WILL A BLIND MODEL HEAR BETTER? ADVANCED AUDIOVISUAL RECOGNITION SYSTEM WITH BRAIN-LIKE COMPENSATING AND GATING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal data (e.g., audio-visual inputs, various medical images) fusion neural networks has draw more attention recently with growing number of models and training techniques being proposed. Despite the success of the multi-modal fusion neural network, we find a interesting "low single-modality robustness" phenomenon. Specifically, a multi-modal trained model may achieve worse performance than single-modal trained model if another modal data are masked. This is like a born blind or deaf person (single-modal trained) surpass the healthy one (multi-modal trained) with only one modality data input, and the multi-modal experience becomes a bias causing negative transfer. It shows that the existing neural networks have lower robustness than the human brain in terms of modal-missing problem. To overcome the defect, in this paper we design a brain-like neural network modeling the processing of audio and visual signals by training it to perform audiovisual speech recognition tasks. Our results demonstrate the computational model's vulnerability to sensory deprivation while promoting this adaption can help in multi-modal processing. Besides, we propose modality mix and gated fusion techniques to get a more robust model with better generalization ability. We ask for more attention on the interaction of signals of different modalities and hope our work will inspire more researchers to study the cross-modal complementary.

1 INTRODUCTION

The artificial intelligence models aiming at human brain have made impressive progresses in many hard tasks. However, many researches have demonstrated the significant gap between the current AI models and human intelligence, such as catastrophic forgetting(Kirkpatrick et al., 2016), texture bias(Geirhos et al., 2019) and adversarial attacks(Goodfellow et al., 2015)(Kurakin et al., 2017)(Madry et al., 2018), and addressing these defects in the models can help us build more robust intelligent systems. Besides, more attention has been paid to the process of the multi-modal data, which is closer to the real-world situation but the heterogeneous data also bring much more difficulties.

As humane beings we naturally receive multi-modal signals, and the ability to combine and processing multiple modals ensure the superior performance in the tasks that require intra-modal information complementary, such as speech recognition task where lip movements help remove the noise in the voice data. More importantly, the human brains are robust enough in dealing with modal-missing problem. People with sensory deficiency can still live effectively in the multi-modal situation, after human-specific training (e.g., sign language for the deafs and braille for the blinds) or natural learning process (e.g., distinguishing the speakers with eyes covered).

Artificial neural networks have achieved great success in multi-modal fusion tasks. However, we find that the current multi-modal neural networks are far inferior to the brain in terms of modal-missing problem, which is called as "weak multi-modality robustness" phenomenon in this paper (Figure 1). Specifically, a model trained with multi-modal data may achieve poor performance if one modal data are masked, even worse than the same model trained with single-modal data. It is kind of like that a born blind or deaf person (trained with single-modal data) surpasses the healthy one (trained with multi-modal data) with only one modality data input, and the multi-modal experience

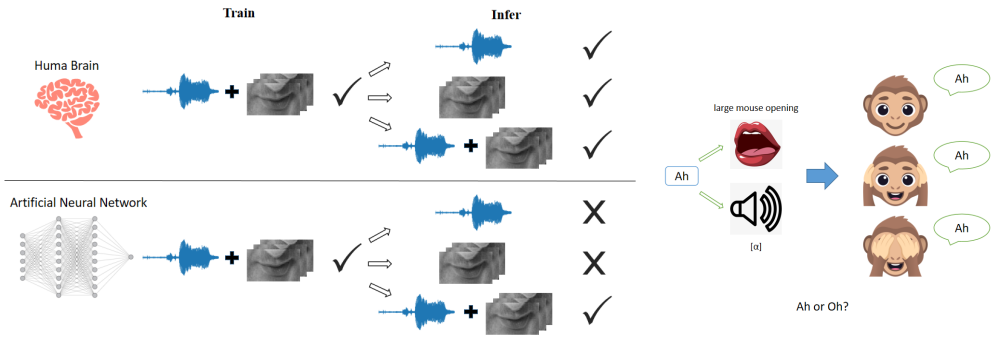


Figure 1: **Left:** Audio-visually trained brains can adapt to single-modal tasks while neural networks failed. **Right:** It's not hard for us to distinguish "Ah" from "Oh" with pure audio or visual signals, but for artificial neural networks, missing modality(ies) may result in the corruption of the whole system.

becomes a bias causing negative transfer. People and animals with temporal or eternal sensory injury can still survive in the multi-modal data, on the other hand, audiovisual complementary has been observed and studied by researchers, uncovering the robustness of cognition for biological neuron systems(Rauschecker, 1995)(Sadato et al., 2004)(Huber et al., 2020).

Rethink the recently proposed and advanced multi-modal neural networks, and less attention has been paid to the robustness of modal-missing problem. In this paper we try to take inspirations from human intelligent system and overcome the defects. We build a compact computation model formulating the process of multi-modal inputs in the brain to study the robustness of such system to missing modalities, and try to reproduce and study the audiovisual compensatory phenomenon discovered in human brains. It is discovered that a blind person usually have better hearing while deaf people have better sight and neuron science have also found the cross-modal plasticity in human brains(Merabet & Pascual-Leone, 2010). We conducted sensory deprivation experiments and do transfer learning on networks and found classically trained model failed to generate well facing modal missing, even with retraining and fine-tuning, corresponding to humans with sensory loss. We also propose to mix training data with uniformly dropout modalities and use a gated fusion to train a model that not only reaches highest accuracy in audio-visual word recognition tasks among our models and also is significantly more robust and can better generalize.

2 Related Works

2.1 Neural Networks Robust to Modal Missing

Multi-modal AI models can use complementary information from different modalities to form a better perspective and avoid the limitation of single view. Models capable of receiving different kinds of information has proved to steadily surpass traditional single ones with proper training techniques(Wang et al., 2020), and more and more multi-modal models and methods have been proposed recent years to tackle different problems in medical image analysis(van Tulder & de Bruijne, 2015), speech recognition(Petridis et al., 2018) and video comprehension(Sun et al., 2019).

However, in real world it's a common case that only data from a part of the whole modalities can be acquired to be fed to the model. An intuitive solution is to synthesis the needed modals using data from available modality(ies)(van Tulder & de Bruijne, 2015), which requires extra models and make the amount of generators increase rapidly as more types of modalities being used since possible modal missing situations grows. Other approaches include making information from different modals more similar to avoid strong bias(Chen et al., 2019)(van Tulder & de Bruijne, 2019), or to fuse only modal-invariant information by modifing the loss function(Chartsias et al., 2018) or using statistic information(Havaei et al., 2016), all of which highlight the importance of a model being robust to missing modals or sensory loss. Most of such research focus on medical imaging since the diversity of imaging techniques(i.e. PET, FLAIR, MRI) but we studied this problem in audio-visual speech recognition task.

2.2 AUDIOVISUAL COMPENSATORY

Humans have the ability to receive, process and combining different kinds of outside signals. The brain firstly process information of different modalities in the respective cortex separately, but complex human behaviors involve the interaction of multiple senses, such as watching videos, playing games, etc. The input from different sense organs can be regarded as data of different modalities, and the brain will eventually combine these data in more advanced brain areas (such as the prefrontal cortex) to form higher level of cognition. Human sensory and cognition system shows high robustness to sensory loss, especially for people with early or born sensory deprivation, most of which can be found compensatory phenomenon in their brains.

Compensation mechanism refers to the mechanism in which parts of the body partially replace the effects of the missing functions when other part(s) of the body is(are) lost due to damaging or disease. The compensation in this article specifically refers to the visual and auditory compensation of human brains, one aspect of which refers to the blind person's hearing becomes more acute with the loss of vision, so that in some cases they can obtain more information than normal people to compensate for the visual information. Another situation is that after deaf people lose their hearing, their vision will become sharper than ordinary people.

The compensating is probably due to that (for blind people) the original visual cortex has been converted to process auditory information, and better information utilization at higher level may also make contributions. For the former case, the biological explanation is the cross-modal plasticity of neurons that mainly refers to the adjustment of the functional connection of different brain areas of the brain(Merabet & Pascual-Leone, 2010)(Huber et al., 2020), thereby enhancing the response of deprived modal neurons to other modal data(Karnekull et al., 2016). This aspect has been confirmed by neuron science observation. The latter one refers to adaption at a higher cognition level. We believe AI researches can take inspiration from human brains' to build a more plastic and robust multi-modal model.

3 AUDIOVISUAL NEURAL MODEL TRAINING

3.1 MODEL STRUCTURE

In order to study audiovisual compensatory with neural networks, a neural network that accept inputs from two modalities is built, with separate primary processes. We adopt the structure of Petridis et al. (2018) and use two parallel processing modules to process visual and auditory information separately. The two modals' signals are only fused and processed near the output layer. The reason for this structure is that, on the one hand, each modal's processing in this structure is parallel, which is close to the modular structure of the partition between the different senses of the brain. On the other hand, unlike some other structures which import complex cross-modal interaction, in this model different modal data adopts a process that does not depend on other modal data, and is highly independent of each other, which avoids extra interference to sense deprivation and is convenient for observing the model's response to different modalities. The structure is visualized in Figure 2.

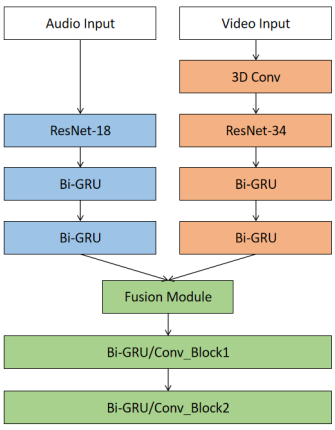


Figure 2: Structure of our multi-modal computational baseline model using two streaming primary processing combined with high-level joint processing.

For visual input, the model uses 3D convolution operation to fuse inter-frame information. Each frame of data obtained is input to a ResNet-34(He et al., 2016) with random initialization. After visual features are extracted, each frame feature becomes a vector, and two layers of bidirectional GRU modules with hidden dimension of 1024 are introduced to extract time series features.

For auditory input, due to the smaller size of each sample, the model is directly input to ResNet-18 to extract features, and then two layers of bidirectional GRU(Cho et al., 2014) with hidden dimension the same as that of visual inputs are used to extract sequence features between frames.

After the feature vectors of the two visual and auditory modalities are extracted, the vectors are fused as input using concat fusion or our proposed Gated Fusion, and two layers of bidirectional GRU modules or convolution blocks are used to obtain the final predictive recognition output.

3.2 Data Preparation

We choose the performance at audiovisual speech recognition task as our criteria for evaluation, and use Lip Reading in the Wild (LRW) dataset(Chung & Zisserman, 2016) to train our model's single-modal feature extractors while further training were conducted in both LRW and OuluVS dataset(Zhao et al., 2009). The reasons why we choose audiovisual task and LRW dataset to train our features are as follows: (1) The amount of data is huge. This data set collects a large amount of BBC program data, with 500 classes, each having up to 1000 video segments (1.16 seconds and 29 frames), which can provide enough data for training a successful and convinced model; (2) The data of each modal is complete. Some lip-reading data sets or video analysis data sets only provide video without audio, and can only learn pure visual tasks, and cannot perform multi-modal training; (3) Natural. The samples of LRW all come from the actual program, rather than being recorded separately. On the one hand, it ensures the diversity and representativeness of the data, on the other hand, it is more in line with the actual situation faced by the model and human beings; (4) The research in LRW is mature. As a well-known data set in the field of audiovisual recognition, the research on this data set is relatively mature, and the accuracy of speech recognition alone can reach more than 95% in most cases. Relative simple model with good enough performance made it convenient to conduct further analysis.

3.3 Preprocessing and Baseline Model Training

The training on LRW dataset followed the procedure of (Petridis et al., 2018), the video inputs are mouth regions of interest (ROI) cropped from each frame in the video, and the audio inputs are audio waveform directly extracted from the raw mp4 file, without extra process. Video inputs are augmented with random crop and horizontal flips. The structure already adapts to make LRW features have the same sizes and the fusion operations need the features from different modals have the same size, so for OuluVS we padded the inputs to the max size found in the dataset, same as the OuluVS_pad in Section 5.3. The complete training procedure includes pretraining in each modality and combined training with multi-modal part only and end-to-end finetuning. Details can be found in Appendix A.1.

The finally trained baseline model reaches accuracy over 97% in the validation set at audio-visual task for LRW dataset, which is similar to the results of paper (Petridis et al., 2018) and the slight drop may comes from not using the extra augmented material. For OuluVS we can reach nearly perfect performance with accuracy over 99%. To have a more comprehensive comparison, we also build models with CNN as multi-model part besides the baseline with Bi-GRU, and some of latter experiments are also conducted on CNN structure[1].

3.4 Born Disability and Acquired Disability Variants

We train lots of various models besides the baseline in the cascade experiments using different structure and input modalities. We name them with different types of disability because we use them to simulate the situation of variant kinds of sensory defects.

For a born blind model, the video inputs are replaced with zeroed-outed inputs (born deaf model is vice versa), we train such models' multi-modal parts 5 epochs. For an acquired blind model, while the input data same as the born blind, the training starts from the previously trained audio-visual model. The models are also trained 5 epochs in the multi-modal parts for a fair comparison with the ones with born-disability.

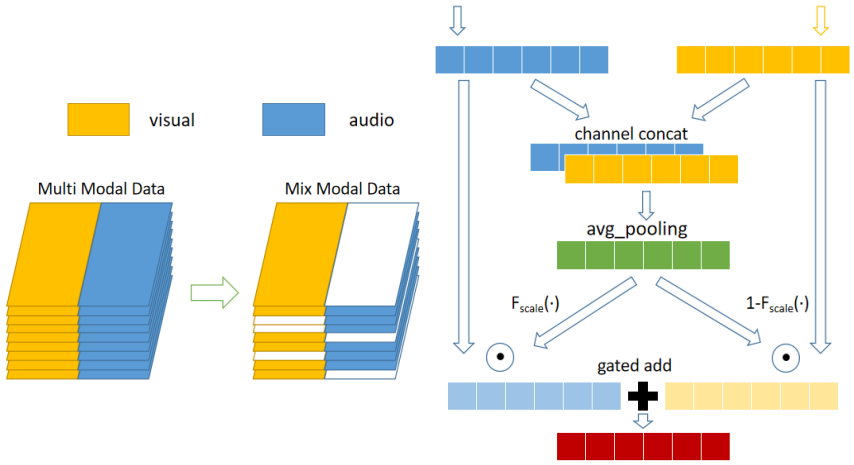
[1]A transformer-based model is also tested but proved to be hard to convergence in this classification task without extra pre-training, so we leave it for future work.

4 MODALITY MIX TRAINING AND GATED-FUSION MODULE

In the experiments, we found that baseline model structure failed to build good connections between modalities and often have two strong bias towards specific modality, which hurts the models' generalization and robustness to missing modality data. In order to tackle the disadvantages, we propose two methods: Modality Mix and Gated Fusion.

4.1 MODALITY MIX

Traditionally audio-visual model training research focus on complex cross-modal interaction and knowledge distillation(Ren et al., 2021), while a few introduced part of single modality as data augmentation(Chung et al., 2017). The work of Chung et al. (2017) introduce mixing only a small part of single-modal data in sentence-level lip-reading, in which a whole modality is eliminated, making it only adapt to attention model that fuse information in temporal dimension. In van Tulder & de Bruijne (2019) the authors proposed modality dropout as one of training techniques, but they use averaging-based fusion thus remove the corresponding modality in the fused features, and they chooses subset of modalities while our method generalized to use one modal only in situations with more modalities than two. As stated before, we take inspiration from the modality bias and propose Modality Mix technique which only need to zero-out the input from the deprived modality, and we use an uniformly mixed data from different cases of sensory loss instead of a majority multi-modal data.



(a) Modality Mix (b) Gated Fusion

Figure 3: **Left:** Samples conducted with Modality Mix where part of the training data have random modality being zeroed out, and all modal's data keep the same size. **Right:** Gated fusion Operation where each feature in the two modalities are added with computed gating weight from the averaged vector.

4.2 GATED FUSION

Usually a multi-modal structured network have special subset in it dealing with modal-specific data stream, and the way to combine the information of such sub-network is defined as fusion. The fusion techniques can be divided into early fusion, late fusion and intermediate fusion according to the place of the fusion module. Traditionally the modalities are treated equally in fusion before further processing in higher layer, and methods include concatenation, addition and dot-product. While these processing methods ignores the inherent difference of different modals, they often results in strong bias towards one more informative and plain modal(Michelsanti et al., 2021). We notice that neural experiments have proved that gating mechanism in the brain are developed to process information flow related to different subjects(Postle, 2005)(Gisiger & Boukadoum, 2011), enabling

the flexibility and quick adaption in various tasks(Monsell, 2003). Consistent with the complementary of modalities and the gating mechanism in human brain, and inspired by the SE-Net(Jie et al., 2017) in image classification, we design an attention-based Gated Fusion module for a more flexible combination of different modals. Similar structure has been proposed in audio-visual speech enhancement, in which only audio data are gated because the purpose of the network. (Chen et al., 2019) proposed a gated fusion method similar to one of the methods used in our experiment, but neither the case nor the purpose of the two researches are overlapped.

5 Experiment Results

5.1 Robustness to Sensory Deprivation

5.1.1 Multi-modal trained models are vulnerable due to incorrigible modality bias

We tested different models' vulnerability to different cases of modal missing in both LRW and OuluVS datasets. As in Section 3.4, the missing modal's input are zeroed out to stimulate sensory loss. The results can be viewed in Table 1. The A and B means the baseline model using concat fusion and that using gated fusion, while B1 and B2 means the model trained end-to-end and the model trained with features fixed. All results are tested on models with highest validation accuracy in the targeted modality(ies), which is audio-visual for baseline, audio-only for the blind and video-only for the deaf. We also studied models modified with our proposed gated-fusion model for a more complete analysis. It shows that for LRW dataset, our gated fusion module can enhance the models ability in audio processing while keeping competitive performance for multi-modal inputs, and in the OuluVS dataset which is much smaller and simpler such operation will not affect the performance much.

Table 1: Comparison of Robustness to Sensory Loss of Different Models: "A" means the baseline model with concat fusion. "B1" means models use gated fusion and "B2" is the same as "B1" models with the single-modal features keep fixed during training.

		LRW			OuluVS		
		AV	AO	VO	AV	AO	VO
Bi-GRU	A	97.87%	39.63%	**5.66%**	**99.50%**	99.00%	10.45%
	B1	97.78%	52.50%	4.53%	**99.50%**	**100.00%**	**16.42%**
	B2	**97.96%**	**68.38%**	2.58%	**99.50%**	93.03%	7.96%
CNN	A	**97.95%**	81.70%	3.78%	**100.00%**	**100.00%**	9.45%
	B1	97.79%	**87.43%**	**4.53%**	**100.00%**	99.50%	**11.94%**
	B2	97.78%	77.90%	1.60%	99.50%	98.51%	11.44%

The results shows that for both recurrent and convolutional models has the obvious problem of vulnerability to sensory loss, with convolution-based models seems to have more bias towards audio signals. As the effect of the gated-fusion model alone, it will improve the models' bias towards the more informative modal with little influence to the multi-modal performance, which is acceptable results since we propose such mechanism to be combined with modality mix to have a better understand of the relationship between modalities in the following transferring situation.

We further analysis the error patterns of the negative samples mistakenly classified by our model, and the examples of the findings can be found in Figure 4. Prediction results from baseline recurrent model with vision missing (above) and hearing missing (below). We can find that for the wrong false results shown in red, the wrong true results in green mostly focus near some specific labels, especially in deaf cases.

5.1.2 Using Modality Mix to enhance the robustness

As described in Section 4.1 and shown in Figure 4, the main mistakes made by audio-visual trained models are their irreformable bias towards specific modality, causing them take silence or black as information. We believe Modality Mix will help the model overcome such behavior and made up to a more robust model, and the training results are shown in Table 2. We can conclude that with

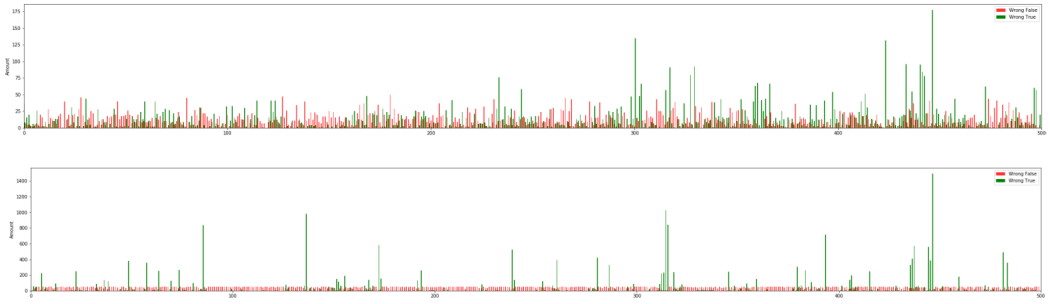


Figure 4: Analysis of negative samples for baseline model in modal missing scenarios. Every position in the x-axis refers to a class, with red bars indicates the number of samples that are mistakenly classified to other classes and green bars means samples from other classes are mistakenly classified to this class.

the Modality Mix the model would still keep the performance in the multi-modal scene but gain significant improvements in the robustness towards modality missing. The slight drop of 0.3% in the models using Gated Fusion and Modality Mix train end-to-end may results from the reduced hidden dimension since we replace concatenation with gated-adding operation, halving the fused features.

Table 2: Robustness to Sensory Loss For Models Trained With Modality Mix

		LRW			OuluVS		
		AV	AO	VO	AV	AO	VO
Concat Fusion	Baseline	97.87%	39.63%	5.66%	99.50%	99.00%	10.45%
	+ Modality Mix	97.85%	95.88%	80.56%	99.50%	100.00%	61.69%
Gated Fusion	Baseline	97.78%	52.50%	4.53%	99.5%	100.00%	16.42%
	+ Modality Mix	97.84%	95.76%	78.60%	99.0%	99.50%	46.27%
Gated Fusion (Feature Fixed)	Baseline	97.96%	68.38%	2.58%	99.5%	93.03%	7.96%
	+ Modality Mix	97.69%	95.69%	78.60%	99.0%	98.51%	18.41%

5.2 Generalization to Missing Modality

5.2.1 Born Disability Models Beat Acquired Disability Ones

For human beings, it's intuitive that people with born disabilities will adapt better than those acquired later, but a fashion in neural networks is conducting transfer learning as a way of training better models, which is also used in cross-modal situations, with SoundNet(Aytar et al., 2016) as an example. In order to test the generalization ability of the models to the sensory deprivation cases, we train and retrain variants with different methods as described in Section 3.4. At first we expect that the models seen multi-modal data shall benefit from the understanding of different models, as researches like Aytar et al. (2016) and Arandjelovic & Zisserman (2017) have shown some promotion. Surprisingly we found that in all kinds of multi-modal models we found usually a trained-from-scratch model outperform the transfer-learned one in the corresponding modality, even with the same procedure of training, which is confusing for a neural network but in consistence with human brains, since our daily experiences support that early-disabled people adapt better than the late-disabled ones, and early studies has also shown that people lost their sense early have better performance than those lost later in Minimum-Audible-Angle Discrimination(Voss et al., 2004), sound-source discrimination(Voss et al., 2008) and auditory episodic recognition(Karnekull et al., 2016) for the blind, and sign language for the blind(Lieberman et al., 2014). Another finding is that the CNN models are much harder to be retrained to fit missing modality scenes, and although such results may come from high learning rate at the beginning of re-training, acquired blind models having worse results in audio-only task even compared with multi-modal model as the start point of training is still out of our expectation.

Table 3: Generalization Ability For Models Multi-Modal Trained Compared With Born Disability

		AV	AO	VO
Bi-GRU	Baseline	**97.87**%	39.63%	5.66%
	Born Blind	89.42%	**94.12**%	0.20%
	Born Deaf	0.25%	0.21%	**76.80**%
	Acquired Blind	22.34%	67.28%	0.20%
	Acquired Deaf	34.36%	0.28%	68.14%
CNN	Baseline	**97.95**%	81.70%	3.78%
	Born Blind	72.76%	**93.09**%	0.18%
	Born Deaf	0.56%	0.20%	**76.41**%
	Acquired Blind	21.20%	67.76%	0.20%
	Acquired Deaf	1.59%	0.28%	68.14%

As in Table 3, models trained with born sensory loss have sometimes achieved 20 percentage of higher accuracy than those trained from an multi-modal model. On the other hand, there appeared a strange result that acquired deaf model beat the blind model which is to the contrary of the models with born disabilities, which is against our intuition that people can easily adapt to audio-only speech recognition while lipreading is much harder. We fix such problem and get better combined single-modal performance with our proposed gated fusion module in further experiments.

5.2.2 Gated Fusion and Modality Mix Experiments

We combined modality mix and gated fusion techniques so as to train a model that can better generalized to modal missing situations. Following the procedure of Section 5.2.1, we conducted similar transfer learning experiments in our models with modality mix and gated fusion. The results are shown in Table 4. Mind that the modality mix doesn't change network structure so the born disability models are the same as the baseline.

From the results we can see that the acquired blind models with gated fusion have much better accuracy than those using concat fusion with little drop in other modality, and the blind and deaf models have the same accuracy relationship as that of the born ones. Although still slightly lower compared to the performance of the born disability models, the modality mix trained models after transfer learning still have better performance than the models with acquired disabilities. Another intriguing discovery is that when we try to make a model trained with modality mix specialized to one modality, both the performance of the other modality and multi-modalities are dropped, the model often clearly fails to get better performance even in the specially trained modal's data (except on the acquired deaf with gated fusion, which still has competitive performance).

Table 4: Comparison of Generalization of Different Models

		AV	AO	VO
Concat Fusion	Baseline with Modality Mix	97.85%	95.88%	80.56%
	Born Blind	89.42%	94.12%	0.20%
	Born Deaf	0.25%	0.21%	76.80%
	Acquired Blind	22.34%	67.28%	0.20%
	Acquired Deaf	34.36%	0.28%	68.14%
	Acquired Blind (Modality Mix)	68.83%	67.60%	0.26%
	Acquired Deaf (Modality Mix)	46.96%	0.24%	77.58%
Gated Fusion	Baseline with Modality Mix	97.84%	95.76%	79.78%
	Born Blind	87.70%	94.32%	0.20%
	Born Deaf	0.30%	0.14%	77.30%
	Acquired Blind	85.82%	73.03%	0.20%
	Acquired Deaf	10.78%	0.23%	67.73%
	Acquired Blind (Modality Mix)	83.73%	91.28%	0.20%
	Acquired Deaf (Modality Mix)	29.39%	0.40%	77.27%

5.3 GENERALIZATION TO DIFFERENT DATASETS

Besides the generalization to missing modalities, another question is whether the proposed techniques can help learn task-agnostic features that can be better generalized to other datasets, without being specifically modified to consider out-of-distribution problem? In order to further evaluate the learned multi-modality features', we tried to retrain the classifier to help fit the model trained in LRW dataset to OuluVS dataset. These two datasets differ in recording condition, head poses and label space, which makes the transferring extraordinarily arduous.

Due to the lengths of records in the OuluVS dataset are not fixed, additional preprocessing are needed to apply our models. We tested our LRW trained models in two differently preprocessed OuluVS datasets, namely OuluVS_pad and OuluVS_resize, where the former one make all samples zero-padded to the max length in the dataset and the latter resized all samples to the same length as the LRW dataset, that is, 29 frames for video using linear interpolation and 19456 vector size for audio using resampling. The results are shown in Table 5. The results shows that combined the gated fusion and modality mix we proposed results in best generalization ability to new datasets. The experiment is preliminary yet and future works may have better analysis and complete comparisons.

Table 5: Generalizing multi-modal features from LRW to OuluVS

	OuluVS_pad	OuluVS_resize
Concat Fusion	21.39%	10.95%
Concat Fusion & Modality Mix	28.86%	12.44%
Gated Fusion	16.92%	7.46%
Gated Fusion & Modality Mix	**34.33**%	**16.42**%

6 CONCLUSION & FUTURE WORK

In this paper we address and analysis the problem that a traditional multi-modal artificial neural network dealing with audiovisual inputs failed to generalize well to corresponding single-modal tasks, which is similar to the common view that born disabilities enables better adaption than those acquired later. We take inspiration from human brains and suppose the lack of understanding the inherent correlation of different modalities in the artificial network being the key of the problem. We use modality mix techniques to enforce adaption of sensory deprivation, and try a clear gating mechanism to simulate the plasticity of neurons in audiovisual compensatory. The detailed experiments prove that modality mix and gated fusion module can help not only the robustness of the models when facing sensory deprivation, but also permit better generalization to single-modal tasks.

A bunch of future works remained for deeper exploration in the similarities and differences between multi-modal neural networks and biological intelligent systems. Firstly, we choose a simplified way of stimulating sensory loss, using zeroed-out specific modality to represent black view or silence, but more methods like regular noise, meaningless activation and unmatched inputs are also worth trial. Secondly, our current models are trained with speech recognition task in which audio signals have huge advantage in classification, and such bias may influence the fairness of our analysis, so we can try to train multi-task models by also introducing tasks where video signals dominate, like speaker recognition or video classification. Finally, we can cooperate with neural scientists to further explore the plasticity and flexibility of human brains to make our computational model more biologically reasonable, and we believe more advanced multi-modal models can be developed if we further investigate the gating and adapting mechanism developed by our brains.

REFERENCES

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018. doi: 10.1109/TPAMI.2018.2889052.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 609–617. IEEE

Computer Society, 2017. doi: 10.1109/ICCV.2017.73. URL https://doi.org/10.1109/ICCV.2017.73.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 892–900, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/7dcd340d84f762eba80aa538b0c527f7-Abstract.html.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman (eds.), *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A. Tsaftaris. Multimodal MR synthesis via modality-invariant latent representation. *IEEE Trans. Medical Imaging*, 37 (3):803–814, 2018. doi: 10.1109/TMI.2017.2764326. URL https://doi.org/10.1109/TMI.2017.2764326.

Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali R. Khan (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part III*, volume 11766 of *Lecture Notes in Computer Science*, pp. 447–456. Springer, 2019. doi: 10.1007/978-3-030-32248-9_50. URL https://doi.org/10.1007/978-3-030-32248-9_50.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1724–1734. ACL, 2014. doi: 10.3115/v1/d14-1179. URL https://doi.org/10.3115/v1/d14-1179.

Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pp. 87–103. Springer, 2016.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3444–3453, 2017. doi: 10.1109/CVPR.2017.367.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Thomas Gisiger and Mounir Boukadoum. Mechanisms gating the flow of information in the cortex: what they might look like and what their uses may be. *Frontiers in computational neuroscience*, 5:1, 2011.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In Sébastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gözde B. Ünal, and William Wells (eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*, volume 9901 of *Lecture Notes in Computer Science*, pp. 469–477, 2016. doi: 10.1007/978-3-319-46723-8_54. URL https://doi.org/10.1007/978-3-319-46723-8_54.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Elizabeth Huber, Kelly Chang, Ivan Alvarez, Aaron Hundle, Holly Bridge, and Ione Fine. Early blindness shapes cortical representations of auditory frequency within auditory cortex. *The Journal of Neuroscience*, 39(26):5143–5152, 2020. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2896-18.2019. URL http://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2896-18.2019.

H. Jie, S. Li, S. Gang, and S. Albanie. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2017.

Stina Cornell Karnekull, Artin Arshamian, Mats E. Nilsson, and Maria Larsson. From perception to metacognition: auditory and olfactory functions in early blind, late blind, and sighted individuals. *Frontiers in psychology*, 7:1450, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, and A. Grabska-Barwinska. Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci U S A*, 114(13):3521–3526, 2016.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJGU3Rodl.

Amy M Lieberman, Arielle Borovsky, Marla Hatrak, and Rachel I Mayberry. Real-time processing of asl signs: Effects of linguistic experience and proficiency. In *Proceedings of the 38th Boston University conference on language development*, pp. 279–291. Cascadilla Press Somerville, MA, 2014.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Lotfi B Merabet and Alvaro Pascual-Leone. Neural reorganization following sensory loss: the opportunity of change. *Nature Reviews Neuroscience*, 11(1):44–52, 2010.

Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396, 2021. doi: 10.1109/TASLP.2021.3066303.

S. Monsell. Task switching. *Trends in Cognitive Sciences*, 7(3):134–140, 2003.

Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pp. 6548–6552. IEEE, 2018. doi: 10.1109/ICASSP.2018.8461326. URL https://doi.org/10.1109/ICASSP.2018.8461326.

B. R. Postle. Delay-period activity in the prefrontal cortex: One function is sensory gating. *j cogn neurosci*, 17(11):1679–1690, 2005.

J. P. Rauschecker. Compensatory plasticity and sensory substitution in the cerebral cortex. *Trends in Neurosciences*, 18(1):36–43, 1995.

Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 13325–13333. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Ren_Learning_From_the_Master_Distilling_Cross-Modal_Advanced_Knowledge_for_Lip_CVPR_2021_paper.html.

N. Sadato, H. Yamada, T. Okada, M. Yoshida, T. Hasegawa, K. I. Matsuki, Y. Yonekura, and H. Itoh. Age-dependent plasticity in the superior temporal sulcus in deaf humans: a functional mri study. *Bmc Neuroscience*, 5(1):1–6, 2004.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 7463–7472. IEEE, 2019. doi: 10.1109/ICCV.2019.00756. URL https://doi.org/10.1109/ICCV.2019.00756.

Gijs van Tulder and Marleen de Bruijne. Why does synthesized data improve multi-sequence classification? In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5-9, 2015, Proceedings, Part I*, volume 9349 of *Lecture Notes in Computer Science*, pp. 531–538. Springer, 2015. doi: 10.1007/978-3-319-24553-9_65. URL https://doi.org/10.1007/978-3-319-24553-9_65.

Gijs van Tulder and Marleen de Bruijne. Learning cross-modality representations from multi-modal images. *IEEE Trans. Medical Imaging*, 38(2):638–648, 2019. doi: 10.1109/TMI.2018.2868977. URL https://doi.org/10.1109/TMI.2018.2868977.

P. Voss, M. Lassonde, F. Gougoux, M. Fortin, J. P. Guillemot, and F. Lepore. Early- and late-onset blind individuals show supra-normal auditory abilities in far-space. *Current Biology*, 14(19): 1734–1738, 2004.

P. Voss, F. Gougoux, R. J. Zatorre, M. Lassonde, and F. Lepore. Differential occipital responses in early- and late-blind individuals during a sound-source discrimination task. *Neuroimage*, 40(2): 746–758, 2008.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 12692–12702. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01271. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_What_Makes_Training_Multi-Modal_Classification_Networks_Hard_CVPR_2020_paper.html.

Guoying Zhao, Mark Barnard, and Matti Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265, 2009.

A APPENDIX

A.1 DETAILED TRAINING PROCEDURE AND HYPERPARAMETERS

We train models using Bi-GRU as the layers processing the fused information in the detailed analysis, but we also train CNN models for observation to prove that the phenomenons aren't on account of the special types of network. The structure of our CNN model is modified from that used in the pretraining of single-modal features.

Firstly the audio stream and video stream are trained separately with the same classification task, with a convolutional-based final two CNN blocks (which are also used in our CNN baseline) instead of Bi-GRU for 30 epoches to train the former layers. Then GRUs are introduced and trained 5 epochs seperately and the whole model is trained end-to-end for 30 epochs.

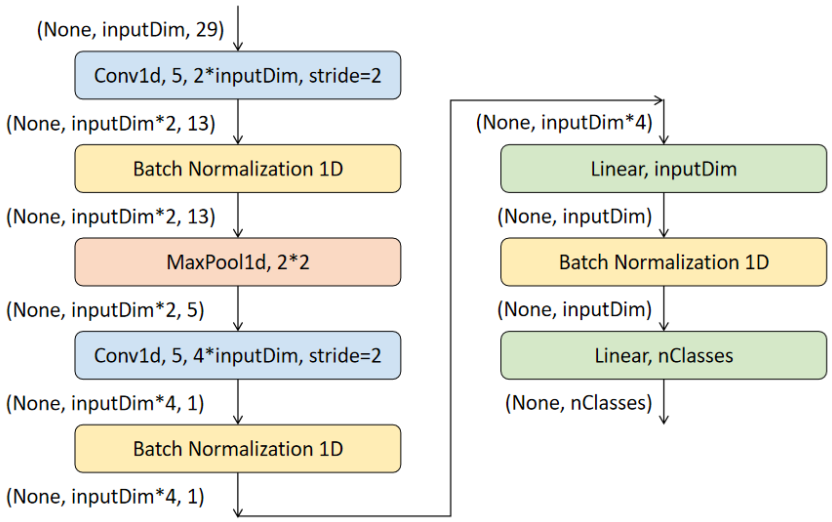


Figure 5: The left and right groups of modules are the two blocks used in the training of single-modal features and our CNN baseline. The LRW dataset has 29 frames for each sample. If more frames are needed such as samples in the OuluVS_pad dataset, the output of the first block will be averaged in the last dimension.

After models for each modality has been trained, the output features from each modalities' Bi-GRU module are concatenated and fed into another two layers of Bi-GRU modules, which will then be trained separately for 5 epochs. The whole network are finally trained end-to-end together for 30 epochs to get the final model. In practice we found that using models without GRU training in the single-modal task resulted better performance in the final multi-modal task and use this trick in all of our following model except those with special notifications.

In the whole processes we use Adam optimizer(Kingma & Ba, 2015) with the initial learning rate of 0.0003 that decays exponentially to its half every 5 epochs. Other Detailed processes can be found in (Petridis et al., 2018) with negligible difference in data augmentation and learning rate settings.

As for the OuluVS dataset, we use the same structures and features pretrained in the LRW dataset before finetuning in the multi-modal situation, and the following processes are the same as those in the LRW dataset.

A.2 ADDITIONAL EXPERIMENT ON SENTENCE-LEVEL AUDIOVISUAL SPEECH RECOGNITION

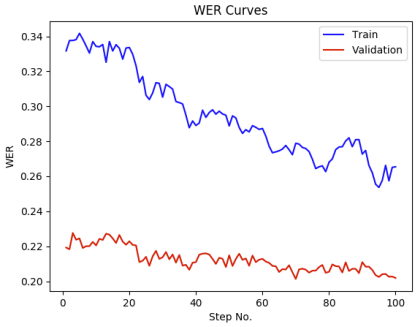
For a further analysis of the multi-modal neural networks doing more complex tasks, we conduct preliminary experiments on the model designed and trained for Lip Reading Sentences in the Wild (LRS2) dataset(Chung et al., 2017). Due to the limit of computing resources we use the two-stream model and the trained weights of Afouras et al. (2018) for sensory deprivation experiments and

did simple transfer learning in which the whole models were trained end-to-end or with feature from each single modal fixed. The pretrained weights are trained with curriculum learning techniques(Bengio et al., 2009) and part of the data also had one modality dropped, so the conditions are more like our experiment in Table 4.

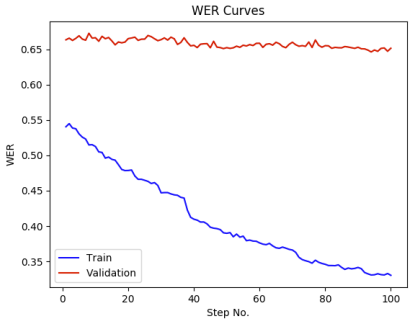
The results are shown in Table 6 and the values are Word Error Rate (WER) with lower value means better performance. We can see that most transfered models failed to reach the performance even with their initial weights. The models were not simply overfitting since the training curves indicating clear improvements with training, as shown in Figure 6.

Table 6: Sensory Deprivation Experiment on LRS2 models

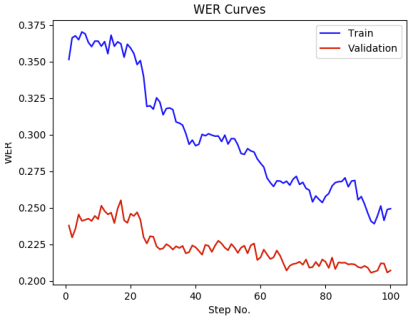
AV	AO	VO	
Original Trained Model	**10.8**%	**12.9**%	**56.3**%
Acquired Blind	12.5%	**12.9**%	59.7%
Acquired Deaf	13.7%	16.3%	56.6%
Acquired Blind (Feature Fixed)	11.8%	13.3%	60.1%
Acquired Deaf (Feature Fixed)	17.0%	15.9%	57.0%



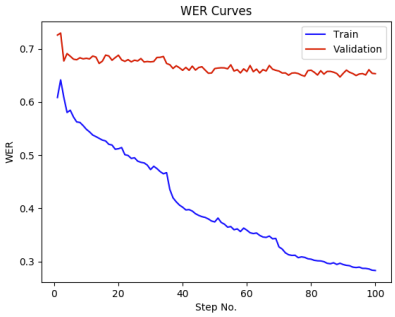
(a) Acquired Blind



(b) Acquired Deaf



(c) Acquired Blind (features fixed)



(d) Acquired Deaf (features fixed)

Figure 6: Training curves of the different models, all of which have clear descendent of loss in the training.

While the results are partly consistent with our experiments in the LRW dataset, we should also point out that the models introduce pretrained visual frontend and language model for the generation of final sentences so the conditions are quite complex, which disturbs our analysis, so the results are not precisely showing the real cases, but enough for an overview of the question.