

Zero-Shot Compositional Concept Learning

Guangyue Xu
Michigan State University
xuguang3@msu.edu

Parisa Kordjamshidi
Michigan State University
kordjams@msu.edu

Joyce Y. Chai
University of Michigan
chaijy@umich.edu

Abstract

In this paper, we study the problem of recognizing compositional *attribute-object* concepts within the zero-shot learning (ZSL) framework. We propose an episode-based cross-attention (EpiCA) network which combines merits of cross-attention mechanism and episode-based training strategy to recognize novel compositional concepts. Firstly, EpiCA bases on cross-attention to correlate *concept-visual* information and utilizes the gated pooling layer to build contextualized representations for both images and concepts. The updated representations are used for a more in-depth multi-modal relevance calculation for concept recognition. Secondly, a two-phase episode training strategy, especially the transductive phase, is adopted to utilize unlabeled test examples to alleviate the low-resource learning problem. Experiments on two widely-used zero-shot compositional learning (ZSCL) benchmarks have demonstrated the effectiveness of the model compared with recent approaches on both conventional and generalized ZSCL settings.

1 Introduction

Humans can recognize novel concepts through composing previously learnt knowledge - known as compositional generalization ability (Lake et al., 2015; Lake and Baroni, 2018). As a key critical capacity to build modern AI systems, this paper investigates the problem of zero-shot compositional learning (ZSCL) focusing on recognizing novel compositional *attribute-object* pairs appeared in the images. For example in Figure 1, suppose the training set has images with compositional concepts *sliced-tomato*, *sliced-cake*, *ripe-apple*, *peeled-apple*, etc. Given a new image, our goal is to assign a novel compositional concept *sliced-apple* to the image by composing the element concepts, *sliced* and *apple*, learned from the training data. Although *sliced* and *apple* have appeared

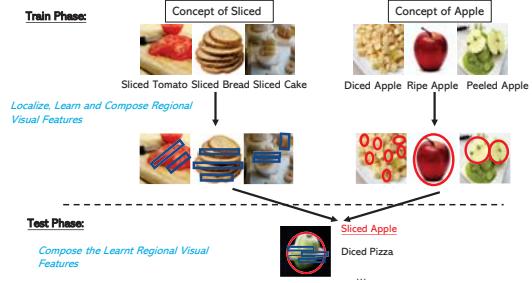


Figure 1: Given the concepts of *sliced* and *apple* in the training phase, our target is to recognize the novel compositional concept *slice apple* which doesn't appear in the training set by decomposing, grounding and composing concept-related visual features.

with other objects or attributes, the combination of this attribute-object pair is not observed in the training set.

This is a challenging problem, because objects with different attributes often have a significant diversity in their visual features. While *red apple* has similar visual features as the *apple* prototype, *sliced apple* presents rather different visual features as shown in Fig 1. Similarly, same attributes can have different visual effects depending on the modified objects. For example, *old* has different visual effect in objects of *old town* compared to objects of *old car*.

Despite recent progress (Misra et al., 2017; Li et al., 2020), previous works still suffer several limitations: (1) Most existing methods adopt metric learning framework by projecting concepts and images into shared latent space, and focus on regularizing the structure of the latent space by adding principled constraints without considering the relationship between concepts and visual features. Our work brings a new perspective, the relevance-based framework inspired by Sung et al., to conduct compositional concept learning. (2) Previous works represent concept and image by the same vector regardless of the context it occurs. However, cross

concept-visual representation often provides more grounded information to help in recognizing objects and attributes which will consequently help in learning their compositions.

Motivated by the above discussions, we propose an Episode-based Cross Attention (EpiCA) network to capture multi-modal interactions and exploit the visual clues to learn novel compositional concepts. Specifically, within each episode, we first adopt cross-attention encoder to fuse the concept-visual information and discover possible relationships between image regions and element concepts which corresponds to the localizing and learning phase in Fig.1. Second, gated pooling layer is introduced to obtain the global representation by selectively aggregating the salient element features corresponding to Fig. 1’s composing phase. Finally, relevance score is calculated based on the updated features to update EpiCA.

The contribution of this work can be summarized as follows: 1) Different from previous work, EpiCA has the ability to learn and ground the attributes and objects in the image by cross-attention mechanism. 2) Episode-based training strategy is introduced to train the model. Moreover, we are among the first works to employ the transductive training to select confident unlabelled examples to gain knowledge about novel compositional concepts. 3) Empirical results show that our framework achieves competitive results on two benchmarks in conventional ZSCL setting. In the more realistic generalized ZSCL setting, our framework significantly outperforms SOTA and achieves over $2\times$ improved performance on several metrics.

2 Related Work

Compositional Concept Learning. As a specific zero-shot learning (ZSL) problem, zero-shot compositional learning (ZSCL) tries to learn complex concepts by composing element concepts. Previous solutions can mainly be categorized as: (1) classifier-based methods train classifiers for element concepts and combine the element classifiers to recognize compositional concepts (Chen and Grauman, 2014; Misra et al., 2017; Li et al., 2019a). (2) metric-based methods learn a shared space by minimizing the distance between the projected visual features and concept features (Nagarajan and Grauman, 2018; Li et al., 2020). (3) GAN-based methods learn to generate samples from the semantic information and transfer ZSCL into a tradi-

tional supervised classification problem (Wei et al., 2019).

Attention Mechanism. The attention mechanism selectively use the salient elements of the data to compose the data representation and is adopted in various visiolinguistic tasks. Cross attention is employed to locate important image regions for text-image matching (Lee et al., 2018). Self-attention and cross-attention are combined at different levels to search images with text feedback (Chen et al., 2020b). More recent works refer Transformer (Vaswani et al., 2017) to design various visiolinguistic attention mechanism (Lu et al., 2019).

Episode-based Training. The data sparsity in low-resource learning problems, including few-shot learning and zero-shot learning, makes the typical fine-tuning strategy in deep learning not adaptable, due to not having enough labeled data and the overfitting problem. Most successful approaches in this field rely on an episode-based training scheme: performing model optimization over batches of tasks instead of batches of data. Through training multiple episodes, the model is expected to progressively accumulate knowledge on predicting the mimetic unseen classes within each episode. Representative work includes Matching network (Vinyals et al., 2016), Prototypical network (Snell et al., 2017) and RelNet (Sung et al., 2018).

The related works to EpiCA are RelNet (Sung et al., 2018) and cvcZSL (Li et al., 2019a). Compared with these methods, we have two improvements including an explicit way to construct episodes which is more consistent with the test scenario and a cross-attention module to fuse and ground more detailed information between the concept space and the visual space.

3 Approach

3.1 Task Definition

Different from the traditional supervised setting where training concepts and test concepts are from the same domain, our problem focuses on recognizing novel compositional concepts of attributes and objects which are not seen during the training phase. Although we have seen all the attributes and objects in the training set, their compositions are novel¹.

We model this problem within the ZSL framework where the dataset is divided into the seen

¹We refer concept as compositional concept, element concept as the attribute and the object in the rest of the paper.

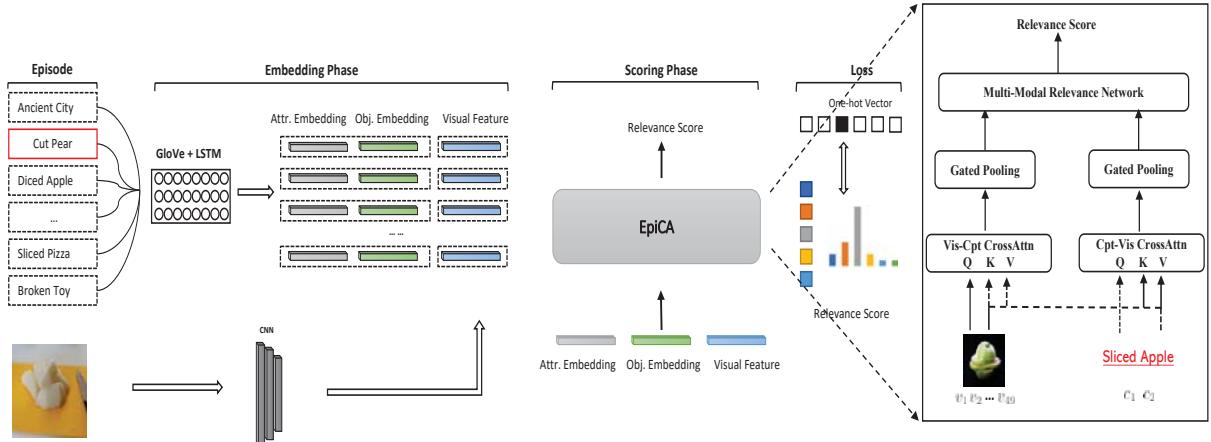


Figure 2: Illustration of the proposed EpiCA framework. It is a two-stage training framework, including the inductive learning and the transductive learning. Both phases are trained on episodes illustrated in Alg. 1.

domain $\mathcal{S} = \{(v_s, y_s) | v_s \in \mathcal{V}^s, y_s \in \mathcal{Y}^s\}$ for training and the unseen domain $\mathcal{U} = \{(v_u, y_u) | v_u \in \mathcal{V}^u, y_u \in \mathcal{Y}^u\}$ for test, where v is the visual feature of image \mathcal{I} which can be extracted using deep convolution networks and y is the corresponding label which consists of an attribute label a and a object label o as $y = (a, o)$ satisfying $a_u \subseteq a_s$, $o_u \subseteq o_s$ and $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$. Moreover, we address the problem in both conventional ZSCL setting and generalized ZSCL setting. In conventional ZSCL, we only consider unseen pairs in the test phase and the target is to learn a mapping function $\mathcal{V} \mapsto \mathcal{Y}^u$. In generalized ZSCL, images with both seen and unseen concepts can appear in the test set, and the mapping function changes to $\mathcal{V} \mapsto \mathcal{Y}^s \cup \mathcal{Y}^u$ which is a more general and realistic setting.

3.2 Overall Framework

As summarized in Fig. 2, EpiCA consists of the cross-attention encoder, gated pooling layer and multi-modal relevance network to compute the relevance score between concepts and images. In order to accumulate the knowledge between images and concepts, EpiCA is trained by episodes including the following two phases:

- *Inductive training phase* constructs episodes from the seen concepts and trains EpiCA based on these constructed episodes.
- *Transductive training phase* employs the self-taught methodology to collect confident pseudo-labeled test items to further fine-tune EpiCA.

3.3 Unimodal Representation

Concept Representation. Given a compositional concept (a, o) , we first transform attribute and object using 300-D *GloVe* (Pennington et al., 2014) separately. Then we use one layer BiLSTM (Hochreiter and Schmidhuber, 1997) to obtain contextualized representation for concepts with d_k hidden units. Instead of using the final state, we maintain the output features for both attribute and object and output feature matrix $C \in \mathbb{R}^{2 \times d_k}$ for each compositional concept.

Image Representation. We extract the visual features using pretrained ResNet (He et al., 2016) from a given image. In order to obtain more detailed visual features for concept recognition, we keep the output from the last convolutional layer of ResNet-18 to represent the image and therefore each image is split into $7 \times 7 = 49$ visual blocks with each block as a 512-dim vector denoted as $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{49})$. Each element represents a region in the image. We further convert v_i with a linear transformation $v_i = \mathbf{W}^\top v_i$, where $\mathbf{W} \in \mathbb{R}^{512 \times d_k}$ is the weight matrix to transfer the image into the joint concept-image space.

3.4 Cross Attention Encoder

Motivation. Previous works usually utilize vector representation for both concepts and images and construct a metric space by pushing aligned images and concepts closer to each other. The potential limitation of such frameworks is that the same vector representations without context information will miss sufficient detailed information needed for grounding and recognizing objects and attributes appeared in the images. We observe that certain vi-

sual blocks in the image can be more related to certain element concept and certain element concept may highlight different visual blocks. Inspired by this observation, our model addresses the previous limitation by introducing cross-attention encoder and constructs more meaningful cross-modality representation for both images and element concepts for compositional concept recognition.

Cross Attention Layer. To fuse and ground information between visual space and concept space, we first design a correlation layer to calculate the correlation map between the two spaces, which is used to guide the generation of the cross attention map. Given an image and a candidate concept, after extracting unimodal representations, the correlation layer computes the semantic relevance between visual blocks $\{v_i\}_{i=1}^{49}$ and element concepts $\{c_j\}_{j=1}^2$ ² with cosine distance and output the final *image-to-concept* relevance matrix as $R \in \mathbb{R}^{49 \times 2}$ with each element r_{ij} calculated using Eq. 1. We can easily have another *concept-to-image* relevance matrix by transposing R .

$$r_{ij} = \left(\frac{v_i}{\|v_i\|_2} \right)^T \left(\frac{c_j}{\|c_j\|_2} \right), i \in [1, 49], j \in [1, 2] \quad (1)$$

In order to obtain attention weights, we need to normalize the relevance score r_{ij} as Eq. 2 as (Chen et al., 2020a).

$$\bar{r}_{ij} = \frac{\text{relu}(r_{ij})}{\sqrt{\sum_{j=1}^n \text{relu}(r_{ij})^2}} \quad (2)$$

After obtaining the normalized attention score, we can calculate the cross-attention representation based on the selected query space Q and the context space V , where $V = K$ in our setting as shown in Fig. 2. Taking *image-to-concept* attention for example, given a visual block feature v_i as query, cross attention encoding is performed over the element concept space C using Eq. 3.

$$\hat{v}_i = \sum_{j=1}^n \alpha_{ij} c_j, \quad \text{s.t.} \quad \alpha_{ij} = \frac{\exp(\lambda \bar{r}_{ij})}{\sum_{j=1}^n \exp(\lambda \bar{r}_{ij})} \quad (3)$$

where λ is the inverse temperature parameter of the softmax function (Chorowski et al., 2015) to control the smoothness of the attention distribution.

²Each compositional concept only has two elements, attribute and object.

Visually-Attended Concept Representation.

The goal of this module is to align and represent concepts with related visual blocks and help further determine the alignment between element concepts and image regions. We use concept embedding as query and collect visual clues using Eq. 3 and the final visually-attended features for compositional concept is $\hat{c} \in \mathbb{R}^{2 \times d_k}$.

Concept-Attended Visual Representation.

An image representation grounded with element concept would be beneficial for compositional concept learning. Following the similar procedure as visually-attended concept representation, we take visual block features as query and concept embedding as context. We can calculate the concept-attended visual representation using Eq. 3. The final result $\hat{v} \in \mathbb{R}^{49 \times d_k}$ represents the concept-attended block visual features with the latent space dimension d_k .

3.5 Gated Pooling Layer

After the cross-attention encoder, the output image features $V = [v_1, \dots, v_{49}] \in \mathbb{R}^{49 \times d_k}$ and concept features $C = [c_1, c_2] \in \mathbb{R}^{2 \times d_k}$ are expected to contain rich cross-modal information. Our target of gated pooling layer is to combine elements to form the final representation for concepts and images separately. Pooling techniques can be directly deployed to obtain such representation. However, we argue that elements should have different effect on the final concept recognition. For example, background visual blocks shouldn't be paid much attention during concept recognition. To address the assumption, we propose gated pooling layer to learn the relative importance of each element and dynamically control the contribution of each element in the final representation. Specially, We apply one linear layers with parameter $W \in \mathbb{R}^{d_k \times 1}$ on the element feature x_i and normalize the output to calculate an attention weight α_i that indicates the relative importance of each element using Eq. 4.

$$x = \sum_i \alpha_i x_i \quad \text{s.t.} \quad \alpha_i = \frac{\exp((W x_i))}{\sum_{k=1}^N \exp((W x_k))} \quad (4)$$

3.6 Multi-Modal Relevance Network

After obtaining the updated features for both images \hat{v}_i and concepts $(\hat{c}, \hat{o})_j$, we introduce the multimodal relevance network shared the spirit as (Sung et al., 2018) to calculate the relevance score

Algorithm 1: Training EpiCA for ZSCL

Input: $\mathcal{D}_{train} = \{(v_m, (a_m, o_m)\}_{m=1}^{|Tr|}$,
 $\mathcal{D}_{test} = \{v_n\}_{i=n}^{|Ts|}$, task size S ,
sample interval t

Output: Multi-Modal Rel. Function f
// Inductive Learning Phase

1 **for** $epoch \leftarrow 1$ **to** E_{ind_max} **do**

2 **for** each image and the corresponding pair in the training set **do**

3 **Construct** an episode
 $[v_p, (a_p, o_p), (a_{n_1}, o_{n_1}), \dots, (a_{n_s}, o_{n_s})]$.

4 **Gated Cross-Attention Encoding**
using Eq. 1, 2, 3 and 4

5 **Calculating** multi-modal relevance score using Eq 5.

6 **Updating** EpiCA.

// Transductive Learning Phase

7 **for** $epoch \leftarrow 1$ **to** E_{trans_max} **do**

8 **if** $epoch \% t == 0$ **then**

9 **Pick** confident samples from unseen set by Eq. 7.

10 **Updating** EpiCA by Eq 9.

as shown in Eq. 5

$$s_{i,j} = g_\phi(\text{concat}[(\hat{v}_i), (\hat{a}, \hat{o})_j]) \quad (5)$$

where g is the relevance function implemented by two layer feed-forward network with trainable parameters ϕ .

In order to train EpiCA, we add Softmax activation on the relevance score to measure the probability of image i belonging to concept j within the current episode as Eq. 5 and update EpiCA using cross-entropy loss.

$$p_j(\hat{v}_i) = \frac{\exp(s_{i,j})}{\sum_{k=1}^C \exp(s_{i,k})} \quad (6)$$

3.7 Training and Prediction

Inductive Training. For each image and the corresponding pair label, we randomly sample negative pairs to form an episode which consists of an image v_p , a positive pair (a_p, o_p) and a pre-defined number n_t of negative pairs in the form of $[v_p, (a_p, o_p), (a_{n_1}, o_{n_1}), \dots, (a_{n_t}, o_{n_t})]$. Then within each episode, we calculate the relevance score between image and all candidate pairs using

Eq. 5. Finally, we calculate the cross entropy loss using Eq. 6 and update EpiCA as shown in Alg. 1.

Transductive Training. The disjointness of the seen/unseen concept space will result in domain shift problems and cause the predictions biasing towards seen concepts as pointed by (Pan and Yang, 2009). Transductive training utilizes the unlabeled test set to alleviate the problem (Dhillon et al., 2019). Specifically, transductive training has a sampling phase to select confident test samples and utilize the generalized cross entropy loss as Eq. 8 to update EpiCA.

Following previous work (Li et al., 2019b), we use threshold-based method as Eq. 7 to pick up confident examples.

$$\frac{p_1(\hat{v}_i)}{p_2(\hat{v}_i)} > \gamma \quad (7)$$

where p is calculated by Eq. 6 and the threshold is the fraction of the highest label probability $p_1(\hat{v}_i)$ and the second highest label probability $p_2(\hat{v}_i)$ which measures the prediction peakiness in current episode. Only confident instances are employed to update EpiCA which is controlled by γ .

Moreover, the recently proposed generalized cross-entropy loss (Zhang and Sabuncu, 2018) is used to calculate the loss for pseudo-labeled test examples as Eq. 8.

$$\mathcal{L}_u = \sum_{(v_i, (a, o)_j) \in \mathcal{U}} \frac{1 - (p_j(\hat{v}_i))^q}{q} \quad (8)$$

where $p_j(\hat{v}_i)$ is the probability of \hat{v}_i belonging to pair $(\hat{a}, \hat{o})_j$ calculated using Eq. 6. $q \in (0, 1]$ is the hyper-parameter related to the noise level of the pseudo labels, with higher noisy pseudo labels requiring larger q .

Finally, the transductive loss is calculated as Eq. 9, where \mathcal{L}_u corresponds to the generalized cross entropy loss from pseudo-labeled test examples and \mathcal{L}_s is the cross entropy loss for the training examples

$$\mathcal{L} = \mathcal{L}_u + \mathcal{L}_s. \quad (9)$$

Prediction. Given a new image with extracted feature v_i , we iterate over all the candidate pairs and select the pair with the highest relevance score as $(\hat{a}, \hat{o}) = \text{argmax}_{\hat{a}, \hat{o}} s_{i,j}(\hat{v}_i, (\hat{a}, \hat{o})_j)$ as Eq. 5 using EpiCA.

4 Experiments

Dataset. We use similar dataset as in (Nagarajan and Grauman, 2018; Purushwalkam et al., 2019) for

both conventional and generalized ZSCL settings with the split shown in Tab. 1. Notably, generalized ZSCL setting has additional validation set for both benchmarks which allows cross-validation to set the hyperparameters. The generalized ZSCL evaluates the models on both seen/unseen sets.

- MIT-States (Isola et al., 2015) has 245 objects and 115 attributes. In conventional ZSCL, the pairs are split into two disjoint sets with 1200 seen pairs and 700 unseen pairs. In generalized ZSCL, the validation set has 600 pairs with 300 pairs seen in the training set and 300 pairs unseen during training and the test set has 800 pairs with 400 pairs seen and remaining 400 pairs unseen in the training set.
- UT-Zappos (Yu and Grauman, 2017) contains images of 12 shoe types as object labels and 16 material types as attribute labels. In conventional ZSCL, the dataset is split into disjoint seen set with 83 pairs and unseen set with 33 pairs. In generalized ZSCL, the 36 pairs in the test set consists 18 seen and 18 unseen pairs. 15 seen pairs and 15 unseen pairs composes the validation set.

Implementation Details. We develop our model based on *PyTorch*. For all experiments, we adopt ResNet-18 pre-trained on ImageNet as the backbone to extract visual features. For *attr-obj* pairs, we encode attributes and objects with 300-dim *GloVe* and fix it during the training process. We randomly sample 50 negative pairs to construct episodes. We use Adam with 10^{-3} as the initial learning rate and multiply the learning rate by 0.5 every 5 epoch and train the network for total 25 epochs. We report the accuracy at the last epoch for conventional ZSCL. For generalized ZSCL, the accuracy is reported based on the validation set. Moreover, the batch size is set to 64, λ in Eq. 3 is set to 9, q in Eq. 8 is set to 0.5 and the threshold in Eq. 7 is set to 10.

Baselines. We compare EpiCA with the following SOTA methods: 1) Analog (Chen and Grauman, 2014) trains a linear SVM classifier for the seen pairs and utilizes Bayesian Probabilistic Tensor Factorization to infer the unseen classifier weights. 2) Redwine (Misra et al., 2017) leverages the compatibility between visual features v and concepts semantic representation to do the recognition. 3) AttOperator (Nagarajan and Grauman, 2018) models composition by treating attributes as matrix op-

	Conventional ZSCL		Generalized ZSCL	
	MIT-States	Zappos	MIT-States	Zappos
# Attr.	115	16	115	16
# Obj.	245	12	245	12
# Train Pair	1262	83	1262	83
# Train Img.	34562	24898	30338	22998
# Test Pair	700	33	800	36
# Test Img.	19191	4228	12995	2914
# Val. Pair			600	30
# Val. Img.			10420	3214

Table 1: Conventional and Generalized Data Split for MIT-States and Zappos Datasets.

erators to modify object state to score the compatibility. 4) GenModel (Nan et al., 2019) adds reconstruction loss to boost the metric-learning performance. 5) TAFE-Net (Wang et al., 2019) extracts visual features based on the pair semantic representation and utilizes a shared classifier to recognize novel concepts. 6) SymNet (Li et al., 2020) builds a transformation framework and adds group theory constraints to its latent space to recognize novel concepts. We report the results according to the papers and the released official code ³ ⁴ of the aforementioned baselines.

Methods	MIT-States(%)	UT-Zappos(%)
Random	0.14	3.0
ANALOG	1.4	18.3
REDWINE	12.5	40.3
ATTOPERATOR	14.2	46.2
GenModel	17.8	48.3
TAFE-Net	16.4	33.2
SymNet	19.9	52.1
EpiCA(Inductive)	15.68	52.56
EpiCA(Transductive)	18.13	55.48

Table 2: Results of Conventional ZSCL setting

4.1 Conventional ZSCL Setting

Quantitive Results. Top-1 accuracy metric is reported in this setting to compare different methods. The top-1 accuracy of the unseen *attr-obj* pairs for conventional ZSCL is presented in Tab. 2. EpiCA outperforms all baselines on Zappos benchmark and exceeds the state-of-the-art by 3.3%. It achieves comparable performance on MITStates benchmark. We will empirically analyze the model’s behavior in later sections.

4.2 Generalized ZSCL Setting

In this setting, following the related work (Purushwalkam et al., 2019), we measure the performance

³<https://github.com/Tushar-N/attributes-as-operators>

⁴<https://github.com/ucbdrive/tafe-net.git>

Model Top k →	Mit-States						UT-Zappos					
	Val AUC			Test AUC			Val AUC			Test AUC		
	1	2	3	1	2	3	1	2	3	1	2	3
AttOperator	2.5	6.2	10.1	1.6	4.7	7.6	21.5	44.2	61.6	25.9	51.3	67.6
RedWine	2.9	7.3	11.8	2.4	5.7	9.3	30.4	52.2	63.5	27.1	54.6	68.8
LabelEmbed+	3.0	7.6	12.2	2.0	5.6	9.4	26.4	49.0	66.1	25.7	52.1	67.8
TMN	3.5	8.1	12.4	2.9	7.1	11.5	36.8	57.1	69.2	29.3	55.3	69.8
SymNet	4.3	9.8	14.8	3.0	7.6	12.3	\	\	\	\	\	\
Inductive EpiCA	7.73	12.19	22.93	6.55	13.07	20.01	25.13	50.19	61.97	25.59	50.06	63.08
Transductive EpiCA	9.01	17.63	24.01	7.18	14.02	21.31	53.18	68.71	77.89	35.04	54.83	70.02

Table 3: AUC in percentage (multiplied by 100) on MIT-States and UT-Zappos. Our *EpiCA* model outperforms the previous methods by a large margin on MIT-States based on most of the metrics on UT-Zappos.

with AUC metric. AUC introduces the concept of calibration bias which is a scalar value added to the predicting scores of unseen pairs. By changing the values of the calibration bias, we can draw an accuracy curve for seen/unseen sets. The area below the curve is the AUC metric as a measurement for the generalized ZSCL system.

Quantitative results. Tab. 3 provides comparisons between our EpiCA model and the previous methods on both the validation and testing sets. As Tab. 3 shows, the EpiCA model outperforms the previous methods by a large margin. On the challenging MIT-States dataset which has about 2000 attribute-object pairs, all the baseline methods have a relatively low AUC score while our model is able to double the performance of the previous methods, indicating its effectiveness.

4.3 Ablation Study

We conduct ablation study on EpiCA and compare its performance in different settings.

Importance of Transductive Learning. The experimental results in Tab. 2 and Tab. 3 show the importance of transductive learning. There are about 2% and 3% performance gains for MIT-States and UT-Zappos in conventional ZSCL. A significant improvement is observed for both datasets in generalized ZSCL. This is within our expectation because 1) our inductive model has accumulated knowledge about the elements of the concept and has the ability to pick confident test examples. 2) after training the model with the confident pseudo-labeled test data, it acquires the knowledge about unseen concepts.

Importance of Cross-Attention (CA) Encoder. To analyze the effect of CA encoder, we remove CA (w/o CA) and use unimodal representations for both concepts and images. From Tab. 4, it can be seen that EpiCA does depend on multi-modal

information to do concept recognition and the results also verifies the rationale to fuse multi-modal information by cross-attention mechanism.

Importance of Gated Pooling (GP) Layer. We replace GP layer by average pooling (w/o GP). Tab. 4 shows the effectiveness of GP in filtering out noisy information. Instead of treating each element equally, GP help selectively suppress and highlight salient elements within each modality.

Importance of Episode Training. We also conduct experiments by removing both CA and GP (w/o GP and CA). In this setting, we concatenate unimodal representation of images and concepts and use 2-layer MLP to calculate the relevance score. Although simple, it still achieves satisfactory results, showing episode training is vital for our EpiCA model.

EpiCA variants	MIT-States(%)	UT-Zappos(%)
Full EpiCA	15.79	52.56
- w/o cross attention (CA)	12.05	42.77
- w/o gated pooling (GP)	13.46	50.47
- w/o GP and CA	14.13	48.76

Table 4: Ablation study of EpiCA components. The episode training and cross-attention encoder are import to our model. Adding gated pooling layer further boosts the accuracy.

4.4 Qualitative Analysis.

Fig. 3 shows some examples and their predicted labels by EpiCA. Although it gives the correct predictions for the two examples in the first row, EpiCA still struggles in distinguishing the similar, even opposite attributes, like *New* and *Old*. For example, the second highest prediction for the image with true label *new truck* is *old car*. The predicted object is reasonable, but the predicted attribute is opposite. Meanwhile, for the incorrect predictions, the predicted labels are meaningful and remain relevant

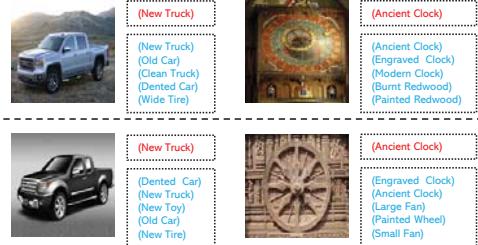


Figure 3: Predicting examples of EpiCA from MIT-States dataset. True label and predicted labels are in red and blue text respectively.

to the image. For example, *Engraved Clock* may be a better label than *Ancient Clock* for the bottom image. These examples show that EpiCA learns the relevance between images and concepts. But the evaluation of the models is hard and in some cases additional information and bias is needed to predict the exact labels occurring in the dataset.

5 Conclusion

In this paper, we propose EpiCA which combines episode-based training and cross-attention mechanism to exploit the alignment between concepts and images to address ZSCL problems. It has led to competitive performance on two benchmark datasets. In generalized ZSCL setting, EpiCA achieves over $2\times$ performance gain compared to the SOTA on several evaluation metrics. However, ZSCL remains a challenging problem. Future work that explores cognitively motivated learning models and incorporates information about relations between objects as well as attributes will be interesting directions to pursue.

References

Chao-Yeh Chen and Kristen Grauman. 2014. Inferring analogous attributes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 200–207.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020a. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12655–12663.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020b. Image search with text feedback by visiolinguistic attention learning. pages 3001–3011.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28:577–585.

Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. 2019. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Phillip Isola, Joseph J Lim, and Edward H Adelson. 2015. Discovering states and transformations in image collections.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International Conference on Machine Learning*, pages 2873–2882.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. pages 201–216.

Kai Li, Martin Renqiang Min, and Yun Fu. 2019a. Rethinking zero-shot learning: A conditional visual classification perspective. *The IEEE International Conference on Computer Vision (ICCV)*.

Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. 2019b. Learning to self-train for semi-supervised few-shot classification. *Advances in Neural Information Processing Systems*, pages 10276–10286.

Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. 2020. Symmetry and group in attribute-object compositions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. pages 13–23.

Ishan Misra, Abhinav Gupta, and Martial Hebert. 2017. From red wine to red tomato: Composition with context. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801.

Tushar Nagarajan and Kristen Grauman. 2018. Attributes as operators. *ECCV*.

Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. 2019. Recognizing unseen attribute-object pair with generative model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8811–8818.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc’Aurelio Ranzato. 2019. Task-driven modular networks for zero-shot compositional learning. *arXiv preprint arXiv:1905.05908*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, pages 4077–4087.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, pages 3630–3638.

Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. 2019. Tafe-net: Task-aware feature embeddings for low shot learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840.

Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. 2019. Adversarial fine-grained composition learning for unseen attribute-object recognition. pages 3741–3749.

Aron Yu and Kristen Grauman. 2017. Semantic jitter: Dense supervision for visual comparisons via synthetic images. *Proceedings of the IEEE International Conference on Computer Vision*, pages 5570–5579.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, pages 8778–8788.