we explore in our paper. One direction that is somewhat related to visual grounding in VQA is learning VQA models which have interpretable visual attentions [44, 55]. However, these approaches usually rely on human annotations of the most influential regions for VQA answer prediction [9, 28], which are expensive to obtain and do not directly associate concept mentions with visual regions. Our work learns to ground concepts mentioned in questions, which facilitates compositional VQA modeling, and does so without using any additional annotations.

3. Skill-Concept Composition in VQA

We propose a novel, compositional view of VQA, called *skill-concept composition* (Fig. 2). *Concepts* are objects and other visually-grounded words or phrases. By *skills*, we refer to the collection of high-level vision understanding processes involved in answering common questions about real-world images. These skills operate on concepts and vary in terms of input/output representation complexity and the necessary reasoning process. Our taxonomy of these skills is extracted from annotating a subset of the VQA v2 questions as well as taking inspiration from prior work on VQA.[2] Skills are generally standalone from each other and have been studied independently in VQA literature (*e.g.*, TextVQA [48], positional reasoning [26], or counting [1]).

We make an important yet intuitive observation about these VQA skills: to answer a question, it often requires the application of only a small number of skills (most often one) to one or more concepts in the image (Fig. 2). This ob-

---

[2]Please refer to Appendix A for a complete skill list.

servation provides an interpretable view of a model's generalization ability to out-of-distribution data: a model should learn that a skill is a separable process that can be applied to different concepts, and that the prediction process should not be tied to specific concepts co-occurring with this skill during training. This explicit notion of skill-concept separation underlies the contributions of this paper, including a new novel-VQA evaluation method which we will introduce next, as well as a new framework to weakly learn VQA models that can answer novel questions (Sec. 4).

Novel-VQA Evaluation. While conceptually intuitive, this skill-concept view offers natural ways to guide the evaluation of VQA models in terms of out-of-distribution data. In our experiments, we evaluate two novel-VQA settings: 1) answering questions on novel compositions of skills and concepts; 2) answering questions about concepts for which the model has not seen any answers before.

Comparison to Existing Evaluations. Our evaluation protocol is different from existing VQA benchmarks that also aim to measure VQA models' generalization ability. VQA-CP [2] builds train-test splits from VQA v2/v1 [6, 16] with distinct answer distributions by greedily dividing the questions based on their annotated question types (*i.e.*, first few words in the question: "*how many*", "*is the*",...) and answers, but this does not capture skill-concept compositions because these question types do not necessarily correspond to skills (*e.g.*, "*Is the dog waiting?*" requires action recognition and "*Is the sky blue?*" requires color recognition yet both have question type "*is the*"), and the same skill-concept composition can appear in training and testing, which violates our novel-VQA setting. TDIUC [30] evaluates VQA accuracy on different categories of task types, without regard for the concepts in the questions. [52] creates testing splits such that at least one word of a question is un-
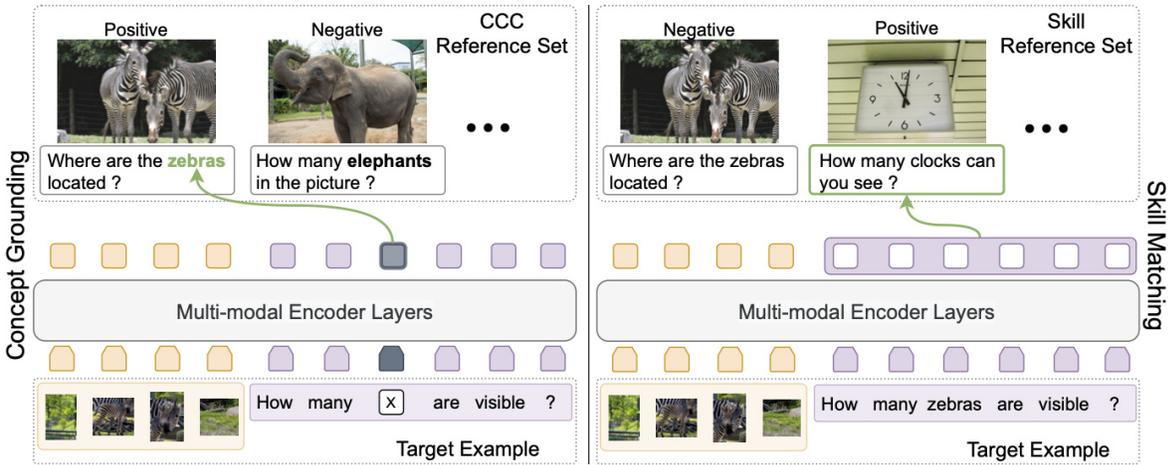
Figure 3. Overview of our approach. Left: We learn to ground concept representations by contrasting the multi-modal representations of a masked concept token in the target example and words in other questions. Right: We encode skills in the summary representations of the question by contrasting with summary representations of other questions with the same (positive) or different (negative) skills.

seen during training, which does not consider skills. None of these benchmarks directly address and evaluate skill-concept compositionality like our evaluation protocol.

Skill-Concept vs. Elementary Compositions. Existing compositional evaluations primarily define compositions as relational reasoning chains associated with questions [26, 29] (Fig. 2), which are suited for learning programs of elementary operations to answer questions. There are two main issues with applying this existing compositional view to real-image question answering. First, the concepts and their attributes are over-simplified and not representative of the diverse visual presentations in the real world. Second, the kinds of compositional questions in these synthetic datasets rarely appear in natural questions about real-world images. Our proposed *skill-concept composition* view is more applicable to real-world VQA and, as a result, can better represent the capabilities that people care about in real-image question answering.

4. Approach

Preliminaries. We assume that we are given a partially labeled dataset of image I, question Q, and answer labels A, where $(I^a, Q^a, A^a) \in \mathcal{D}^a$ has labels and $(I^u, Q^u) \in \mathcal{D}^u$ does not. Typically, models are trained with a VQA loss [51] using the labeled dataset, $\mathcal{D}^a$. Given an example (I, Q), image region features, $g_v(I) = \{v_1, ..., v_M\}$, and question token embeddings, $g_w(Q) = \{x_1, ..., x_N\}$, are extracted and input to a multi-modal encoder to produce multi-modal representations of both modalities, $f(g_v(I), g_w(Q)) = (\{z_m\}_{m=1}^{M}, \{h_i\}_{i=1}^{N})$, where z_m and h_i are the image and text multi-modal representations, respectively. An answer is predicted by pooling the encoded representations to a single representation (or

using a CLS token as input [11]), which is then input to a softmax output layer. We build upon this basic VQA setup to learn skill-concept separation, and uniquely take advantage of both labeled and unlabeled data.

Overview. We aim to learn separable skills and concepts, such that we can compose them to answer novel questions. To do so, the model should recognize that concepts mentioned in the question are manifested by their appearances in the image (*i.e.*, grounding) and that skills should be identifiable regardless of the concepts in the question or image. Gathering supervision for identifying concepts in the question, grounding them in the image, and labeling questions with skills would be very costly. Therefore, we propose to learn skill-concept separation in a self-supervised manner using contrastive learning [7, 41]. Illustrated in Fig. 3, we train the model with two additional contrastive objectives jointly with the VQA objective: *concept grounding* (Sec. 4.1), which learns grounded concept representations, and *skill matching* (Sec. 4.2), which encodes concept-agnostic representations of skills. For each of our objectives, the model is presented with a *target example* and a *reference set* of positive and negative examples sampled from carefully curated *candidate references*. Each objective trains the model to make the representation from the target example similar to those of the positive ones. We expound our training procedure for learning these objectives jointly with VQA in Sec. 4.3. In the following, for brevity, we put specific details of functions/settings in Appendix B.

4.1. Concept Grounding

To learn grounded representations of concepts, we mask the concept mention from the target question and then train the model to recover this concept mention, using the multi-

Positive Examples
How many trees?   *What is on the trees?*

Negative Examples
What type of vehicle is in the picture?   *What kind of animal is shown in the image?*

What kind of building is in the distance?   *What kind of building is in the background?*

Target Concept

What kind of trees are in this photo?
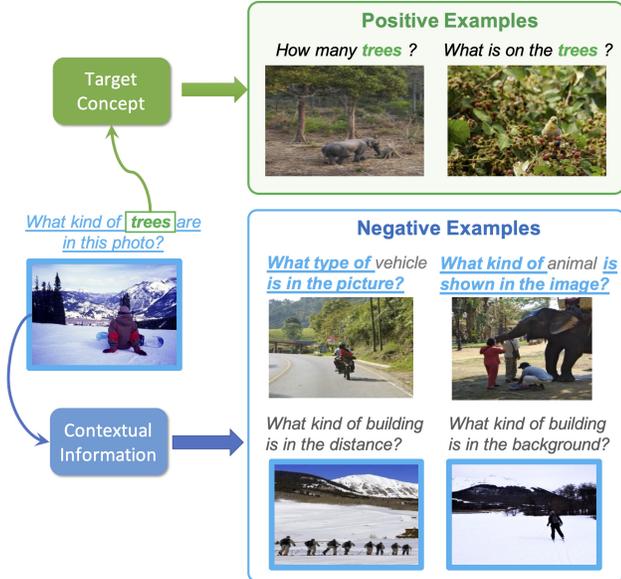
Contextual Information

Figure 4. Target example (left) and a reference set of positive/negative examples (right) from our CCC references.

modal contextual information, by pointing to the same concept mention in examples in the reference set (Fig. 3).

Concept Discovery. We first identify the concept words that can be grounded in images. While this can be done with different methods [12, 20, 31], we simply use heuristics. We use POS tagging and lemmatization [21] to identify the 400 most frequent nouns in VQA v2 and then we filter out concepts that cannot be grounded (*e.g.*, "*time*"). For a given question, Q, we want to find examples that have co-occurring mentions of a concept and the appearance of that concept in the image. It is likely that if a question about an image mentions a concept, then that concept may appear in the image [15]. Therefore, we identify the set of questions that mention the same concept, c, call it $\tilde{\mathcal{R}}_g^+(I, Q, c)$, which we consider as candidate positive references for Q. The set of questions not mentioning any of the same concepts are considered as candidate negative references, call it $\tilde{\mathcal{R}}_g^-(I, Q, c)$. To increase the likelihood of the concept appearing in the image, we employ a set of NLP-based heuristics to remove questions whose images may not contain the concept, such as counting questions with an answer of "0".

Concept-Context Contrastive (CCC) References. Given a target question, Q, and a target concept mention, c, in Q, we could simply create reference sets by randomly sampling positive and negative examples [3, 41] from $\tilde{\mathcal{R}}_g^+(I, Q, c)$ and $\tilde{\mathcal{R}}_g^-(I, Q, c)$, respectively, based solely on if the question contains c or not. However, we propose a novel reference example filtering strategy to encourage concept grounding. Our motivation is that, during VQA training, a concept often co-occurs with certain types of visual

scenes or language priors. So the positive and negative examples should force the model to not rely on superficial cues when contrasted against the target example and, instead, look at the correct visual regions. Our solution is to build sets of refined reference candidates, $\mathcal{R}_g^+(I, Q, c)$ and $\mathcal{R}_g^-(I, Q, c)$, for each (I, Q, c) tuple to ensure that the co-occurrence factor present in the dataset can be reduced. As shown in Fig. 4, we want to find positive examples that also contain the concept "*tree*", but with distinct visual scenes and questions from the target. For negative examples, we seek distractors that are similar to the target in terms of the question or visual scene (*e.g.*, mountains with skiers in Fig. 4), but do not reference "*tree*". To achieve this, we first represent the context of c by masking out c in the question and inputting the masked questions and the image into off-the-shelf feature extractors to obtain question context representation, q, and image representation, v.[3] We measure the contextual similarity by:

$$\xi = \beta \cos(q, q') + (1 - \beta) \cos(v, v'), \qquad (1)$$

where β is a scalar and (v, q) and (v', q') are the representations from target and candidate examples, respectively.

To select positive examples from $\tilde{\mathcal{R}}_g^+(I, Q, c)$, we use $\beta = 0.6$ and sample a set, $\mathcal{R}_g^+(I, Q, c)$, of N^+ examples that *minimize* ξ as our candidate positive examples for (I, Q, c). For negatives, we apply two settings of β that *maximize* ξ: $\beta = 0.7$, which favors examples with more textual similarity, and $\beta = 0.3$, which prioritizes images with similar visual context. We select N^- examples from each setting as our candidate negative examples, $\mathcal{R}_g^-(I, Q, c)$. Illustrated in Fig. 4, when sampling reference sets from these two sets of candidates, the examples encourage the model to learn the specific correspondence between the concept mention in the question and its appearance in the image. Intuitively, the model must to learn to ground the concept mention in the presence of the distractors.

Concept Grounding Loss. Let (I, Q) and c be the target example and target concept mention, respectively, and let $\mathcal{X} = \{(I_k, Q_k)\}_{k=1}^K$ be a reference set. Let k^* be the index of the positive example in $\mathcal{X}$ sampled from $\mathcal{R}_g^+(I, Q, c)$, while the other $K - 1$ examples are negative examples from $\mathcal{R}_g^-(I, Q, c)$. Let w_i be the token in Q that refers to the concept c. We mask out w_i and input this masked version of the question along with the corresponding image into the model, f, which outputs multi-modal representations from which we extract the representation of the masked concept token, h_i. Next, we individually feed the examples from $\mathcal{X}$ into the model to obtain each token representation $\hat{h}_{k,j}$, where j is the index of a token in Q_k. Let $\hat{h}_{k^*, j^*}$ be the representation of the concept mention in the positive example's question. Our grounding loss is an NCE objective [7, 41]

---

[3]We use BERT [11] for questions and ResNet101 [18] for images.

that requires the model to match the multi-modal representation of the masked concept mention to the representation of the same concept mention in the reference set:

$$\mathcal{L}_g = -\log \frac{\exp(\text{sim}(\phi_g(h_i), \phi_g(\hat{h}_{k^*,j^*})))}{\sum_{k,j} \exp(\text{sim}(\phi_g(h_i), \phi_g(\hat{h}_{k,j})))}, \quad (2)$$

where ϕ_g is a learned projection function and $\text{sim}(\cdot, \cdot)$ is a similarity function (*e.g.*, dot product or cosine similarity). To correctly match h_i with $\hat{h}_{k^*,j^*}$, the model must encode the visual features that match between the images of these examples in both token representations. Our CCC references encourage these representations of the concept mention in the positive example and the masked concept mention to be grounded to the right visual regions as the model cannot rely on superficial textual or visual co-occurrences.

4.2. Skill Matching

Contrary to visual concepts, the essential skill needed to answer a certain question is largely independent of image appearances and mentions of concepts in the question. For example, *counting* questions should share a similar process to produce an answer: image areas associated with the subjects of counting are summarized to make the count prediction. This process should be independent of the type of objects being asked about. In other words, we seek to learn summary representations of questions that share the essential steps to infer the answer and are invariant to concepts.

Skill References. A straightforward approach to learn skills is to annotate questions which explicitly require the same reasoning steps. This annotation is readily available on synthetic datasets [26, 29], but not available on datasets involving real-world images and questions. Instead, we propose to mine sets of contrasting examples to learn which questions require the same/different skills, matching questions with the same skills. Since the skills required for the question are typically indicated by the words of the question, we identify questions that are semantically similar. Essentially, questions that require the same skill (*e.g.*, "*What color ...*") should be related to one another, regardless of the specific concept mentions in the question. So, for each question, we mask out the concept words and we compute their BERT [11] representations. For a given (I, Q), the set of positive reference examples, $\mathcal{R}_s^+(I, Q)$, are sampled from the top-200 most similar questions using BERT representation, and the set of negative examples, $\mathcal{R}_s^-(I, Q)$, are randomly chosen from the rest of the dataset.

Skill Matching Loss. For a given target example, (I, Q), let h be a summary representation of the target question. This can be computed using a special input token like BERT [11] or via a pooling operation on all question token representations output from the encoder. We sample a reference set of image-question pairs, $\{(I_l, Q_l)\}_{l=1}^L$, where

the positive example, Q_{l^*} from $\mathcal{R}_s^+(I, Q)$, shares the same skill as the target question, and the rest of the reference set are negative examples from $\mathcal{R}_s^-(I, Q)$. Let $\hat{h}_l$ be a summary representation for a question in the reference set. Shown in Fig. 3, our skill matching loss is defined as

$$\mathcal{L}_s = -\log \frac{\exp(\text{sim}(\phi_s(h), \phi_s(\hat{h}_{l^*})))}{\sum_l \exp(\text{sim}(\phi_s(h), \phi_s(\hat{h}_l)))}, \quad (3)$$

where $\hat{h}_{l^*}$ is the positive example representation and ϕ_s is another learned projection function.

4.3. Training Procedure

With our losses, we use a multi-tasking learning procedure [13, 40], where at each step we employ our objectives with probability p_{sep} or not with probability $1 - p_{\text{sep}}$. During training, we always first sample an instance from the labeled data, $\mathcal{D}^a$, and update the model by minimizing the VQA objective. If at the current iteration we do not use our skill and concept objectives, then we only use the VQA objective. Otherwise, we first use the VQA objective and then apply our other objectives. Both objectives are computed in the same fashion: for $\mathcal{L}_g$ (or $\mathcal{L}_s$), we sample a target example from $\mathcal{D}^a \cup \mathcal{D}^u$ along with N_r^+ positive examples from $\mathcal{R}_g^+$ (or $\mathcal{R}_s^+$) as well as N_r^- negative examples from $\mathcal{R}_g^-$ (or $\mathcal{R}_s^-$), combine the sampled references to form the current reference set, and compute the loss term. We then sum $\mathcal{L}_g$ and $\mathcal{L}_s$, and update the model to minimize the negative sum.

5. Experiments

Data and Settings. We run our experiments on VQA v2 [16], which contains real images, human-written questions, and a variety of skills required to answer the questions. Since the goal of this work is to examine a model's performance on different types of novel questions, it requires the availability of answer annotations for the test data. Since the annotations of test-dev and test-std sets of VQA v2 are not publicly available, we use questions from the validation set for testing as is common [2, 45]. We do not train or tune hyperparameters with the validation set; it is strictly used for evaluation. We compare performance using the VQA accuracy [6] on different splits of novel questions. Details are provided in Appendix C.

Model Comparisons. We select a set of recent VQA models to benchmark their novel-VQA performance. The first category is compositional models [5, 22, 23, 46]. We use **StackNMN** [22] and **XNM** [46], which are designed to handle compositional questions, like those in CLEVR [29], and have state-of-the-art performance on these datasets while also being applicable to real images without supervision from functional programs or image scene graphs.

The second type of model we experiment with is transformer-based [8, 11, 35, 38, 39, 50, 53, 56]. We

Model	Counting				Color		Subcat.	Overall
	animal	{animals}	{vehicles}	{electronics}	animal	{dishware}	vegetable	
XNM [46]	56.02	48.32	44.35	51.94	77.22	65.73	57.33	57.27
StackNMN [22]	54.22	47.56	46.10	52.83	76.57	69.22	57.17	57.67
X-Att [50]	58.94	56.28	46.30	57.05	73.15	67.29	57.25	59.47
X-BERT [8]	63.58	54.58	42.34	56.84	75.88	70.31	58.96	60.36
Base	62.57	59.19	48.33	61.84	76.57	72.91	58.33	62.82
Ours	**65.16**	**59.87**	**50.75**	**62.21**	**77.45**	**73.76**	**61.04**	**64.32**

Table 1. VQA accuracy on novel skill-concept compositions. The highest and the lowest numbers of each experiment are emphasized.

Model	lamp	fruit	fridge	surfer	flag	skateb.	oven	sheep	banana	zebra	Overall
XNM [46]	53.69	50.23	57.98	72.68	36.58	70.16	53.49	54.96	**52.35**	61.50	56.36
StackNMN [22]	54.27	46.10	58.97	74.10	**41.31**	74.11	56.30	57.12	50.98	61.25	57.45
X-Att [50]	46.26	33.10	51.52	67.68	31.53	69.73	51.69	49.83	41.71	**64.93**	50.80
X-BERT [8]	44.43	30.72	50.83	61.60	32.46	66.05	48.10	50.32	43.10	57.06	48.47
Base	55.14	52.99	59.06	74.12	39.05	71.67	56.60	63.31	49.83	56.05	57.78
Ours	**57.40**	**54.40**	**60.92**	**74.36**	40.15	**75.27**	**59.91**	**64.04**	50.78	60.77	**59.80**

Table 2. VQA Accuracy on individual novel-concept split. *skateb.* refers to *skateboarder*.

use two top-performing transformer architectures from this model family: 1) a two-stream, cross-attention model [50] (**X-Att**), which has modality specific branches and cross-attentions in early layers followed by multi-modal layers later in the network; and 2) a vision-and-language transformer model [8] (**X-BERT**) that acts as multi-modal encoder throughout the entire network. For fair comparison, we do not use pre-training, same as our model, since we are specifically interested in the generalization ability of data-efficient models without requiring large-scale (*e.g.,* 9M+ image-text pairs), in-domain data external to VQA [19, 47].

Lastly, for our base model, we employ a variant of the standard multi-modal transformer where image features can be attended by both the CLS token and language features. When the base model is trained without the proposed skill-concept contrastive losses, it serves as a baseline model (denoted by **Base**).

Detailed settings are in Appendix C and Appendix D.

5.1. Novel Skill-Concept Composition VQA

We select three prevalent and common skills present in VQA v2: counting, color querying and subcategory recognition. For each skill, we remove the data labels for its co-occurring questions with one concept or a set of multiple concepts which can form a distinct category from training, and then test on these compositions. The concepts (or concept groups) are randomly sampled with two criteria: each skill-concept composition contains reliable amount of test data to measure accuracy and the compositions have diverse coverage across the dataset (more details in Appendix A).

Tab. 1 shows the VQA accuracy on each of the novel compositional subset. Interestingly, although neural module networks are designed to explicitly break down the question answering process into subtasks, which in principle should help with adapting these subtasks to new questions

and thus generalize better, they yield lower performance than transformer models. This may be due to the effective feature learning capacity of self-attention mechanisms. Among all transformer models, our base encoder achieves competitive performance to existing networks, demonstrating that it is a strong baseline among multi-modal transformers. Finally, our contrastive learning framework outperforms the baseline and all other approaches across each novel composition set. This supports the effectiveness of our framework for generalization to new compositions.

5.2. Novel-Concept VQA

For this experiment, we are interested in the setting where models are never trained to answer questions about a concept but can make use of the unlabeled image-question pairs, and then are tested on questions that have mentions of this given concept. Similar to the previous experiment, these concepts were sampled to maximize coverage as well as maintaining a reasonable test size. This setting is more challenging than the previous experiment since the model misses the VQA training supervision on any questions that have the given concept, as opposed to any questions that have both the given concept and a certain skill.

We provide qualitative examples of novel-concept VQA in Appendix E and report quantitative results in Tab. 2. For this more challenging setting of novel question answering, on average, two of the existing transformer architectures underperform other models by a noticeable margin. This may suggest that the transformer architectures, which perform well on large-scale vision and language pre-training, may have difficulty specializing to the VQA task. The Base model slightly outperforms neural module networks. Lastly, our framework again outperforms all models on average, demonstrating its value in improving VQA generalization ability on novel concepts.
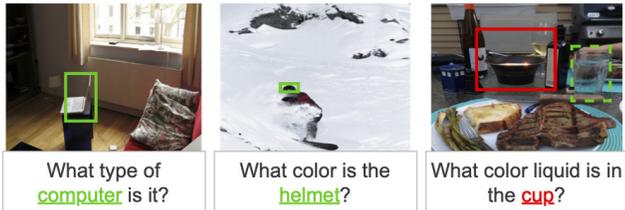
5.3. Analysis

Figure 5. Correct and incorrect grounding examples. We visualize the most similar visual region to the concept in the question.

Concept Grounding. Since our approach learns to ground concepts without strong supervision, we would like to test its grounding abilities directly. To obtain an evaluation set, we manually annotate 320 image-question-concept tuples with the visual regions in the image that corresponds to the concept in each tuple. Candidate visual regions are found using Faster-RCNN [43]. We use recall@5 as our grounding metric, considering a grounding correct if the correct visual region falls within the top 5 most similar visual regions to the target concept token. The model trained with our framework achieves a grounding recall of **59.12**, compared to **43.71** of Base. Note that our framework obtained this improvement with no additional training data for grounding. As shown in Fig. 5, our model can often correctly ground a variety of objects, but can be fooled by ambiguous looking concepts like the candle in the incorrect example. Further, it is challenging to learn to differentiate concepts that almost always co-occur (*e.g.*, "*shirt*" and "*person*").

Loss Ablation. We ablate our losses by sampling three novel compositions and three novel concepts and report their average performance in Tab. 3. Adding our losses leads to consistent gains, with top performance achieved with our full framework. When used alone, our grounding loss seems to contribute a larger benefit compared to the skill loss. Nonetheless, the best performance is achieved by combining the two components, further supporting the value of skill and concept separation. We also experiment with a masked language modeling (MLM) objective [11] that replaces our losses. Our objectives perform better than the MLM objective, implying that the improvements our objectives offer are not simply due to additional data.

CCC Reference Sets. To study the effects of our CCC reference set selection strategy, we compare it with the commonly used random sampling method [3, 41] and report novel-concept VQA results in Tab. 4. We train both models with our full framework, the only difference being the reference set construction method for the concept loss. Both models improve upon the Base model, with our reference set construction method offering more consistent gains.

Existing Benchmarks. We also evaluate on VQA-CP [2]

Model	Avg. Novel Count	Avg. Novel Concepts
Base	58.03	60.42
Base+MLM	58.41	60.25
Base+$\mathcal{L}_s$	58.83	61.85
Base+$\mathcal{L}_g$	59.80	62.06
Ours	**60.71**	**63.19**

Table 3. Effect of using different losses on novel skill-concept composition and novel-concept VQA.

Model	Lamp	Fruit	Fridge	Surfer	Flag
Base	55.14	52.99	59.06	74.12	39.05
Random	**+2.80**	+0.49	-0.16	-0.14	+0.50
CCC (Ours)	+2.26	**+1.41**	**+0.98**	**+0.24**	**+1.10**

Table 4. Comparing different reference set construction schemes for concept learning across five different concepts.

Model	VQA-CP	Test-dev	Test-std
Base	40.98	69.60	69.99
Ours	**41.71**	**69.78**	**70.09**

Table 5. Single model VQA performance on VQA-CP [2] and VQA v2 test-dev and test-std splits. Both models see the exact same training data (*i.e.*, no compositions/concepts are removed).

and the test-dev/test-std splits of VQA v2 (Tab. 5). While we see gains in general, notably, our approach is able to improve on VQA-CP without extra annotations, ensembling/tuning, or a performance drop on VQA v2.

6. Conclusions

We propose a new setting for generalization in VQA: measuring the ability to compose the skills needed to answer a question and the visual concepts that should be grounded to the image. We show that existing approaches have difficulty generalizing to unseen compositions of these two factors. We present a novel approach that implicitly disentangles skills and concepts, while grounding concepts visually, using a contrastive learning procedure. Our approach is able to learn from unlabeled VQA data in order to answer questions about previously unseen concepts. Results on the VQA v2 show that the proposed framework can achieve state-of-the-art performance on novel skill-concept compositions as well as generalize from unlabeled data.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *AAAI*, 2019. 3

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018. 2, 3, 6, 8

[3] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *CVPR*, 2019. 2, 3, 5, 8

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1, 2

[5] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016. 6

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2, 3, 6

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 4, 5

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 2, 6, 7

[9] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *EMNLP*, 2016. 3

[10] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *ICCV*, 2019. 3

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2, 4, 5, 6, 8

[12] Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander Berg, et al. Detecting visual text. In *NAACL*, 2012. 5

[13] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *ACL*, 2015. 6

[14] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015. 3

[15] Siddha Ganju, Olga Russakovsky, and Abhinav Gupta. What's in a question: Using visual questions as a form of supervision. In *CVPR*, 2017. 5

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 1, 2, 3, 6

[17] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *ECCV*, 2020. 2, 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[19] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv preprint arXiv:2102.00529*, 2021. 2, 7

[20] Jack Hessel, David Mimno, and Lillian Lee. Quantifying the visual concreteness of words and topics in multimodal datasets. In *NAACL*, 2018. 5

[21] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 5

[22] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *ECCV*, 2018. 2, 6, 7

[23] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, 2017. 2, 6

[24] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[25] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *ICLR*, 2018. 1, 2

[26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2, 3, 4, 6

[27] Drew A Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. In *NeurIPS*, 2019. 2

[28] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*, 2018. 3

[29] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 2, 4, 6

[30] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, 2017. 2, 3

[31] Gitit Kehat and James Pustejovsky. Integrating vision and language datasets to measure word concreteness. In *IJCNLP*, 2017. 5

[32] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018. 1, 2

[33] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 2017. 1

[34] Wendy Lehnert. Human and computational question answering. *Cognitive Science*, 1977. 1

[35] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 6

[36] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 2

[37] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[38] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2, 6

[39] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2, 6

[40] Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. In *ICLR*, 2016. 6

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4, 5, 8

[42] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*, 2018. 2

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 8

[44] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *ICCV*, 2019. 2, 3

[45] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *CVPR*, 2019. 6

[46] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, 2019. 2, 6, 7

[47] Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*, 2020. 2, 7

[48] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 3

[49] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. 2

[50] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 2, 6, 7

[51] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*, 2018. 4

[52] Damien Teney and Anton van den Hengel. Zero-shot visual question answering. *arXiv preprint arXiv:1611.05546*, 2016. 3

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 6

[54] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. In *ICLR*, 2016. 1

[55] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. In *NeurIPS*, 2019. 2, 3

[56] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 1, 2, 6

[57] Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. Vision skills needed to answer visual questions. *Proc. ACM Hum.-Comput. Interact.*, 2020. 1