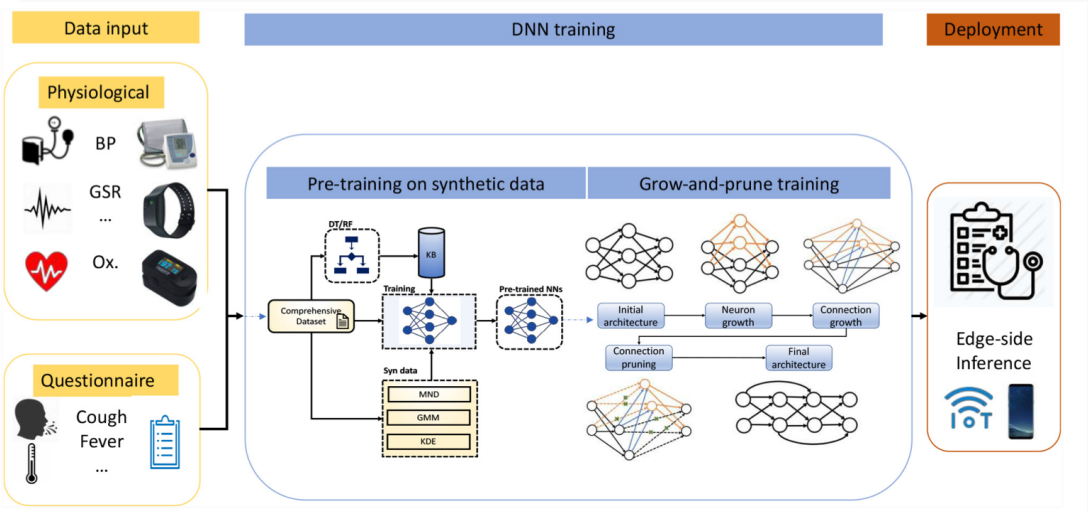
IEEE CONSUMER TECHNOLOGY SOCIETY (CTSOC)

ISSUE: APRIL 2022

CTSOC-NCT NEWS ON CONSUMER TECHNOLOGY



Schematic diagram of the CovidDeep framework (GSR: Galvanic skin response, Ox.: oxygen saturation, BP: blood pressure, DT/RF: decision tree/random forest, NN: neural network, KB: knowledge-base, MND: multi-variate Normal distribution, GMM: Gaussian mixture model, KDE: kernel density estimation).

IEEE CTSoc CONSUMER TECHNOLOGY SOCIETY

TABLE OF CONTENTS

EDITOR'S NOTE

On behalf of the Editorial Board of IEEE CTSoc News on Consumer Technology (NCT) and my co-editors, Yafei Hou and Luca Romeo, I am delighted to introduce the April 2022 issue of the News on Consumer Technology (NCT).

This issue starts with a cover story which presents a SARS-CoV-2/COVID-19 detection framework, published in IEEE Transactions on Consumer Electronics, as a demonstration of easy-to-use, accurate, and pervasive framework that can be easily deployed on edge devices (e.g., smartphones or smartwatches) as well as servers, with the possibility to be extended to larger deployment scenarios.

Next, an interview with Prof. Zhihui Lu from Fuduan University, China, presents his vision on the research for wireless service and cloud computing technologies. This issue ends with a featured article brought by Dr. Bei Liu and Dr. Jianlong Fu of Microsoft Research Asia, sharing us a new paradigm for multi-modality understanding

Besides, I would like to bring to your attention the online registration of ICCE-TW 2022, an annual flagship conference of IEEE CTSoc, is now open. You are encouraged to attend either virtually or in person!

Happy reading!

Wen-Huang Cheng
Editor-in-Chief

ARTICLE TITLE

CovidDeep: SARS-CoV-2/COVID-19 Test Based on Wearable Medical Sensors and Efficient Neural Networks

AUTHOR(S)

Shayan Hassantabar , Novati Stefano, Vishweshwar Ghanakota, Alessandra Ferrari, Gregory N. Nicola , Raffaele Bruno, Ignazio R. Marino, Kenza Hamidouche , and Niraj K. Jha , Fellow, IEEE

JOURNAL TITLE

IEEE Transactions on Consumer Electronics

JOURNAL VOLUME AND ISSUE

Volume: 67, Issue: 4

DATE OF THE ARTICLE

November 2021

PAGE NUMBERS FOR THE ARTICLE

244-256

In response to COVID-19, governments around the world issued social distancing and self-isolation orders. This led to a significant increase in unemployment across diverse economic sectors. As a result, COVID-19 triggered an economic recession in a large number of countries. Reverse Transcription-Polymerase Chain Reaction (RT-PCR) is currently the gold standard for SARS-CoV-2 detection. The RT-PCR test is invasive and uncomfortable, and non-reusable testing kits have led to significant supply chain deficiencies. SARS-CoV-2 infection can also be assessed with an anti-body test. The anti- body test is also invasive, requiring venipuncture which, in combination with a several-day processing time, makes it less ideal for rapid mass screening. In the current economic and social situation, there is a great need for an alternative SARS- CoV-2/COVID-19 detection method that is easily accessible to the public for repeated testing with high accuracy. In this paper, the authors propose CovidDeep, an easy-to-use, accurate, and pervasive SARS-CoV-2/COVID-19 detection framework. It combines features extracted from physiological signals using wearable medical sensors and simple-to-answer questions in a smartphone application-based questionnaire with efficient DNNs. The framework uses synthetic data generation to alleviate the need for large datasets. Training of CovidDeep DNNs based on the grow-and-prune synthesis paradigm enables them to learn both the weights and the architecture during training. Hence, these DNNs can be easily deployed on edge devices (e.g., smartphones or smartwatches) as well as servers. CovidDeep was evaluated based on data collected from 87 individuals. The highest accuracy it achieves is 98.1%. They also obtained high enough test accuracies for many different sets of sensor/questionnaire data categories. Thus, users can choose the DNN model that is based on the sensors that are most conveniently accessible to them from the market. With more data collected from larger deployment scenarios, the accuracy of CovidDeep DNNs can be improved further through incremental learning.

INTERVIEW WITH PROF. LU ZHIHUI, FUDAN UNIVERSITY CHINA

Editor: Yafei Hou

Prof. Lu Zhihui

Zhihui Lu is a Professor at School of Computer Science, Fudan University. He received Ph. D from Fudan University in 2004, and he is a member of the IEEE and China computer federation's service computing specialized committee. His research interests are cloud computing and service computing technology, big data architecture, mobile edge computing, and IoT distributed system.

What are the major missions and main research topics of your team?

Cloud Computing, Distributed/Parallel Computing, and Mobile Edge Computing.

You have got a lot of research funds from both industrial also government. Could you briefly introduce your projects and which has been the most challenging so far?

For projects, I have not only hosted some funds supported by government, but also hosted some projects from the industrial companies. These projects include National Key Research and Development Programs, National Natural Science Foundations and Shanghai Science and Technology Innovation Action Plan Projects, and Enterprise cooperation project. I think the most challenging project is our current National Natural Science Foundation "The research of collaborative processing technology of intelligent tasks in edge-cloud orchestrated architecture".

The hybrid architecture of edge and cloud computing provides important support for edge intelligent task processing, and can process the highly dispersed massive data generated by terminal devices at the edge. However, there are still some problems to be solved. For example, the mismatch between resource-intensive intelligent model and resource-limited devices, the low efficiency of model collaborative scheduling strategy in a heterogeneous environment, and the lack of security and reliability of edge nodes and end devices. Aiming at these problems, this project will firstly help to improve the development of collaborative processing technology of data intelligent tasks under edge-cloud orchestrated architecture. Secondly, key algorithms and models support for the effective utilization of resources, efficient scheduling of models, and reliable guarantee of data storage in the edge-cloud collaborative environment will be provided. Finally, theoretical foundations and valuable practical explorations for the industrial application of edge computing will be consolidated, facilitating the application of basic research.

What are the main research directions of wireless service and cloud computing technologies for the next decade from the your view?

I think the main research directions of wireless service and cloud computing technologies for the next decade are mobile edge computing, edge-AI and Cloud-Edge resource collaborative scheduling etc.

The technologies of AI and IoT are definitely driving forces for future wireless system. In your opinion, how do you think these technologies will change the research directions of wireless service and cloud computing technologies and smart device of consumer electronics?

I think the technologies of AI and IoT can be considered as a complementary package towards the research directions of wireless service and cloud computing technologies and smart device of consumer electronics. From this perspective, it is essential to understand the role of these significant components that will provide a comprehensive vision for the worldwide smart city project in the near future. It is also essential to consider the emerging technologies-based intelligent applications for better lifestyle and more optimized solutions in our daily life.

In your opinion, which research topics are more important or practical for the wireless network systems in next decade?

I think the following research topics are more important or practical for the wireless network systems:

1. Develop a scheduling technique when users are dynamic
2. Develop a conflict-free scheduling algorithm which runs on the base station for which the tasks will be assigned from mobile user to a particular sensor node
3. Using data filtering or data compression method, unwanted sensory data can be minimized.

Could you provide some comments on the integration of wireless applications or high-performance content delivery in large-scale smart devices or consumer electronic equipment for Society 5.0?

For the integration of wireless applications or high-performance content delivery, we mainly focus on the intelligence edge computing. In large-scale smart devices or consumer electronic equipment for Society 5.0, Intelligence Edge Computing (IEC) is the key enabler of emerging 5G technologies networks and beyond. IEC is considered to be a promising backbone of future services and wireless communication systems in 5G integration. In addition, IEC enables various use cases and applications, including autonomous vehicles, augmented and virtual reality, big data analytic, and other customer-oriented services. Moreover, it is one of the 5G technologies that most enhanced market drivers in different fields such as customer service, healthcare, education methods, IoT in agriculture and energy sustainability.

How do you think the integration of your research results on the consumer electronic systems?

Our research results involve cloud-edge hybrid architecture, edge AI, etc. We believe these technologies can be combined with the consumer electronic systems to promote the development of the industry.

Do you have some messages of encouragement to young researchers potentially interested in your research field?

From the beginning of your research, the choices you make in what you pursue will have a major impact on when you can begin your research work. You will need to balance your passion for an area or particular project with a realistic appraisal of how long the project will take to complete. Nevertheless, science is more of a calling than simply a job, and it is your passion for the work that will sustain you throughout your life. For my research field (i.e. distributed computing, edge computing etc.), there will be many challenges and opportunities to promote the development of the industry and change the definition of the world. Hence, try to do it, young man.

MULTI-MODAL PRE-TRAINING: A NEW PARADIGM FOR MULTI-MODALITY UNDERSTANDING

Bei Liu
Microsoft Research Asia
Beijing, China
bei.liu@microsoft.com

Jianlong Fu
Microsoft Research Asia
Beijing, China
jianf@microsoft.com

Abstract

Pre-training has been an emerging topic that provides a way to learn powerful representation for downstream tasks in many fields (e.g., natural language processing, computing vision). In the last few years, we have witnessed many research works on multi-modal pre-training, especially in the visionlanguage domain. Pre-training models achieve state-of-the-art performances in many downstream tasks. They outperform traditional models by a large margin with a very simple design and demonstrate the superiority of pre-training on a large scale of data. In this article, we will guide you to see the power of multi-modal pre-training and introduce our exploration in this direction.

I. WHAT IS MULTI-MODAL PRE-TRAINING?

A modality in the context of human-computer interaction is defined as the classification of a single independent channel of sensory input/output between the human and the computer [6]. Multi-modality focuses on studying the integration of multiple communicative modalities, such as vision, language, audition, depth, etc. Technologies in a single modality have achieved a high level and machines can even outperform human in some tasks (e.g., image classification, language translation, speech recognition). However, the real world is multi-modality and we humans interact with the world in multiple senses. To accelerate the progress of human-like AI, multi-modality learning becomes much more critical and attracts more attention from industrial scenarios in recent years. Using vision and language modalities for example, there are many tasks proposed, like automatically generate one or several sentence in natural language given visual signals (i.e., image/video captioning [4], visual storytelling [7], imagebased poem generation [9]) or vice versa, automatically generate visual signals guided by language [1], vision-language alignment (i.e., image/video-text retrieval, image language grounding, video temporal localization), visual question and answering [10], etc.

Pre-training has shown great potential in many

domains, especially with the development of deep learning. Compared with traditional training with hand-crafted features, deep learning requires much more amount of data to learn the hidden features. While many tasks as mentioned above have limited supervised data in certain scenarios, a model pre-trained with a large scale of data provides a much better initial representation for faster convergence and higher performance. Forexample, many computer vision models use a backbone pretrained with ImageNet [2], and many recent natural language models utilize BERT [3] for initialization. Multi-modality pre-training is a new paradigm that provides better crossmodal representation for downstream multi-modality tasks by learning from a large scale of multi-modal data with welldesigned pre-training tasks. Figure 1 shows a general pipeline of a multi-modal pre-training. Each modality is fed to its own encoder for representation learning, and a multi-modal representation learning model is designed for joint learning of multiple modalities with designed general pre-training tasks. Pre-training model is used as initialization and then fine-tuned on different downstream tasks to achieve good performance.

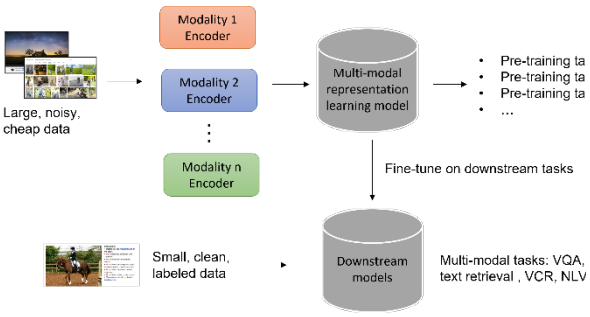


Fig. 1. Multi-modal pre-training pipeline for multi-modality tasks, using vision-language as an example.

II. WHY WE NEED MULTI-MODAL PRE-TRAINING?

In the multi-modality domain, the amount of high-quality data is limited and the annotation of multi-modal data is very costly. For example, MSCOCO [8], which is a widely-used image-description dataset for image-text retrieval and image captioning, costs 108,000 dollars to label 5 sentences for all 120,000 images. On the other side, there are a large amount of weakly paired data available on the Internet, such as imagetext pairs, and video-transcript-audio data. The power of pretraining is intuitive. A machine that has seen a lot of data in advance and performs well at some pre-defined multi-modal tasks can better perform multi-modal tasks compared to the one trained from scratch.

III. HOW MULTI–MODAL PRE-TRAINING WORKS?

The key to multi-modality learning is the alignment between different modalities. This is challenging due to many aspects. First, the representation of each modality is different which makes the alignment difficult to learn. For example, the representation of images (i.e., RGB) is real-valued and dense while language (i.e., word token) is represented in discrete and sparse form. Second, different modalities are not exactly matched which makes the learning of alignment even harder. For example, a sentence for an image can only indicate part of the information in the image and we cannot picture a whole video with only its audio. Third, it is difficult to directly evaluate the goodness of the alignment learned by a pre-trained model.

In this section, we will introduce three research works that we have done to tackle the above challenges.

A. End-to-end image-language pre-training

In early works of image-language pre-training, image representation is usually fixed by using region-based image features following previous works on image-language tasks (e.g., image captioning, image question and answering). Having an image, we first extract regions of objects in the image and use the visual features of these regions as input for multi-modal learning. However, there are three drawbacks to using regionbased features. First, region-based features only

focus on the foreground objects in the images while neglecting the context in the background of images. Context is not that critical for object-centered tasks (i.e., image classification and object detection) while for language-related tasks, context is much more important. Second, the visual representation of images is limited to the pre-defined categories while the semantics in the language domain is much larger. Third, as the object detector used to extract visual features is too heavy to be jointly optimized with multi-modality learning, the extracted region features cannot be optimized for target cross-modal tasks.

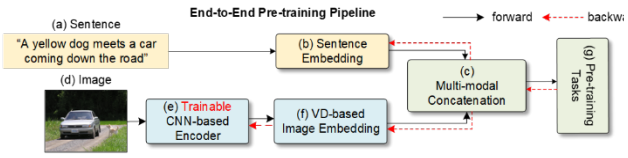


Fig. 2. SOHO: the first end-to-end vision-language pre-training framework[5].

To overcome the above shortages of pre-extracted regional features, we propose the first end-to-end image-language pretraining model to See Out of tHe bOx (SOHO) [5] for better cross-modality learning in CVPR 2021. As illustrated in Figure 2, for an image-text pair, we use Transformerbased text embedding as a language encoder and a trainable CNN-based visual encoder to extract visual representation. In this design, we do not need an object detection model and the information we can learn from images is not limited to pre-defined categories. The visual backbone is optimized in an end-to-end fashion and visual features can be updated in alignment with language.

Different from language modality where each word has its own particular meaning, pixels in images often share the same semantics. To better align image and text in the semantic level, we group pixels at the feature level with similar features into one item to indicate a consistent semantic. This is achieved by applying a visual dictionary (VD)-based image embedding to the image encoder outputs. Text embedding and VD-based embedding are then concatenated for three designed pre-training tasks: image-text matching, masked language modeling, and masked vision modeling. Through this work, we find that end-to-end learning can result in a better representation of multimodality.

B. High-resolution video-language pre-training

Data is one of the main factors in deep learning-based models. In the video-language domain, the datasets are limited in either scale or scope. Early datasets that use videos and annotated descriptions are limited in scale due to the heavy cost of annotation. The most used large-scale video-language dataset for pre-training (i.e., HowTo100M) consists of only instructional videos with their transcripts. Thus in a videolanguage pre-training work accepted by CVPR 2022 [11], we collect a video-language dataset (HD-VILA-100M) to overcome both limitations.

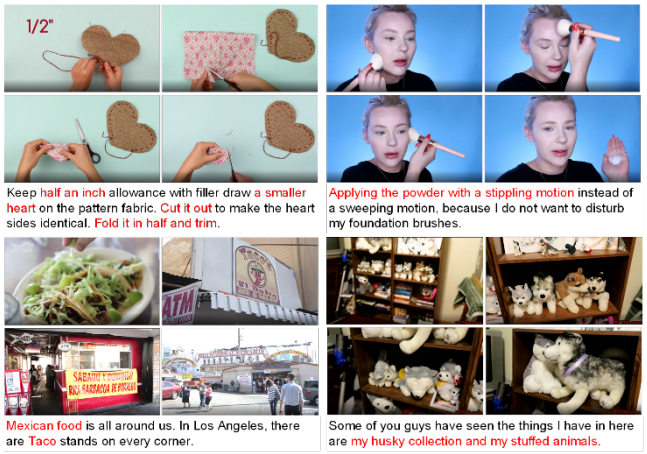


Fig. 3. One example of video-language pair in HD-VILA-100M [11].

We use transcripts along with videos from YouTube as the source of our dataset. Figure 3 shows an example in HDVILA- 100M dataset. HD-VILA-100M has three key properties. First, it is one of the largest video-language datasets. It includes 100 million video clips and transcript pairs from 3.3 million videos. It covers 371.5K hours in total which is 2.8 times than HowTo100M dataset in duration. The average length of each sentence is 13.4 which is about 8 times longer than HowTo100M. This ensures the richness of semantic in language. Second, all the videos are in high resolution with 720p. The quality of videos is

much higher than most video datasets that are 240p or 360p. Third, the dataset is diverse and balanced in consideration of topics as shown in Figure 4. It covers 15 popular categories on YouTube and the number of video clips in each category is balanced.

To efficiently utilize the high-resolution videos in pretraining, we propose to form a hybrid image sequence which consists of one high-resolution frame and several surrounding low-resolution frames from a video clip. The hybrid image sequence is then fed into a novel hybrid video encoder that learns spatiotemporal information with a hybrid Transformer. Since the alignment between videos and transcripts is not as high as video-description pairs, we adopt contrastive learning to ensure paired data are close to each other while unpaired ones are far from each other.

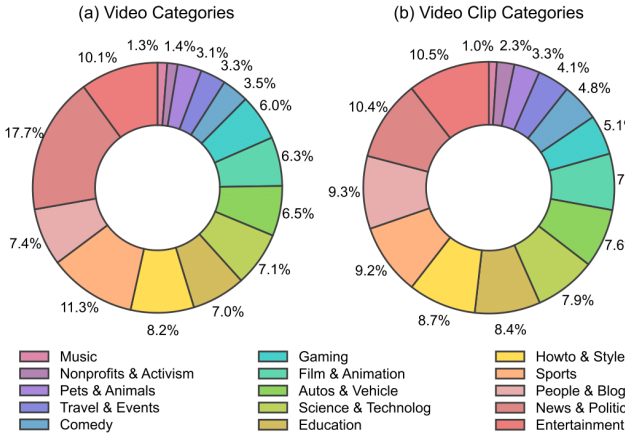


Fig. 4. Distribution of categories in HD-VILA-100M.

C. Probing inter-modality in vision-language pre-training

It is essential to learn the relation between different modalities in multi-modality tasks. In the vision-language domain, learning the inter-modal alignment between visual information and language semantics is very important. For language, the structured text with grammar makes it easy to learn the intrarelation of words. While for images, image features with CNN-based backbones (e.g., grid or regional feature) lacks the global relationship learning between different semantics. The inconsistent representation of visual and language modalities adds the burden of intra-modal learning on the visual side and inter-modal learning of both modalities encapsulated in the multi-modal module. This makes the learning of alignment even hard.

In our paper published in NeurIPS 2021 [12], we propose the first fully Transformer-based image-language pre-training model as shown in Figure 5. By adopting self-attention for visual feature learning, the spatial inductive bias is not introduced and we can learn long-range global relations of visual semantics before joint learning. This ensures the multi-modal Transformer is more specialized for cross-modal joint learning. To further measure the fusion quality of inter-modality learning, we propose the Inter-Modality Flow (IMF) metric to compute the information flow between two modalities.

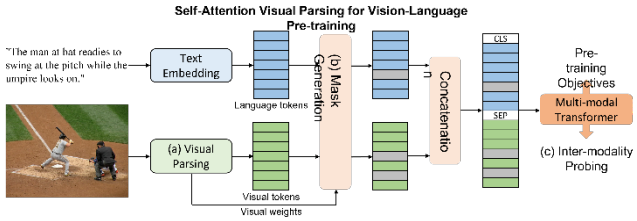


Fig. 5. The first fully Transformer-based image-language pre-training model[12].

IV. WHAT IS THE NEXT?

We can see many research works in multi-modality pretraining in the past few years, especially in the vision-language domain. However, how to efficiently use pre-trained models in industry scenarios still faces many challenges. First of all, pre-training models are usually too heavy while real-time computing is often required in real applications. Secondly, how to bridge the domain gap between pre-training data and the real data in the wild (e.g., health care, navigation, digital human) remains a problem. Moreover, as we have claimed above, the real world is about much more modalities than vision and language. How to effectively learn the joint representation and alignment between more than two modalities is worth studying. For example, in

the domain of embodied AI, more modalities (such as depth, action, segmentation, etc.) are involved. How to learn the alignment between different modalities when it is more complex will be a problem. No matter how difficult the path ahead is, we still believe multimodality pre-training is the right way we need to walk towards the real AI.

REFERENCES

[1] Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin,Xiaoyu Qi, Chunting Wang, and Jin Zhou. Neural storyboard artist: Visualizing stories with coherent image sequences. In ACM MM, pages 2236–2244, 2019.
[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248–255. Ieee, 2009.
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[4] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In ACM MM, pages 1138–1147, 2021.
[5] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In CVPR, pages 12976–12985, 2021.
[6] Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh, and Mo Nours Arab. Human-computer interaction: Overview on state of the art. International Journal on Smart Sensing and Intelligent Systems, 1(1), 2017.
[7] Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. Emotion reinforced visual storytelling. In ICMR, pages 297–305, 2019.
[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755. Springer, 2014.
[9] Bei Liu, Jianlong Fu, Makoto P Kato, and Masatoshi Yoshikawa. Beyond narrative description: Generating poetry from images by multiadversarial training. In ACM MM, pages 783–791, 2018.
[10] Bei Liu, Zhicheng Huang, Zhaoyang Zeng, Zheyu Chen, and Jianlong Fu. Learning rich image region representation for visual question answering. arXiv preprint arXiv:1910.13077, 2019.
[11] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In CVPR, 2022.
[12] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. NeurIPS, 34, 2021.