**Slide 1:**

Hello, I would like to present thesis about “ABAW based on facial video processing”. My supervisor is Professor Savchenko Andrey Vladimirovich.

**Slide 2:**

The increment of the facial videos is quite remarkable in recent years! It come from various sources such as video calls, social media like TikTok and Facebook Reel, and any camera system. Especially, TikTok has caught the attention of not just users, but also businesses. With this explosive growth in facial video data, we are faced with an exciting challenge. We need effective techniques to analyze and understand users within these videos. It's not just about capturing the data; we need to make sense of it all.

**Slide 3:**

Affective Behavior analysis in-the-wild is the technique that refers to the study and analysis of human emotions, expressions, and behaviors in the real-world. Not only face detection, face recognition, but also multiple methods, etc.,

**Slide 4:**

And of course, understand what users want allows us to do many thing. And a huge benefit could be gained by applying ABAW methods.

However, there are still existing challenges. The biggest challenge is related to personal identity information (PII) and data privacy. It is possible if your facial data is leaked and a hacker got it, he could perform a deepfake attack on your family. This issue makes collecting enough real-life data really difficult. This leads to the model performance degradation. The model performs excellent in laboratory environments, but failed in real-life condition.

**Slide 5:**

So my thesis, study the application of Affective behavior analysis in-the-wild based on facial video processing. The first objective is to demonstrate the benefit of using ABAW in a practical use case, enhance effectiveness for the outdoor advertising.

The second objective study the issue related to the data privacy during the development of an ABAW application.

Synthetic data is the first possoble solution for this issue. In the first task, I propose an ensemble approach for fer using synthetic dataset.

Running application on edge device is second solution to deal with data privacy problem. By performing all the necessary computations on the edge device, we can keep the data secure without any concerns about potential leaks during data transmission over the network. So in this task, I propose a video analytics application, which use ABAW, running on Jetson Nano.

**Slide 6:**

I have structured the remaining presentation into three distinct parts, each comprising a literature review, a proposed solution, and experimental results.

**Slide 7:**

Let’s revisit the market size of digital and outdoor advertising. It normally measured by an impressive billion USD. The success of social media and mobile platforms lies in their ability to deliver targeted and real-time advertisements by capturing and leveraging our behavioral data.

On the other hand, outdoor advertising is a distinct marketing channel. However, unlike digital advertising channels, the content in outdoor advertising is typically delivered randomly.

It lacks the ability to know their customer. So if we could apply an ABAW method such as age gender recognition to recognize the user, we could suggest the content better than the random approach. This could help to increase the effectiveness of outdoor advertising.

**Slide 8:**

So I propose to use Impression score to measure the effectiveness of the advertising campaign. I compare the Impression score of using agwe gender regconition with the random approach.

Let’s say we have an input face image with age 22, generation: Gen Z, gender: Female information. In a random approach, we randomly select an advertising content belonging to the group like Gen Y, Male. So we compare the correct category and the selected category. If it is equal, then we count 1 Match for the Impression score, otherwise we count zero.

On the other hand, in our proposed approach, we used an age-gender recognition model to classify this facial image into a proper group, let’s say Gen Z and Female. So because those information are equal, we count 1 Match for the Impression score. We select the open-source and lightweight age-gender recognition models, because we want those models to run on Jetson Nano.

So, we used the UTKFace dataset, which contains 40,000 face images with information of age, gender, as the test dataset.

**Slide 9:**

The results show that using the age-gender recognition model achieved the Impressions score at 0.334, while random approach only achieved at 0.082. This is a significant improvement in terms of Impressions score. Those images illustrate the output results from a selected frame, when running our application on Jetson Nano.

I hope this demonstrates the benefits of using ABAW in a practical use case.

**Slide 10:**

Let’s move to the second task: facial expression recognition using synthetic dataset. So facial expression recognition refers to the technique of classifying the expressions on face images from digital images or video frames into various categories such as anger, fear, surprise, sadness, happiness and so on.

**Slide 11:**

This table summarizes the most popular datasets for facial expression recognition tasks. You could see the number of images is not big, compared to 8 billion people in the world. And again, all concerns about real-life dataset are mentioned here. So in this task we use a synthetic dataset. There are several advantages of synthetic dataset. such as easily generated by a computer program, cost-effective, faster and easier to control generation conditions.

**Slide 12:**

So we use synthetic dataset from the Learning from Synthetic Dataset (LSD) competition in the 4th ABAW workshop. The LSD training dataset was used to train the models.

To evaluate model performance, f1-score, as the same metric from the competition, was used. LSD validation dataset was used as the synthetic test dataset, and a sample selected from the Multi-Task Learning competition was used as the real-life test dataset.

Those models were implemented on Jetson Nano to evaluate the inference speed. We want to use this measurement to select the optimal FER to use in the end-to-end video analytic application in the next task. We repeat 100 times the inference task on a random input tensor with shape (1,3,224,224), then measuring the average frame per second.

**Slide 13:**

We asserted the solutions used by top performers of the LSD competition. The first one is enet, the EfficientNetB0 which is recently state-of-the-art facial expression recognition model on AffectNet dataset. The second one is DAN, a transformer-based model, which is well-known for its advantage on generalization. The last one is GUS, a graph convolution used in the 2nd solution.

In our proposed solution, we try to combine those solutions to a single model to utilize their advantages. The first ensemble model is enet-DAN, which is the combination of enet and DAN. The second one is enet-GUS, which is the combination of enet and GUS. The last one is enet-DAN GUS, which is the combination of three single solutions.

**Slide 14:**

In general, the results show that our proposed solutions outperform any single model. More specifically, the enet-DAN GUS achieved the highest f1 score of 0.771 on the original validation dataset (LSD). The enet-DAN achieved the highest f1-score of 0.419 on the MTL dataset. We could see a degradation between model performance on synthetic test dataset and real-life test dataset. This highlights that we need to make the synthetic dataset more generalized to close this gap as much as possible.

About inference speed, in general, all models achieved a near real-time speed on Jetson Nano. The original GUS achieved the highest speed of 47.62 fps. The slowest speed is achieved by the ensemble enet-DAN GUS due to its complex architecture.

**Slide 15:**

The last task is to explore and evaluate the deployment process in a selected edge device (Jetson Nano). Jetson Nano is a low-power single board computer design for running AI applications. It has several advantages: power efficient, has AI capability with 128 GPU cores, runs Linux OS and supports many deep learning frameworks such as tensorflow, pytorch, opencv, onnx. Its price is low, and the most important thing is that we can run all required computations locally and keep data safe without transmission over the network.

However, it has some disadvantages. The first one is its limited resources. Only 4GB RAM, 4 CPU and GPU architecture is old. Its Linux version lacks many common libraries. Its Python version is old, hence limiting the ability to use the latest update from many deep learning frameworks.

**Slide 16:**

So I proposed a simple application architecture running on Jetson Nano. The first module is the video analytics module, which consists of a common webcam to capture frames of the environment, a face detector, a face tracking and reID, the age-gender recognition model and facial expression recognition model from previous tasks. All those computations were performed on Jetson Nano. Only final results such as average age, gender are stored for further application.

The second module is the back end module, consisting of a mongodb cloud server to keep the metadata of the latest frame, and a simple fastAPI server to retrieve this data for further use.

For simplicity, I created a simple front-end app with ReactJS, which retrieves this data and changes the video content based on age and gender information.

**Slide 17:**

This video demo records how the application works in practice. It demonstrates how to use age-gender recognition to get user information from a specific location, and play the contents in another device at different locations. The facial expression recognition could also allow us to calculate the satisfaction of users by capturing their emotional information.

However, the frame speed drops significantly when running in an end-to-end application. in case only one face appeared in the frame, the frame speed is only 12 fps. In case 4 faces appeared in the frame, the frame speed dropped under 5 fps.

This is due to the limitation of computation resources on Jetson Nano. However, I believe this frame speed still works for many applications. Like an ad in Youtube takes normally 6 seconds, a Tik Tok video takes a minimum 15 seconds. Then depending on the use case, it is still applicable.

**Slide 18:**

In conclusion, I have completed all tasks from the beginning. I demonstrate the benefit of using age-gender recognition could improve the Impressions score for the outdoor advertising case.

I explore the use of synthetic data and propose an ensemble approach to improve the performance of this task.

Finally, I proposed an end-to-end video analytics application running on jetson nano to deal with the data privacy concerns, as well as showcase the feasibility of applying affective behavior analysis in-the-wild methods by using edge devices.

**Slide 19:**

So, that’s all. Thanks for your attention.

My presentation consists of 4 parts: in the background I will go through a brief about the definition of facial expression recognition, its application and methodology. In the second part, experimental research, I will introduce about the dataset and the suggested solutions were used in my project. In the third part, experimental results will summarize the results of my research. And the last part, my conclusion about my project.

So facial expression recognition is the task of classifying the expressions on face images from digital images or video frames into various categories such as anger, fear, surprise, sadness, happiness and so on. FER can be used in a variety of applications, such as security, marketing, health care, robotics, and gaming. In security, facial expression recognition can be used to identify suspicious behavior such as potential terrorists. In marketing, facial expression recognition can help identify and target potential customers.

There are several datasets for facial expression recognition tasks, many of them are collected in laboratory condition, some are collected from movie or web. The data collection is a challenge for this facial expression recognition task. having sufficient labeled training data that include as many variations of the populations and environments as possible is important for the design of a deep expression recognition system.

So you can see that the real-life data almost, were collected from lab, movie and web environment, which has limited amount of available data and inconsistent labels. Because the data collection requires a complex condition, time and cost-consuming, synthetic data is a solution, that easily generated by a computer program with a lower cost, faster and easier to control and set up the conditions. And, synthetic data has become increasingly popular not only in facial expression recognition task, but also in many deep learning tasks due to its ability to generate more accurate, realistic, and diverse training datasets for neural networks.

So, this is the general flow of my experimental research. There are 3 main parts, data source selection, data pre-processing, training & evaluation.

For the first part, the datasets were received from the Learning from Synthetic Data (LSD) competition, organized in 4th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW). As you can see, for the LSD challenge, the organizer have selected some specific frames-images from the Aff-Wild2 database and manipulated to create the LSD datasets. In total, there are 277K images in training dataset and 4.7k in validation dataset. Because the test dataset was not public, I have sampled some images from the MTL learning datasets, and used as the test set to evaluate the facial expression recognition in my research. In LSD challenge, the performance measure is f1 score, and the baseline results on validation and test sets are 0.5 and 0.3.

So, this is some sample from the LSD and MTL dataset. You can see there are samples with blur, and the image size is only 112x112, which is quite small. This is the reason that I have this additional pre-processing process, to enhance the quality and upscale the images with a scale factor by 2. This is how the image quality changed after applying the super-resolution algorithm.

The third part, the most important part in my paper, I have use the LSD training dataset to train the models, then use the LSD validation set, and the MTL datasets to evaluate performance of the experiment models. To evaluate model performance, f1 score, accuracy score and inference time per image were used.

My suggested solutions for the facial expression recognition models could be categorized to “single model” and “ensemble model”. In “single model”, the first suggested model is a lightweight model in EfficientNet family, received from the hsemotion github repo, which is the state-of-the-art in AffectNet dataset. This is also the top performer in the LSD challenge. The second model is a transformer-based model from Distract you Attention Network paper, which was used in 3rd place solution in LSD challenge. This model is also the state-of-the-art model in AffectNet in 2021. The last single model is a graph-based model, Graph Embedded Uncertainty Suppressing, which was used in 2nd place solution in LSD challenge. In “ensemble model”, I tried different combination of the single model, such as combination of DAN and the state-of-the-art EfficientNet, combination of GUS and the EfficientNet, combination of DAN, GUS and EfficientNet, etc.,

In the experimental results, firstly let’s compare the f1 score of different models on the synthetic LSD validation dataset. The first observation is the single model could not achieve high score, and the ensemble models outperform any single models. The top performer is the ensemble of DAN, GUS and state-of-the-art EfficientNet, which accuracy score is 0.84 and f1 score is 0.77 on LSD validation set. The second top performer is the ensemble of GUS and state-of-the-art EfficientNet. In general, ensemble models with GUS and EfficientNet achieve higher score than others.

Now, let’s go to the results on MTL dataset. The first observation is that there is high gap between model performance between the MTL dataset and LSD dataset. No model has f1 score higher than 0.5. The highest accuracy score is only 0.58 and f1 score is only 0.49. This could be because the lack of generalization of the synthetic dataset, then the experiment models were overfitting on the synthetic dataset. However, similar result was observed, that the ensemble models still outperforms the single models. The ensemble model with GUS and EfficientNet still perform better than others.

Now, let’s go to the inference time. The first observation is that single model with the backbone resnet18 are the fastest runner, however, its accuracy and f1 score are not high compared to others. So this there is a trade-off between quality and speed. From remain models, we could see that lightweight model run faster than other ensemble models, and this is still the advantage of lightweight model in real-life application for faster inference, including it’s easier to deploy. And although the inference time of GUS and EfficientNet is higher than the single EfficientNet, but this gap is not too much and it could be a potential for the application of ensemble of the graph-based model and state-of-the-art EfficientNet in real-time use cases.

To conclusion, the main insight that the ensemble models outperform other single model, especially the combination of graph-based model and state-of-the-art lightweight model. It is potential to combine the single model with other complex model such as graph-based or transformer-based to achieve better performance in facial expression recognition task. A minor insight that although still have some limitation about generalization, using synthetic data can accelerate the development model for facial expression task.

Thank you for listening to my presentation on "Facial Expression Recognition using Synthesis Dataset". I hope you found it informative and insightful.

Hello and welcome to my term paper presentation on "Facial Expression Recognition using Synthesis Dataset". My supervisor is Professor Savchenko Andrey Vladimirovich.

My project aims to experiment different facial expression recognition models in a synthetic dataset to achieve better performance. I'll be discussing 4 parts:

* Background on facial expression recognition, its applications and methodology.
* Experimental research, including the dataset and solutions used in my project.
* Experimental results, summarizing the outcomes of my research.
* My conclusion on the project.

So, what is facial expression recognition? It is the task of categorizing facial expressions into emotions such as anger, fear, surprise, sadness, happiness, etc. from digital images or video frames. This technology has numerous applications, including security, marketing, healthcare, robotics, and gaming. For example, in security, facial expression recognition can be used to identify suspicious behavior.

This is the most popular datasets for facial expression recognition. You could see all of them were collected from the lab environment, movie or web.

Because collecting enough labeled training data that represents diverse populations and environments is a challenge with many constraints, and requires high effort and cost. This is where synthetic data comes in as a solution. Synthetic data is easily generated by a computer program, is cost-effective, faster and easier to control conditions, and has become increasingly popular in various deep learning tasks, including facial expression recognition.

In my research, I used datasets from the LSD competition, organized in the 4th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW). There were 277K images in the training dataset and 4.7K in the validation dataset. Because the test dataset was not public, I used images from the MTL learning dataset as the test set to evaluate the model performance. In this LSD competition, the performance was measured using the f1 score, with a baseline result of 0.5 on the validation set and 0.3 on the test set.

There are samples from the LSD and MTL datasets, and you could see the images are blur, not sharpen, and the input size is small. This is the reason that I performed an additional data pre-processing, to enhance the quality and upscale the images using a super-resolution algorithm. You can see how the quality of images change before and after applying the super-resolution algorithm.

I then trained the models using the LSD training dataset and evaluated performance in remain datasets using the f1 score, accuracy score, and inference time per image.

My proposal for facial expression recognition models includes "single model" and "ensemble model" approaches. The "single model" approach includes three models: a lightweight EfficientNet model, which is state-of-the-art on AffectNet dataset, a transformer-based model (DAN) from the Distract You Attention Network paper, and a graph-based model (GUS). Those models were used in the top performer’s solutions. The "ensemble model" approach combines different models, such as DAN and EfficientNet, GUS and EfficientNet, and DAN, GUS, and EfficientNet.

Experiment results on LSD validation set show that ensemble models outperform single models, with the best performance achieved by the ensemble of DAN, GUS, and EfficientNet, with accuracy score of 0.84 and f1 score of 0.77. The second best performer is the ensemble of GUS and EfficientNet.

On MTL dataset, no model had f1 score higher than 0.5, and the highest accuracy score was only 0.58 and f1 score was only 0.49. However, the ensemble models still performed better than single models, with the best performance achieved by the ensemble of GUS and EfficientNet.

In terms of inference time, the single model with a ResNet18 backbone was the fastest, but its accuracy and f1 score were lower than other models. The lightweight model was faster than the ensemble models. The inference time for GUS and EfficientNet was higher than for single EfficientNet, but the gap was not significant.

In conclusion, ensemble models outperform single models, and combining single models with more complex models, such as graph-based model, can improve performance in facial expression recognition but still keeping the low inference time. The second conclusion, although there are limitations such as generalization, using synthetic data can accelerate the model development process.

Thank you for listening to my presentation on "Facial Expression Recognition using Synthesis Dataset". I hope you found it informative and insightful.