TECHNICAL UNIVERSITY OF MUNICH

TUM Data Innovation Lab

"Active Scene Understanding for Video Labeling" - Public Version

Authors	Morgane Ayle, Andreas Döring, Andrei Goncharov, Md Siyam Sajeeb Khan, John Rachwan
Mentors	M.Sc. Keesiu Wong (Design AI)
	M.Sc. Frederik Mattwich (Design AI)
Co-Mentor	M.Sc. Marija Tepegjozova (Department of Mathematics)
Project Lead	Dr. Ricardo Acevedo Cabra (Department of Mathematics)
Supervisor	Prof. Dr. Massimo Fornasier (Department of Mathematics)

Feb 2021

Abstract

Annotation of video data requires a large amount of manual labor. Annotations are short descriptions of the events occurring in a certain scene. Additionally, they are essential to the successful archiving of video data. We introduce a novel architecture: the Multi-Modal Transformer (MMT). It is a transformer architecture that leverages multiple modalities of a certain video to: (1) clip a video into smaller segments that correspond to different events, and (2) provide annotations that describe these segments. We evaluate our model and show that it surpasses the state of the art competitors on all metrics. Contrary to the usual development of Deep Learning models, it is more beneficial for the model to adapt and learn continuously when it is deployed in production. Indeed annotations are customer-specific and the model should be able to meet these specific needs. Consequently, we propose Adaptive Long Term Captioning (ALTC), an online learning strategy that enables our model to both learn to predict new words, and preserve its memory in the long and short term. We show that ALTC not only reaches the same accuracy as offline learning but surpasses it.

Contents

1 Introduction

1.1 Motivation

The amount of data generated and stored is steadily increasing across different domains. Notably, many companies from various industries rely on the storage of large amounts of videos in their databases. Depending on the company's needs, the videos need to be annotated to facilitate certain downstream tasks such as archiving, video retrieval, video analysis, etc. This task is typically done by human annotators, which renders the whole annotation process slow, inefficient, very costly for companies, boring for employees, and leads to only partially annotated databases. The goal of this project is to speed up and automate the annotation process while still achieving human quality annotations by automatically captioning videos using natural language sentences. Said annotations are very customer-centric, hence, our end product should be able to continuously learn the customers' preferences.

1.2 Objectives

The goal of this project is to create a full technical prototype, that is the core of a minimum viable product (MVP). This report focuses on the Deep Learning model that powers the product. We divided the development of our model into 4 main phases:

1. Research: Identify the most promising approaches for video understanding and captioning.
2. Prototyping: Develop, implement and evaluate the necessary individual modules on a public dataset.
3. Integration: Combine the different modules into a unified continual learning framework.
4. Deployment: Integrate the unified framework into the MVP's back end architecture.

1.3 Our Approach

We identify Dense Video Captioning, the task of analyzing a video and captioning different proposed sequences, as the closest solution to our problem from the literature. We create a multi-modal transformer that leverages multiple modalities to extract high-quality annotations. Our proposed model surpasses the state of the art in Dense Video Captioning when tested on the dataset ActivityNet Captions [16]. Moreover, to allow our model to adapt to customer-specific needs, we propose a novel continual learning mechanism: Adaptive Long Term Captioning (ALTC). ALTC not only matches the performance of the model if it were trained offline but surpasses it due to its capability of extending its own vocabulary when introduced with new words.

2 Related Work

2.1 Dense Video Captioning

2.1.1 Introduction

Video content analysis is a complex field which includes a diverse set of tasks, such as action recognition, object recognition and tracking, sentiment analysis or question answering. We chose the task of *video captioning* as a core of the product. Video captioning is the task of generating a text description for a given video clip. The current models are solving this task by recognizing the activity which is present in the clip and generating the natural language sentence which is describing this activity alongside other attributes. In order to analyze long video sequences which contain multiple scenes with different activities, it is necessary to divide the original video into a set of segments and generate captions for every segment, which is known as *dense video captioning*. The classical approach to this task is to analyze only the visual component of the video content, but other modalities such as audio or speech have shown to largely contribute to the understanding of the semantic meaning of the video content. Analyzing multiple modalities can improve the quality of the captions significantly. This approach is known as *multi-modal dense video captioning*.

2.1.2 Literature Review

The problem of Dense Video Captioning (DVC) was introduced by Krishna et al. [16] alongside a new dataset called ActivityNet Captions. The task was naturally divided into two subtasks: generating the segment proposals from the original video (proposal module) and then captioning the proposals (captioning module). The spatiotemporal features [13] were extracted from the original videos and used as an input to the modules. Both proposal and captioning modules were based on the LSTM networks [7] to capture contextual information from past and future events.

Zhou et al. [33] proposed an end-to-end Transformer model for DVC. The encoder was used to encode the video into appropriate representations, the proposal decoder formed event proposals and the captioning decoder employed a masking network to restrict its attention to the proposal event over the encoding feature. In addition, the model employed a self-attention mechanism [28], which enabled the use of an efficient non-recurrent structure during encoding.

Moving to the multi-modal DVC, Rahman et al. [24] were first to include the audio modality into DVC. Shi et al. [27] added the speech modality to improve the understanding of cooking videos, using Transformer for video encoding, the pre-trained BERT model [6] to create the embeddings for subtitles, and LSTM models for proposal and captioning modules.

Iashin et al. [12] went further and used the mixture of three modalities: video, audio and speech. Separate Transformers were used in the captioning module to process the features extracted from the modalities. The output of the Transformers were concatenated and fed into several fully-connected layers in order to predict the caption sequence. The pre-trained Bi-directional SingleStream Temporal action proposals network (Bi-SST) proposed in [3] was employed as a proposal module. The video and audio features were

extracted using the pre-trained I3D [4] and VGGish [9] models respectively, and the external ASR module (youtube caption generator) was used to generate the speech transcript. This model was extended by Chadha et al. [5] by building a common-sense knowledge base using contextual cues to infer causal relationships between objects in the video. Common-sense vectors generated by the knowledge base generator were concatenated with the I3D features and fed together into the Transformer.
Iashin et al. proposed another version of the DVC model using a Bi-Modal Transformer [11]. In this model, only two modalities were used as an input (video and audio), but the Transformer architecture was adapted for a bi-modal input. The audio and video features are extracted and passed through the bi-modal N-layered encoder to produce bi-modal sequence representations utilizing novel bi-modal multi-headed attention blocks to fuse the features from both sequences. The pre-trained bi-modal encoder was also used as a feature extractor for the proposal generation module. Additionally, the GloVe [23] pre-trained model was used for word embeddings. The results of this model were considered SOTA at the end of the year 2020 based on the ActivityNet benchmark results.

2.1.3 Datasets

There are multiple commonly used dense video captioning datasets. Some of these datasets, such as YouCook II [32], are application specific, while others are open and contain different kinds of video types [1]. Our model should have a broad understanding of various video types, so that it can act as a base for future customer-specific video types. Hence, application specific datasets were disregarded. Out of the remaining datasets, ActivityNet Captions [16] stood out. It has 20,000 videos amounting to 849 hours with 100,000 descriptive sentences, each with unique start and end times. The average video length is 180 seconds and the sentences have an average length of 13.48 words [16]. It is thereby the largest open dataset for dense video captioning [1] and contains longer videos than many of the other datasets, which is why we chose it.

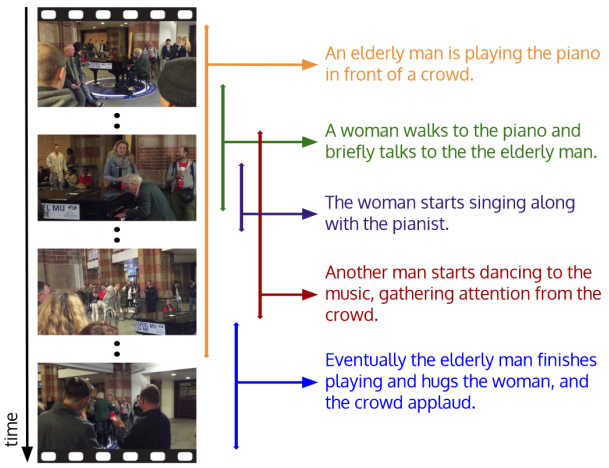

Figure 1: ActivityNet Captions dataset with multiple temporally localized events. [16]

2.2 Online Learning

2.2.1 Introduction

In Supervised Learning, an agent is trained once on labelled data and then deployed in the real world. While this works well for static settings, our work is very customer dependant and the model should be able to adapt to the customer's needs and preferences. In this paradigm, known as continual learning, the model is trained online on a single pass through the data stream that cannot be assumed to be i.i.d. Streaming Learning

causes conventional Neural Networks to fail for two main reasons: (1) They are trained with multiple passes through the dataset; and (2) non-i.i.d. data will cause catastrophic forgetting [14], the concept of forgetting previously acquired knowledge when subjected to new and different knowledge. An old fix to both of these issues is coined "Rehearsal". When a new example arrives, it is mixed with old examples that the model has already seen, and then this mixture is used to update the model. Full rehearsal is memory intensive and slow as it has to store all previous data the model has seen. In our use case, this data increases a lot as video warehouses tend to have a large amount of videos. That being said, Full Rehearsal [8] has been shown to prevent catastrophic forgetting in many continual learning settings.
During the past few years we have witnessed a renewed and growing attention to Continual Learning [20]. However, the field is still very premature, focusing mainly on standard classification problems and thus the proposed methods are task-dependant and directly using them in a different setting such as Dense Video Captioning is not possible. Therefore, both adaptation and testing are required in order to confirm a method's validity in new tasks.

2.2.2 Literature Review

The sudden interest in Continual Learning (CL) and its applications, especially in the sense of deep architectures, has led to rapid progress and initial research directions, leaving the research community without common terminology and specific objectives. In line with [14] and [31], here we suggest a three-way fuzzy categorization of the most common CL strategies:

- **Architectural techniques**: Use complex architectures, layers, activation functions, and/or weight-freezing strategies to mitigate forgetting.
- **Regularization strategies**: The loss function is expanded with regularization terms in order to facilitate selective consolidation of the weights that are essential for preserving past memories. This includes fundamental methods for regularization such as weight sparsification, dropout or early stopping.
- **Rehearsal strategies**: The model is regularly fine-tuned with past knowledge to reinforce links to memories it has already acquired. An easy solution is to store part of the previous training data and to interconnect it with new training data. Pseudo-rehearsal of generative models is a more difficult approach.

In Figure 2, we show the multiple reviewed methods in the categorization we have just described. Although more methods are being discovered in the respective categories, research has been extensive in their combination, especially at the intersections of the three categories.
One of the first architectural techniques suggested is Progressive Neural Networks (PNN)[26], which is based on a combination of parameter freezing and network expansion. Although PNN has been shown to be successful on short series of simple tasks, the number of parameters of the model tends to increase at least linearly with the number of tasks, making it difficult to use for long sequences. CopyWeights with Re-init (CWR) [18], which has recently been proposed, is a simpler and lighter counterpart to PNN (at the expense of

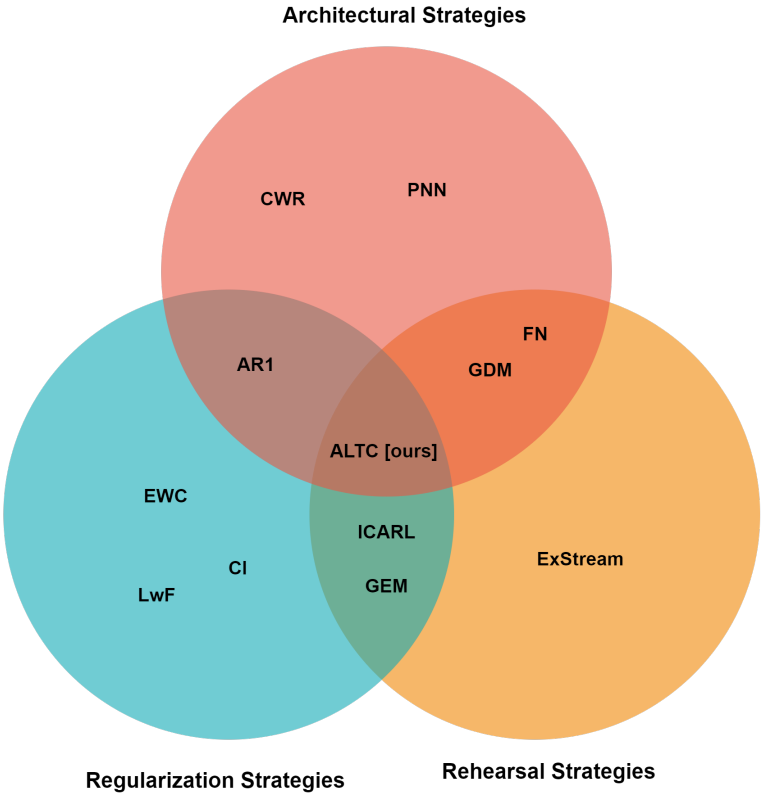

Figure 2: Taxonomy of Continual Learning Strategies

lower flexibility) with a fixed number of shared parameters and has already been shown to be useful for longer task sequences.
Learning Without Forgetting (LWF) [17] is a regularization technique that tries to maintain the accuracy of the model on old tasks by enforcing consistency of performance through distillation of knowledge [10]. Elastic Weights Consolidation (EWC) [15] and Synaptic Intelligence (SI) [31] are other well known regularization techniques, each expressed around a weighted quadratic loss of regularization that penalizes moving weights. ICARL [25] and GEM [19] are illustrated at the intersection of rehearsal and regularization strategies. The former requires an external fixed memory to store a subset of old task data based on an elaborate sample selection procedure, but also employs a distillation step acting as a regularization. The latter, referred to as Gradient Episodic Memory, uses a fixed memory to store a subset of old patterns and apply regularization constraints to the optimization of losses, aimed not only at regulating forgetting, but also at improving accuracy on previous tasks while learning the subsequent ones (a "positive backward transfer" phenomenon). Recent research on the memory-efficient implementation of pure rehearsal strategies is given in [8], where a modern partitioning-based stream clustering strategy called ExStream has been shown to be very competitive with a full rehearsal approach (storing all previous data) and other memory management techniques.

3 Results and Discussion

3.1 Evaluation Metrics

For evaluating the quality of our generated video captions we choose two of the most used evaluation metrics: BLEU [22] and METEOR [2].

BLEU: Bilingual Evaluation Understudy (BLEU), was proposed to measure the quality of machine translated sentences at a corpus level. However, it has also been extensively used in video captioning tasks. BLEU evaluates a machine generated caption with the ground truth or the reference caption(s) by forming n-grams (a group of n adjacent words) from the generated caption and looking for matches at the n-gram level in the ground truth caption(s). This accounts for the BLEU-N score where N actually corresponds to the number of n-grams used while matching. For example, BLEU-3 and BLEU-4 (also referred to as B@3 and B@4) uses 3-grams and 4-grams respectively to match the generated caption with the ground truth. These are the most used measurements for evaluating the quality of captions. In practice the logarithm of BLEU score is used [1]:

$$log\ BLEU = min(1 - \frac{l_r}{l_c}, 0) + \sum_{n=1}^{N} w_n\ logp_n \tag{1}$$

where l_r/l_c corresponds to the ratio between the lengths of the reference caption(s) and the machine generated caption respectively. w_n refers to the positive weights and p_n refers to the geometric mean of the modified n-gram precisions. BLEU is a precision based metric, it tends to favor shorter captions. To rectify this, a brevity penalty is used (second part of the equation) which penalizes captions which are shorter than the reference captions. The BLEU score ranges from 0 to 1 with 0 meaning there is no correspondence between the ground truth caption and the generated caption and 1 meaning they are exactly the same.

METEOR: Metric for Evaluation of Translation with Explicit Ordering (METEOR), was proposed in 2005 to rectify the shortcomings of the precision based metric BLEU. Contrary to the BLEU, METEOR introduced a recall based evaluation scheme which eradicated the shortcoming of the exact word matching mechanism of BLEU. In addition to the exact word matching, METEOR also matches word stems, synonyms and paraphrases. The METEOR score comprises of uni-gram based precision (P), recall (R) and a F-score. The precision and recall are combined in the following way [1]:

$$P = \frac{m_{cr}}{m_{ct}},\quad R = \frac{m_{cr}}{m_{rt}},\quad F_{mean} = \frac{10PR}{R + 9P} \tag{2}$$

Here m_{cr} corresponds to the number of uni grams that were matched in both generated and the reference captions, m_{ct} stands for the uni-gram count in the generated and m_{rt} is the total uni-grams present in the reference caption respectively. Like BLEU, the METEOR score also ranges from 0 to 1 with 0 meaning no correspondence to 1 meaning exact correspondence between the two captions.

3.2 Final Results

3.2.1 Offline Learning

We report our model's results using the B@3, B@4, and METEOR metrics described in Section 3.1. Our model is trained and tested on the regular ActivityNet Captions, which consists of 100k temporally localized sentences for 20k YouTube videos. The dataset is split into 50/25/25% parts for training, validation, and testing. Since ActivityNet Captions is a challenge, the test set is not provided with the ground truth and hence we will be reporting our results on the validation set. Moreover, the dataset is a collection of Youtube videos. This means some of the videos have been removed since the inception of the challenge. Therefore, we only possess 91% of the original training dataset. All metrics are averaged for every video with temporal Intersection over Union thresholds: [0.3,0.5,0.7,0.9]. The original evaluation script had a mistake as was discovered by [21]. Henceforth we use the updated evaluation script in our reported results.

Comparison to the State of the Art. We compare our Multi-Modal Transformer with multi-headed proposal generator to other methods in the literature of DVC [21, 17, 33, 29, 12, 11, 24, 30, 16]. The results of the comparison for captioning both ground truth and learned proposals are shown in Table 1. Evaluating captioning modules is still an open research problem and the METEOR metric is only a proxy. Therefore, we omit the results of certain models that employ Reinforcement Learning (RL) to optimize the METEOR specifically, especially because any model can be adapted to perform RL after Supervised Learning (SL). Instead, we report these models' scores after solely training in a supervised learning setting.
According to the results, our model outperforms all other models on both ground truth and Learned Proposals. Although, this is still not a fair comparison since our model uses only 91% of the training dataset.

	Full Dataset Available	GT Proposals			Learned Proposals		
		B@3	B@4	METEOR	B@3	B@4	METEOR
Mun et al.	Yes	-	-	-	-	-	6.92
Krishna et al.	Yes	4.09	1.60	8.88	1.90	0.71	5.69
Li et al.	Yes	4.51	1.71	9.31	2.05	0.74	6.14
Zhou et al.	Yes	5.78	2.71	11.16	2.91	1.44	6.91
Wang et al.	Yes	-	-	10.89	2.27	1.13	6.10
Teng et al.	No	-	-	11.49	-	-	7.65
Iashin et al.	No	4.52	1.98	11.07	2.53	1.01	7.46
Rahman et al.	No	3.04	1.46	7.23	1.85	0.90	4.93
BMT	No	4.63	1.99	10.90	3.84	1.88	8.44
MMT [Ours]	No	**5.83**	**2.86**	**11.72**	**4.00**	**2.01**	**9.43**

Table 1: Experimental Results of State of the Art DVC models compared to MMT trained on ActivityNet Captions on both GT and Learned Proposals

3.2.2 Online Learning

We evaluate ALTC on the original BMT model [11]. This is due to the concurrent development of both the improved model and the online learning strategy. Therefore, we had to use a fixed model to train and test the validity of our Online Learning methods. We report our model's results using the B@3, B@4, and METEOR metrics. We use the ActivityNet Captions dataset with a different split. First we train an offline BMT model on the full training dataset. Then we split the validation set into a 80/20 split. We train our model using ALTC on 80% of the validation set and report our results on the remaining 20%. Since ALTC is a captioning strategy, it is more beneficial to report our results on the captioning module using ground truth proposals. We additionally compare our online model to two offline baselines: (1) BMT trained only on the orignal training dataset, and (2) BMT trained on the original training dataset and 80% of the validation set.

Ablation Study. We perform an ablation study to show the impact of each component separately. Our results are shown in Table 2. We can see that the BMT model which is additionally trained on the 80% of the validation dataset in an offline manner achieves a METEOR score of 11.12. Our goal would be to be able to match this performance when we train on the same 80% of the data in an online manner. We can see that our novel online strategy ALTC shows an improvement over the offline model.
In short, our proposed online strategy, ALTC, not only matches the performance of our offline learning baseline, but surpasses it.

Method	B@3	B@4	METEOR
BMT(offline + Train Set)	4.9	2.3	10.51
BMT(offline + Train Set + 80% Val Set)	5.66	2.71	11.12
ALTC	**5.7**	**2.8**	**11.23**

Table 2: Online Learning Experimental Results

4 Conclusion

In this report, we introduced a novel Dense Video Captioning model, the Multi-Modal Transformer, that suggests important segments of a video and provides captions for each of them. Our modal leverages multiple modalities to perform the aforementioned tasks. We show that our model surpasses all of the current state of the art in dense video captioning on B@3, B@4, and METEOR scores evaluated on the ActivityNet Captions dataset. Additionally, we proposed an online learning strategy, Adaptive Long Term Captioning, that leverages adaptations of architectural, rehearsal, and regularization strategies to allow the model to adapt to the customer's needs and continuously learn and improve after being deployed in production. We show that ALTC not only matches the performance of offline learning, but, with the added capability of learning new words, surpasses it.

References

[1] Nayyer Aafaq et al. "Video description: A survey of methods, datasets, and evaluation metrics". In: *ACM Computing Surveys (CSUR)* 52.6 (2019), pp. 1–37.

[2] Satanjeev Banerjee and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005, pp. 65–72.

[3] S. Buch et al. "SST: Single-Stream Temporal Action Proposals". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6373–6382. DOI: 10.1109/CVPR.2017.675.

[4] Joao Carreira and Andrew Zisserman. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. 2018. arXiv: 1705.07750 [cs.CV].

[5] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. "iPerceive: Applying Common-Sense Reasoning to Multi-Modal Dense Video Captioning and Video Question Answering". In: *arXiv preprint arXiv:2011.07735* (2020).

[6] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[7] Jeff Donahue et al. *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. 2016. arXiv: 1411.4389 [cs.CV].

[8] Tyler L. Hayes, Nathan D. Cahill, and Christopher Kanan. *Memory Efficient Experience Replay for Streaming Learning*. 2019. arXiv: 1809.05922 [cs.LG].

[9] Shawn Hershey et al. *CNN Architectures for Large-Scale Audio Classification*. 2017. arXiv: 1609.09430 [cs.SD].

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[11] Vladimir Iashin and Esa Rahtu. *A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer*. 2020. arXiv: 2005.08271 [cs.CV].

[12] Vladimir Iashin and Esa Rahtu. *Multi-modal Dense Video Captioning*. 2020. arXiv: 2003.07758 [cs.CV].

[13] S. Ji et al. "3D Convolutional Neural Networks for Human Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 221–231. DOI: 10.1109/TPAMI.2012.59.

[14] Ronald Kemker et al. "Measuring Catastrophic Forgetting in Neural Networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/11651.

[15] James Kirkpatrick et al. "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the National Academy of Sciences* 114.13 (2017), pp. 3521–3526. ISSN: 0027-8424. DOI: 10.1073/pnas.1611835114. eprint: https://www.pnas.org/content/114/13/3521.full.pdf. URL: https://www.pnas.org/content/114/13/3521.

[16] Ranjay Krishna et al. *Dense-Captioning Events in Videos*. 2017. arXiv: 1705.00754 [cs.CV].

[17] Zhizhong Li and Derek Hoiem. *Learning without Forgetting*. 2017. arXiv: 1606.09282 [cs.CV].

[18] Vincenzo Lomonaco and Davide Maltoni. "CORe50: a New Dataset and Benchmark for Continuous Object Recognition". In: *Proceedings of the 1st Annual Conference on Robot Learning*. Ed. by Sergey Levine, Vincent Vanhoucke, and Ken Goldberg. Vol. 78. Proceedings of Machine Learning Research. PMLR, Nov. 2017, pp. 17–26. URL: http://proceedings.mlr.press/v78/lomonaco17a.html.

[19] David Lopez-Paz and Marc'Aurelio Ranzato. "Gradient Episodic Memory for Continual Learning". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017, pp. 6467–6476. URL: https://proceedings.neurips.cc/paper/2017/file/f87522788a2be2d171666752f97ddebb-Paper.pdf.

[20] Davide Maltoni and Vincenzo Lomonaco. *Continuous Learning in Single-Incremental-Task Scenarios*. 2019. arXiv: 1806.08568 [cs.LG].

[21] Jonghwan Mun et al. "Streamlined Dense Video Captioning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[22] Kishore Papineni et al. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

[24] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. *Watch, Listen and Tell: Multi-modal Weakly Supervised Dense Event Captioning*. 2019. arXiv: 1909.09944 [cs.CV].

[25] S. Rebuffi et al. "iCaRL: Incremental Classifier and Representation Learning". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5533–5542. DOI: 10.1109/CVPR.2017.587.

[26] Andrei A. Rusu et al. *Progressive Neural Networks*. 2016. arXiv: 1606.04671 [cs.LG].

[27] Botian Shi et al. "Dense Procedure Captioning in Narrated Instructional Videos". In: Jan. 2019, pp. 6382–6391. DOI: 10.18653/v1/P19-1641.

[28] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[29] Tan Wang et al. "Visual commonsense r-cnn". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10760–10770.

[30] Teng Wang, Huicheng Zheng, and Mingjing Yu. *Dense-Captioning Events in Videos: SYSU Submission to ActivityNet Challenge 2020*. 2020. arXiv: 2006.11693 [cs.CV].

[31] Friedemann Zenke, Ben Poole, and Surya Ganguli. "Continual Learning Through Synaptic Intelligence". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 3987–3995. URL: http://proceedings.mlr.press/v70/zenke17a.html.

[32] Luowei Zhou, Chenliang Xu, and Jason J. Corso. *Towards Automatic Learning of Procedures from Web Instructional Videos*. 2017. arXiv: 1703.09788 [cs.CV].

[33] Luowei Zhou et al. *End-to-End Dense Video Captioning with Masked Transformer*. 2018. arXiv: 1804.00819 [cs.CV].