DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

Day27

Q1. Learning to Caption Images with Two-Stream Attention and Sentence Auto-Encoder

Answer:

Understanding the world around us via visual representations, and communicating this extracted visual information via language is one of the fundamental skills of human intelligence. The goal of recreating a similar level of intellectual ability in artificial intelligence(AI) has motivated researchers from computer vision and natural language communities to introduce the problem of automatic image captioning. Image captioning, which is to describe the content of an image in the natural language, has been an active area of research and widely applied to video and image understanding in multiple domains. The ideal model for this challenging task must have two characteristics: understanding of an image content well and generating descriptive sentences which is coherent with the image content. Many image captioning methods propose various encoder-decoder models to satisfy these needs where encoder extracts the embedding from an image, and decoder generates the text based on the embedding. These two parts are typically built with a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN), respectively.

Fig.: This Image captioning decoder with two-stream attention and the Auxillary decoder “finds” and “localizes” relevant words better than general caption-attention baselines

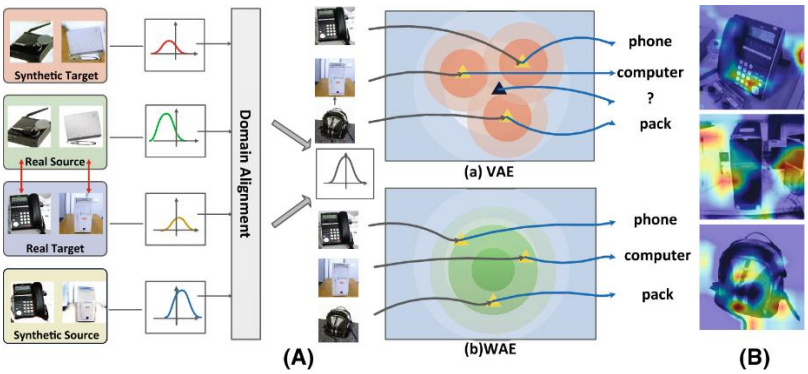
One of the challenging question in encoder-decoder architectures is how to design interface that controls the information flow between a CNN and RNN. While early work employs static representation for interface such that the CNN compresses an entire image into a fixed vector, and an RNN decodes representation into natural language sentences, this strategy is shown to perform poorly when target sentence is prolonged, and the image is reasonably cluttered. Inspired from, Xu *et al*. propose the powerful dynamic interface, namely attention mechanism, that identifies relevant parts of a image embedding to estimate the next word. RNN model then predicts the word based on the context vector associated with the related image regions and the previously generated words. The attentional interface is shown to obtain significant performance improvements over static one, and

since then, it has become the key component in all state-of-the-art(SOTA) image captioning models. Although this interface is substantially effective and flexible, it comes with critical shortcoming.

Nevertheless, visual representations that are learned by Convolutional Neural Network(CNNs) have been rapidly improving the state-of-the-art(SOTA) recognition performance in various image recognition tasks in past few years. They can still be inaccurate when applied to noisy images and perform poorly to describe their visual contents. Such noisy representations can lead to incorrect association between words and images regions and potentially drive the language model to poor textual descriptions. To address these shortcomings, we propose two improvements that can be used in standard encoder-decoder based image captioning framework.

First, we propose the novel and powerful attention mechanism that can more accurately attend to relevant image regions and better cope with ambiguities between words and image regions. It automatically identifies *latent categories* that capture high-level semantic concepts based on visual and textual cues, as illustrated in the second fig. The two-stream attention is modeled as a neural network where each stream specializes in orthogonal tasks: the first one soft-labels each image region with the latent categories, and the second one finds the most relevant area for each group. Then their predictions are combined to obtain a context vector that is passed to a decoder.

Second, inspired by sequence-to-sequence (seq2seq) machine translation methods, we introduce a new regularization technique that forces the image encoder coupled with the attention block to generate a more robust context vector for the following RNN model. In particular, we design and train an additional seq2seq sentence auto-encoder model (“SAE”) that first reads in a whole sentence as input, generates the fixed dimensional vector, then the vector is further used to reconstruct input sentence. SAE is trained to learn structure of the input (sentence) space in an offline manner, Once it is trained, we freeze its parameters and incorporate *only* its decoder part (SAE-Dec) to our captioning model (“IC”) as the auxiliary decoder branch. SAE-Dec is employed along with the original image captioning decoder (“IC-Dec”) to output target sentences during training and removed in test time. We show that the proposed SAE-Dec regularizer improves the captioning performance for IC-Dec and does not bring any additional computation load in test time.



Q2.Explain PQ-NET.

Answer:

PQ-NET: A Generative Part Seq2Seq Network for 3D Shapes.Learning generative models of 3D shapes is a crucial problem in both computer vision and computer graphics. While graphics are mainly concerned with 3D shape modeling, in inverse graphics, a significant line of work in computer vision, one aims to infer, often from a single image, a disentangled representation wrt 3D shape and scene structures. Lately, there has been a steady stream of works on developing deep neural networks for 3D shape generation using different shape representations, e.g., voxel grids, point clouds, meshes, and, most recently, implicit functions. However, most of these works produce *unstructured* 3D shapes, even though object perception is generally believed to be a process of a *structural understanding*, i.e., to infer shape parts, their compositions, and inter-part relations.

In this paper, we introduce a deep neural network that represents and generates 3D shapes via *sequential part assembly*, as shown in both Fig. In a way, we regard assembly sequence as a "sentence," which organizes and describes the parts constituting the 3D shape. Our approach is inspired, in part, by the resemblance between speech and shape perception, as suggested by the seminal work of Biederman on recognition-by-components (RBC). Another related observation is that the phase structure rules for language parsing, first introduced by Noam Chomsky, take on the view that sentence is both a linear string of words and a hierarchical structure with phrases nested in phrases. In the context of shape structure presentations, our network adheres to linear part orders, while other works have opted for *hierarchical* part organizations.

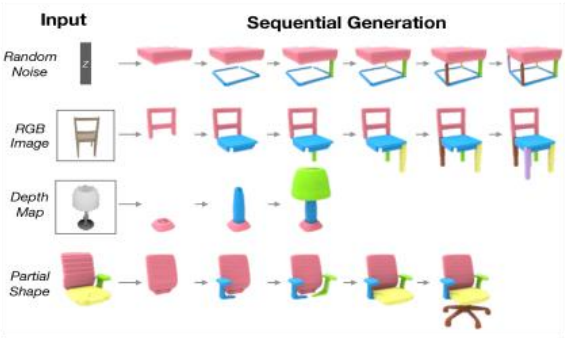


Fig 1: Our network, PQ-NET, learns 3D shape representation as a *sequential part assembly*. It can be adapted to generative tasks such as random 3D shape generation, single-view 3D reconstruction (from RGB or depth images), and shape completion.

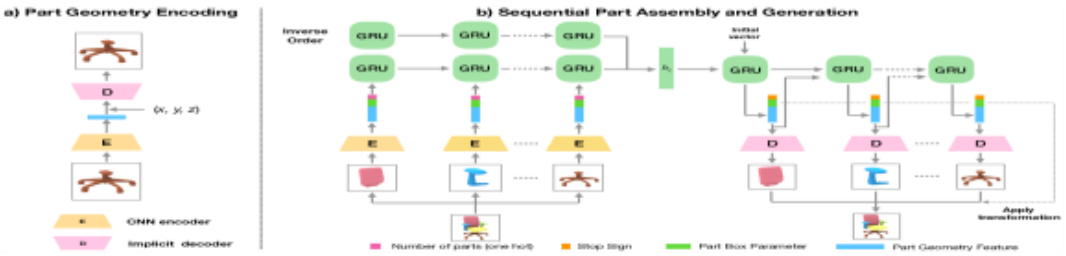


Fig2: The architecture of PQ-NET: our part Seq2Seq generative network for 3D shapes.

The input to our network is a 3D shape segmented into parts, where each part is first encoded into a feature representation using a part autoencoder; see Fig2(a). The core component of our network is a *Seq2Seq* autoencoder, which encodes a sequence of part features into the latent vector of fixed size, and the decoder reconstructs the 3D shape, one part at the time, resulting in sequential assembly; see Fig 2(b). With its part-wise Seq2Seq architecture, our network is coined *PQ-NET*. The latent space formed by Seq2Seq encoder enables us to adapt the decoder to perform several generative tasks,

including shape autoencoding, interpolation, new shape generation, and single-view 3D reconstruction, where all generated shapes are composed of meaningful parts.

As training data, we take the segmented 3D shapes from PartNet, which was built on ShapeNet. It is important to note that we do not enforce any particular part order or consistency across input shapes. The shape parts are always specified in the file following some linear order in the dataset; our network takes whatever part order that is in a shapefile. We train the part and Seq2Seq autoencoders of PQ-NET separately, either per shape category or across all shape categories, of PartNet.

Our part autoencoder adapts IM-NET to encode shape parts, rather than whole shapes, with the decoder producing an implicit field. The part Seq2Seq autoencoder follows a similar architecture as the original Seq2Seq network developed for machine translation. Specifically, the encoder is a bidirectional stacked recurrent neural network (RNN) that inputs two sequences of part features, in opposite orders, and outputs a latent vector. The decoder is also a stacked RNN, which decodes the latent vector representing the whole shape into a sequential part assembly.

PQ-NET is the first *fully generative* network that learns a 3D shape representation in the form of sequential part assembly. The only prior part sequence model was 3D-PRNN, which generates part boxes, not their geometry — our network jointly encodes and decodes part structure and geometry. PQ-NET can be easily adapted to various generative tasks, including shape autoencoding, novel shape generation, structured single-view 3D reconstruction from both RGB and depth images, and shape completion. Through extensive experiments, we demonstrate that performance and output quality of our network is comparable or superior to state-of-the-art generative models, including 3D-PRNN, IM-NET, and StructureNet.

Q3. What is EDIT?

Answer:

EDIT: Exemplar-Domain Aware Image-to-Image Translation

A scene can be expressed in various manners using sketches, semantic maps, photographs, and painting,artworks, to name just a few. The way that one portrays the scene and expresses his/her vision is the so-called style, which can reflect the characteristic of either a class/domain or a specific case. Image-to-image translation (I2IT) refers to the process of converting an image I of a particular style to another of the target style St with the content preserved. Formally, seeking the desired translator T can be written in the following form:

$$\min \quad \mathcal{C}(I_t, I) + \mathcal{S}(I_t, S_t) \quad \text{with} \quad I_t := \mathcal{T}(I, S_t), \tag{1}$$

where $\mathcal{C}(I_t, I)$ is to measure the content difference between the translated I_t and the original I, while $\mathcal{S}(I_t, S_t)$ is to enforce the style of I_t following that indicated by S_t.

Figure 1: Several results by the proposed EDIT. Our EDIT can take arbitrary exemplars as reference for translating images across multiple domains, including photo-painting, shoe-edge, and semantic map-facade in *one* model.

With the emergence of deep techniques, a variety of I2IT strategies have been proposed with excellent progress made over the last decade. In what follows, we briefly review contemporary works along two main technical lines, *i.e.*, one-to-one translation and many-to-many translation.

One-to-one Translation. Methods in this category aim at mapping images from a source domain to a target domain. Benefiting from the generative adversarial networks (GANs), the style of translated results satisfies the distribution of the target domain Y, achieved by S(It,St):=D(It,Y), where D(It, Y) represents a discriminator to distinguish if It is real with respect to Y. An early attempt by Isola *et al.* uses conditional GANs to learn mappings between two domains. The paired data supervise the content preservation, *i.e.*, C(It, I):=C(It, Igtt) with Igtt, the ground-truth target. However, in real-world situations, acquiring such paired datasets, if not impossible, is impractical. To alleviate the pressure from data, inspired by the concept of cycle consistency, cycleGAN in an unsupervised fashion was proposed, which adopts C(It,I):=C(FY⟶X(FX⟶Y(I)),I) with FX⟶Y the generator from domain X to Y and FY⟶X the reverse one. Afterward, StarGAN further extends the translation between two domains that cross multiple areas in a single model. Though the effectiveness of the mentioned methods has been witnessed by a broad spectrum of specific applications such as photo-

caricature, making up-makeup removal, and face manipulation, their main drawback comes from the nature of deterministic (uncontrollable) one-to-one mapping.

Many-to-many Translation. The goal of approaches in this class is to transfer the style controlled by an exemplar image to a source image with content maintained. Arguably, the most representative work goes to, which uses the pre-trained VGG16 network to extract the content and style features, then transfer style information by minimizing the distance between Gram matrices constructed from the generated image and the exemplar E, say S(It,St):=S(Gram(It),Gram(E)). Since then, numerous applications on the 3D scene, face swap, portrait stylization and font design have been done. Furthermore, several schemes have also been developed towards relieving limitations in terms of speed and flexibility. For example, to tackle the requirement of training for every new exemplar (style), Shen *et al.* built a meta-network, which takes in the style image and produces a corresponding image transformation network directly. Risser *et al.* proposed the histogram loss to improve the training instability. Huang and Belongie designed a more suitable normalization manner, *i.e.*, AdaIN, for style transfer. Li *et al.* replaced the Gram matrices with an alternative distribution alignment manner from the perspective of domain adaption. Johnson *et al.* trained the network with a specific style image and multiple content images while keeping the parameters at the inference stage. Chen *et al.* introduced a style-bank layer containing several filter-banks, each of which represents a specific style. Gu *et al.* proposed a style loss to make parameterized, and non-parameterized methods complement each other. Huang *et al.*designed a new temporal loss to ensure the style consistency between frames of a video. Also, to mitigate the deterministic nature of one-to-one translation, several works, for instance, advocated to separately take care of content c(I) and style s(I) subject to $I \simeq c(I) \circ s(I)$ with $\circ$ the combined operation. They manage to control the translated results by combining the content of an image with the style of the target, *i.e.*, $c(I) \circ s(E)$. Besides, one domain pair requires one independent model, their performance, as observed from comparisons, is inferior to our method in visual quality, diversity, and style preservation. Please see the above Fig. , For example produced by our approach that handles multiple domains and arbitrary exemplars in a unified model.

Q4. What is Doctor2Vec?

Answer:

Doctor2Vec: Dynamic Doctor Representation Learning for
Clinical Trial Recruitment

The rapid growth of electronic health record (EHR) data and other health data enables the training of complex deep learning models to learn patient representations for disease diagnosis, risk prediction,

patient subtyping, and medication recommendation. However, almost all current works focus on modeling patients. Deep neural networks for doctor representation learning are lacking.

Doctors play pivotal roles in connecting patients and treatments, including recruiting patients into clinical trials for drug development and treating and caring for their patients. Thus an effective doctor representation will better support a broader range of health analytic tasks. For example, identifying the right doctors to conduct the trial *site selection* to improve the chance of completion of the trials [hurtado2017improving] and doctor recommendation for patients.

In this work, we focus on studying the *clinical trial recruitment* problem using doctor representation learning. Current standard practice calculates the median enrollment rate. Enrollment rate of a doctor is the number of patients enrolled by a doctor to the trial. For the therapeutic area as the predicted enrollment success rate for whole participating doctors, which is often incorrect. Also, some develop a multi-step manual matching process for site selection, which is labor-intensive. Recently, deep neural networks were applied on site selection tasks via static medical concept embedding using only frequent medical codes and simple term matching to trials. Despite the success, two challenges remain open.

1. Existing works do not capture the time-evolving patterns of doctors experience and expertise encoded in EHR data of patients that the doctor have seen;
2. Current jobs learn a static doctor representation. However, in practice, given a trial for a particular disease, the doctor's experience of relevant diseases are more important. Hence the doctor representation should change based on the corresponding trial representation.

To fill the gap, we propose Doctor2Vec, which simultaneously learns i) doctor representations from longitudinal patient EHR data and ii) trial embedding from the multimodal trial description. In particular, Doctor2Vec leverages a dynamic memory network where the observations of patients seen by the doctor are stored as memory while trial embedding serves as queries for retrieving from the mind. Doctor2Vec has the following contributions.

1. **Patient embedding as a memory for dynamic doctors representation learning**. We represent doctors' evolving experience based on the representations from the doctors' patients. The patient representations are stored as a memory for dynamic doctor representation extraction.
2. **Trial embedding as a query for improved doctors selection**. We learn hierarchical clinical trial embedding where the unstructured trial descriptions were embedded using BERT [devlin2018bert]. The trial embedding serves as queries of the memory network and will attend over patient representation and dynamically assign weights based on the relevance of doctor experience and trial representation to obtain the final context vector for an optimized doctor representation for a specific test.

We evaluated Doctor2Vec using large scale real-world EHR and trial data for predicting trial enrollment rates of doctors. Doctor2Vec demonstrated improved performance in the site selection task over the best baselines by up to 8.7% in PR-AUC. We also showed that the Doctor2Vec embedding could be transferred to benefit data insufficiency settings, including trial recruitment in less populated/newly explored countries or for rare diseases. Experimental results show for the country transfer, Doctor2Vec achieved 13.7% relative improvement in PR-AUC over the best baseline. While for embedding transfer to unique disease trials, Doctor2Vec made 8.1%relative improvements in PR-AUC over the best benchmark.

Q5. Explain PAG-Net.

Answer:

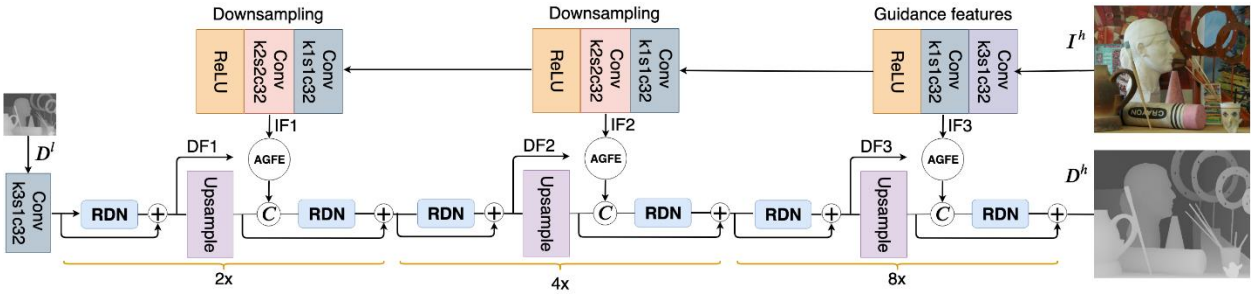
PAG: Progressive Attention Guided Depth Super-resolution Network

A geometric description of a scene, the high-quality depth map is useful in many computer vision applications, such as 3D reconstruction, virtual reality, scene understanding, intelligent driving, and robot navigation. Literature mainly contains two classes of techniques for depth information acquisition, which are passive methods and active sensors. Firstly, passive methods infer depth maps from the most widely used dense stereo matching algorithms, but they are time-consuming. Despite the advances in technology, the depth information from passive methods is still inaccurate in occluded and low-texture regions. The acquisition of high-quality depth maps is more challenging to obtain than RGB images.

Depth acquisition from active sensors has become increasingly popular in our daily life and ubiquitous to many consumer applications, due to their simplicity, portability, and inexpensive. Unlike passive methods, the depth of a scene can be acquired in real-time, and they are more robust in low-textured regions by low-cost sensors such as Time-of-Flight camera and Microsoft Kinect. Current sensing techniques measure depth information of a scene by using echoed light rays from the stage. Time-of-Flight sensor (ToF) is one of the mainstream types which computes depth at each pixel between camera and subject, by measuring the round trip time. Although depth-sensing technology has attracted much attention, it still suffers from several quality degradations.

Depth information captured by ToF sensors suffers from low-spatial resolutions (e.g., 176×144, 200×200 or 512×424) and noise when compared with the corresponding color images. Due to the offset between projector and sensor, depth maps captured by Microsoft Kinect sensors contain structural missing along discontinuities and random missing at homogeneous regions. These issues restrict the use of depth maps in the development of depth-dependent applications. High-quality depth is significant in many computer vision applications. Therefore, there is a need for restoration of

depth maps before using in applications. In this work, we consider the problem of depth map super-resolution from a given low-resolution depth map and its corresponding high-resolution color image.



Existing depth super-resolution (DSR) methods can be roughly categorized into three groups: filter design-based, optimization-based, and learning-based algorithms. Many of the existing techniques assumed that a corresponding high-resolution color image helps to improve the quality of depth maps and used aligned RGB image as guidance for depth SR. However, significant artifacts including texture copying and edge blurring, may occur when the assumption violated. The color texture will be transferred to the super-resolved depth maps if the smooth surface contains rich textures in the corresponding color image. Secondly, depth and color edges might not align in all the cases. Subsequently, it leads to ambiguity. Hence, there is a need for optimal guidance for the high-resolution depth map.

Although there have been many algorithms proposed in the literature for the depth super-resolution (DSR), most of them still suffer from edge-blurring and texture copying artifacts. In this paper, we offer a novel method for attention guided depth map super-resolution. It is based on dense residual networks and involves a unique attention mechanism. The attention used here to suppress the texture copying problem arises due to improper guidance by RGB images and transfer only the salient features from the guidance stream. The attention module mainly involves providing spatial attention to the guidance image based on the depth features. The entire architecture for the example of super-resolution by the factor of 8 is shown in Above Fig.

Q6. An End-to-End Audio Classification System based on Raw Waveforms and Mix-Training Strategy

Answer:

Sound is the indispensable medium for information transmission of surrounding environment. When some sounds happen, such as baby crying, glass breaking, and so on, we usually expect that we can "hear" sounds immediately, even if we are not around. In this case, an audio classification that aims to

predict whether an acoustic event appears has gained significant attention in recent years. It has many practical applications in remote surveillance, home automation, and public security.

In real life, an audio clip usually contains multiple overlapping sounds, and types of sounds are various, ranging from natural soundscapes to human activities. It is challenging to predict a presence or absence of audio events in an audio clip. Audio Set is the common large-scale dataset in this task, which contains about two million multi-label audio clips covering 527 classes. Recently, some methods have been proposed to learn audio tags on this dataset. Among them, a multi-level attention model achieved state-of-the-art(SOTA) performance, which outperforms Google's baseline. However, the shortcoming of these models is that the input signal is the published bottleneck feature, which causes information loss. Considering that the actual length of sound events is different and the handcrafted features may throw away relevant information at a short time scale, raw waveforms containing more valuable information is a better choice for multi-label classification. In the audio tagging task of DCASE 2017, 2018 challenge, some works [5, 6] combined handcrafted features with raw waveforms as input signal on a small dataset consisting of 17 or 41 classes. To our knowledge, none of the works proposes an end-to-end network taking raw waveforms as input in the Audio Set classification task.

In this paper, we propose a classification system based on two variants of ResNet, which directly extracts features from raw waveforms. Firstly, we use one-dimension (1D) ResNet for feature extraction. Then, two-dimension (2D) ResNet with multi-level prediction and attention structure is used for classification. For obtaining better classification performance further, a mix-training strategy is implemented in our system. In this training process, the network is trained with mixed data, which extends training distribution and then transferred to the target domain using raw data.

In this work, the main contributions are as follows:

1. The novel end-to-end Audio Set classification system is proposed. To best of our knowledge, it is first time to take raw waveforms as input on Audio Set and combine 1D ResNet with 2D ResNet for feature extraction at different time scales.
2. A mix-training strategy is introduced to improve the utilization of limited training data effectively. Experiments show that it is robust in multi-label audio classification compared to the existing data augmentation methods.

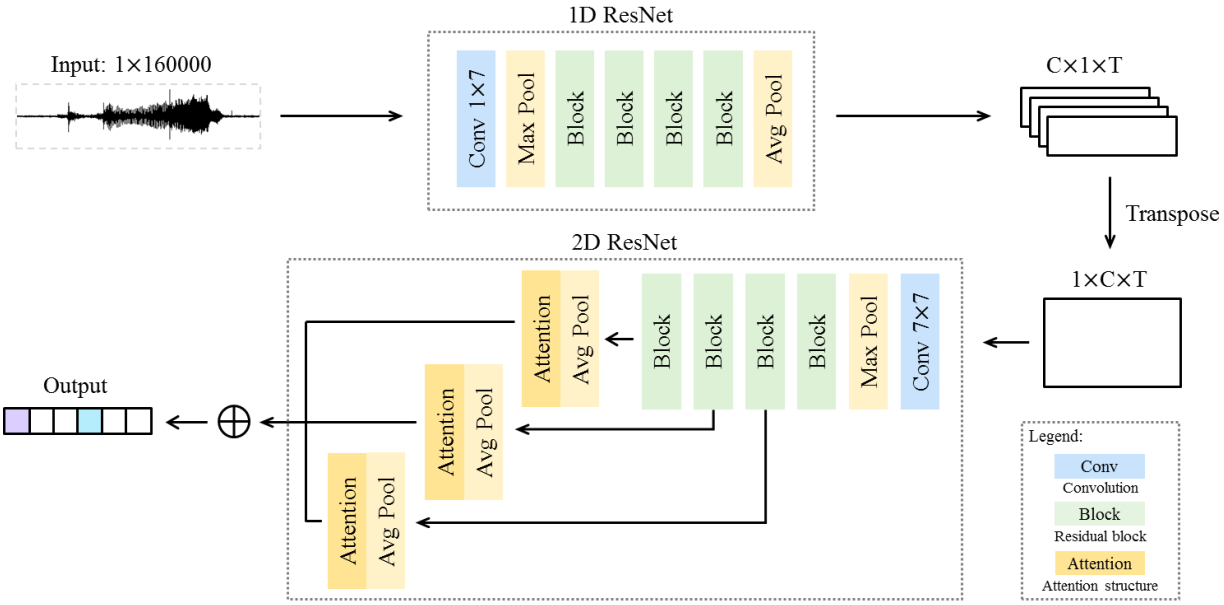


Figure 1: Architecture of the end-to-end audio classification network. The raw waveform (1D vector) is the input signal. First, the 1D ResNet is applied to extract audio features. Then, the elements are transposed from C×1×Tto 1×C×T. Finally, a 2D ResNet with a multi-level prediction structure performs audio classification. The output of network has multiple labels and is the mean of multi-level prediction results. The Block is composed of n bottleneck blocks, where n is related to a number of layers in ResNet.

Q7. What is Cnak? : Cluster Number Assisted *K-means*

Answer:

Cnak stands for Cluster Number Assisted K-means

In cluster analysis, it is required to group the set of data points in a multi-dimensional space so that data points in same group are more similar to each other than to those in other groups. These groups are called clusters. Various distance functions may be used to compute degree of dissimilarity or similarity among these data points. Typically Euclidean distance function is widely used in clustering. This unsupervised technique aims to increase homogeneity in the group and heterogeneity between groups. Several clustering methods with different characteristics have been proposed for different purposes. Some well-known methods include partition-based clustering, hierarchical clustering [Hierarchy1963], spectral clustering [onspectral2001], density-based clustering [DBSCAN]. However, they require the knowledge of cluster number for a given dataset a
priori [Lloyd57; onspectral2001; DBSCAN; DBCLASD; DENCLUE].

Nevertheless, estimation of the number of clusters is difficult problem as the underlying data distribution is unknown. Readers can find several existing techniques for determining cluster number in [survey_cluster_number2017; R3_Chiang2010]. We have followed the terminology used in R3_Chiang2010 for categorizing different methods for the prediction of cluster numbers. In this work, we choose to focus only on three approaches: 1) variance-based approach, 2) Structural approach, and 3) the Resampling approach. Variance-based plans are based on measuring compactness within a cluster. Structural approaches include between-cluster separation as well as within-cluster variance. We have chosen these approaches as they are either more suitable for handling big data, or appear in a comparative study by several researchers. Some well-known approaches are Calinski-Harabaz [CH], Silhouette Coefficient [sil], Davies-Bouldin [DB], Jump [jump], Gap statistic [gap], etc. These approaches are not appropriate for handling big data, as they are computationally intensive and require ample storage space. It requires a scalable solution [kluster2018; ISI_LL_LML2018] for identifying the number of clusters. Resampling-based approaches can be considered in such scenario. Recently, the concept of stability in clustering has become popular. A few methods [instability2012; CV_A] utilize the idea of clustering robustness against the randomness in the choice of sampled datasets to explore clustering stability.

Q8. What is D3S?

Answer:

D3S – A Discriminative Single Shot Segmentation Tracker.Visual object tracking is one of the core computer vision problems. The most common formulation considers the task of reporting the target location in each frame of the video given a single training image. Currently, the dominant tracking paradigm, performing best in evaluations [kristan_vot2017, kristan_vot2018], is correlation bounding box tracking where the target represented by a multi-channel rectangular template is localized by cross-correlation between the template and a search region.

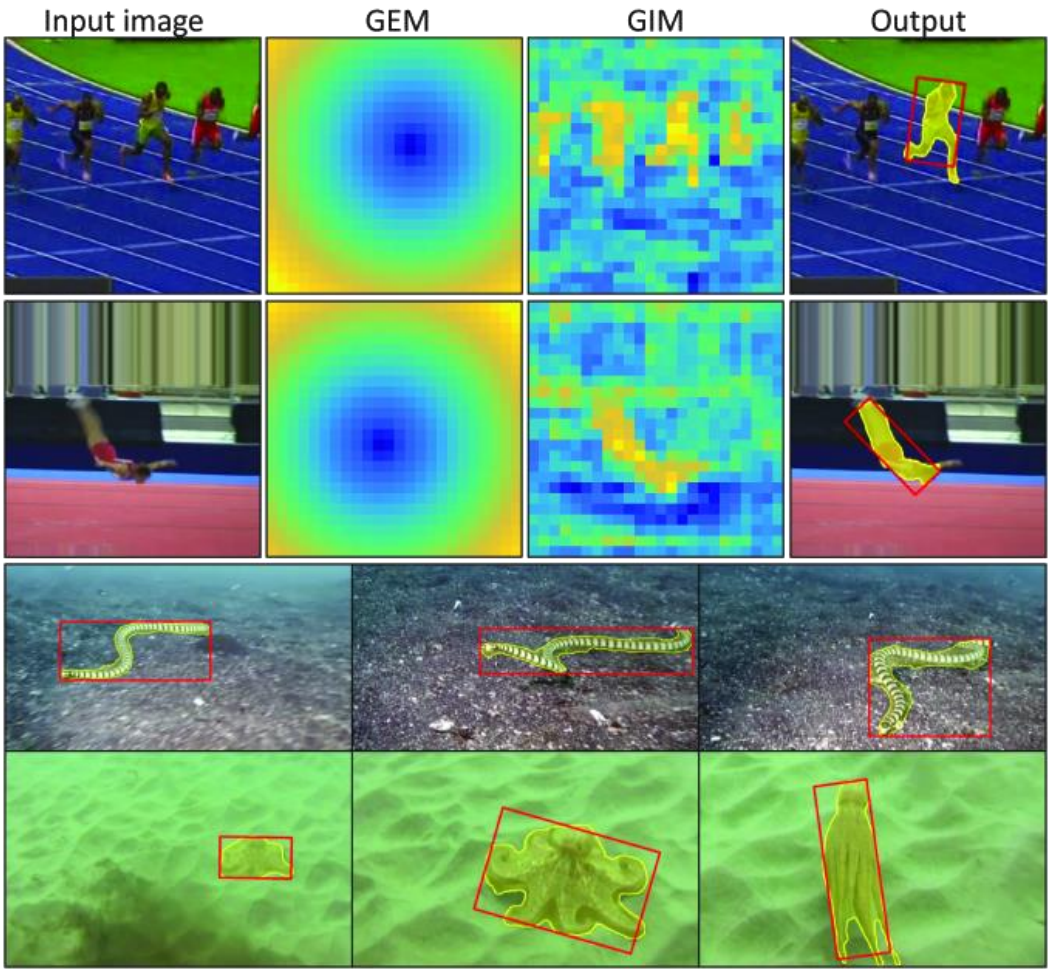


Figure 1: The D3S tracker represents the target by two models with complementary geometric properties, one invariant to a wide range of transformations, including non-rigid deformations (GIM - geometrically invariant model), the other assuming a rigid object with motion well approximated by a euclidean change (GEM - geometrically constrained Euclidean model). The D3S, exploiting the complementary strengths of GIM and GEM, provides both state-of-the-art localization and accurate segmentation, even in the presence of substantial deformation.

State-of-the-art template-based trackers apply an efficient brute-force search for target localization. Such a strategy is appropriate for low-dimensional transformations like translation and scale change but becomes inefficient for more general situations, e.g. such that induce an aspect ratio change and rotation. As a compromise, modern trackers combine approximate exhaustive search with sampling and bounding box refinement/regression networks for aspect ratio estimation. However, these approaches are restricted to axis-aligned rectangles.

Estimation of high-dimensional template-based transformation is unreliable when a bounding box is a sparse approximation of the target. This is common – consider, e.g. elongated, rotating, deformable

objects, or a person with spread out hands. In these cases, the most accurate and well-defined target location model is a binary per-pixel segmentation mask. If such output is required, tracking becomes the video object segmentation task recently popularized by DAVIS and YoutubeVOS challenges.

Unlike in tracking, video object segmentation challenges typically consider large target observed for less than 100 frames with low background distractor presence. Top video object segmentation approaches thus fare poorly in short-term tracking scenarios where the target covers a fraction of the image, substantially changes its appearance over a more extended period, and moves over a cluttered background. Best trackers apply visual model adaptation, but in the case of segmentation errors, it leads to irrecoverable tracking failure. Because of this, in the past, segmentation has played only an auxiliary role in template-based trackers, constrained DCF learning and tracking by 3D model

construction.

Recently, the SiamRPN tracker has been extended to produce high-quality segmentation masks in two stages – SiamRPN branches first localize the target bounding box, and then segmentation mask is computed only within this region by another branch. The two-stage processing misses the opportunity to treat localization and segmentation jointly to increase robustness. Another drawback is that a fixed template is used that cannot be discriminatively adapted to the changing scene.

We propose a new single-shot discriminative segmentation tracker, D3S, that addresses the limitations as mentioned above. Two discriminative visual models encode the target – one is adaptive and highly discriminative but geometrically constrained to a euclidean motion (GEM), while the other is invariant to a broad range of transformation (GIM, geometrically invariant model), see above Fig.

GIM sacrifices spatial relations to allow target localization under significant deformation. On the other hand, GEM predicts the only position but discriminatively adapts to the target and acts as a selector between possibly multiple target segmentations inferred by GIM. In contrast to related trackers [siammask_cvpr19, siamrpn_cvpr2019, atom_cvpr19], the primary output of D3S is the segmentation map computed in a single pass through the network, which is trained end-to-end for segmentation only.

Some applications and most tracking benchmarks require reporting the target location as a bounding box. As a secondary contribution, we propose an effective method for interpreting the segmentation mask as a rotated rectangle. This avoids an error-prone greedy search and naturally addresses changes in location, scale, aspect ratio, and rotation.

D3S outperforms all state-of-the-art trackers on most of the significant tracking benchmarks [kristan_vot2016, kristan_vot2018, got10k, muller_trackingnet] despite not being trained for bounding box tracking. In video object segmentation benchmarks [davis16, davis17], D3S

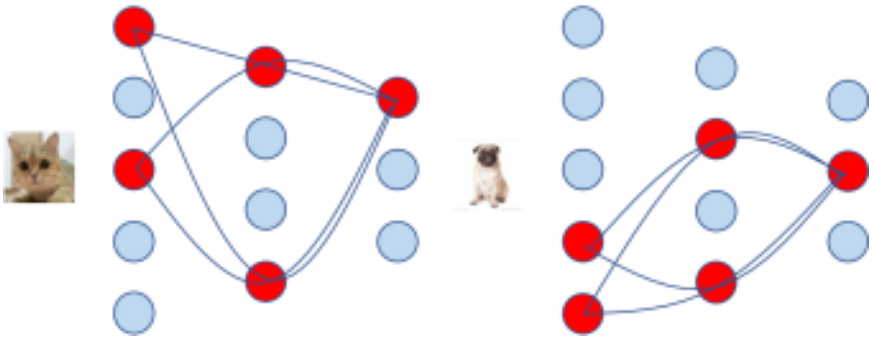
outperforms the leading segmentation tracker [siammask_cvpr19] and performs on par with top video object segmentation algorithms (often tuned to a specific domain), yet running orders of magnitude faster. Note that the D3S is not re-trained for different benchmarks – a single pre-trained version shows remarkable generalization ability, and versatility.PyTorch implementation will be made available.

Q9. What is DRNet?

Answer:

DRNet stands for Dissect and Reconstruct the Convolutional Neural Network via Interpretable Manners Convolutional neural networks (CNNs) have been broadly applied on various visual tasks due to its superior performance ([vgg], [resnet], [densenet]). But the huge computation burden prevents convolutional neural networks from running on mobile devices. Some works had been done to prune neural networks into smaller ones ([slimming], [pruning1], [pruning2]). Also, there are too many lightweight network structures ([mobilenet], [mobilenetv2], [shufflenet]) were proposed to adapt convolutional neural networks to computational limited mobile devices. However, these methods usually require running a whole pre-trained network, whatever the task is. i.e., the first task requires the discrimination power of cats and dogs, and the second task requires the discrimination power of apples and watermelons. If one has a CNN which was pre-trained on ImageNet, he must run the whole CNN on each task, which is usually time-consuming and computation wasted.

Our work focuses on an underlying problem, i.e., can we run only parts of a CNN? To achieve this goal, we need to find a method to dissect the whole network into pieces and reconstruct some of these pieces according to specific tasks. The reconstructed CNN should have a smaller computation cost and better performance. Meanwhile, the process of generating this substructure should be quick and easy. Therefore this technology can be applied on mobile devices and small robots such as cell-phones and uncrewed aerial vehicles. Using these technologies, these devices only need to store one complete CNN and some information about the substructure generating program. When specific tasks come, these devices can create a smaller substructure in an instant and run on it, rather than run the whole original CNN.



In this paper, we proposed a novel and interpretable algorithm to generate these smaller substructures. An interpretable way of CNN inspires our method. As shown in Figure 1: the original CNN has many channels, but not every channel is useful for the discrimination of every class. What we need to do is to find the channels relevant to every type and combine them for the specific task. This method looks similar to the previous work: structured network pruning ([slimming], [pruning3], [pruning4]). However, all of these pruning methods need fine-tuning, which is time-consuming and not allowed on mobile devices. And these pruning methods are usually lack of interpretability which is much needed by human-beings when using CNNs. Therefore, we do not mean to propose a pruning method and make CNN smaller, but to find the best channels for each class, and combine them for specific tasks. Our approach not only can be used on VGG and ResNet but also on some light structures such as MobileNetV2. Also, we make this process quick and interpretable.