AUTOMATIC PAIN ASSESSMENT IN FETUSES THROUGH TRANSFER LEARNING

THIAGO MELO DE OLIVEIRA

AUTOMATIC PAIN ASSESSMENT IN FETUSES THROUGH TRANSFER LEARNING

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Nivio Ziviani
Coorientador: Adriano Veloso

Belo Horizonte
Março de 2020

THIAGO MELO DE OLIVEIRA

AUTOMATIC PAIN ASSESSMENT IN FETUSES THROUGH TRANSFER LEARNING

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Nivio Ziviani
Co-Advisor: Adriano Veloso

Belo Horizonte
March 2020

Oliveira, Thiago Melo de

D1234p Automatic pain assessment in fetuses through transfer learning / Thiago Melo de Oliveira. — Belo Horizonte, 2020
xxii, 42 f. : il. ; 29cm

Dissertação (mestrado) — Federal University of Minas Gerais
Orientador: Nivio Ziviani

1. Insert Keywords Here. I. Título.

CDU 519.6*82.10

[Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

Se a imagem da folha de aprovação precisar ser ajustada, use:
`approval=[`*ajuste*`][`*escala*`]{`*nome do arquivo*`}`
onde *ajuste* é uma distância para deslocar a imagem para baixo
e *escala* é um fator de escala para a imagem. Por exemplo:
`approval=[-2cm][0.9]{`*nome do arquivo*`}`
desloca a imagem 2cm para cima e a escala em 90%.

I dedicate this work to everyone that was on my side during this journey, especially my fiancee Maria, my mother Suely, and my sisters Cínthia and Magda.

Acknowledgments

I thank my mother Suely and my family for all the support they have given me and for always believing in my choices. I also thank my fiancee Maria, for being available in all the moments I needed the most. Without them, this journey would not be possible.

I thank my friends at Kunumi and the company itself for creating this rich environment, which made me a better professional and a better person. I also thank the Fetal Pain Study Group from USP, who made this research feasible.

Finally, I thank my advisors Nivio and Adriano, for the trust and for the opportunity given when this was just a dream.

"Imagination will often carry us to worlds that never were.
But without it we go nowhere."
(Carl Sagan)

Abstract

Prolonged exposure to pain circumstances can have many side-effects on the life of a fetus and cause negative developmental consequences. Thus, pain assessment and management is made necessary to identify these scenarios early on. Even though numerous pain scales exist to help assess pain in neonates, until recently, no such method existed for detecting pain in fetuses. Based on these scales, some research has been developed to automatically assess pain through the means of analyzing images with computational help. Still, no such work had been developed for fetuses as well.

In this scenario, we propose the use of deep convolutional neural networks to construct a learning model capable of automatically detecting the presence of pain in fetuses. We do so through the evaluation of their facial expressions in images collected from 4-D ultrasound machines. By taking advantage of transfer learning, we used a network pre-trained on the task of face recognition, and confirmed that transferring from a similar task performed better than if made from a general-purpose dataset.

We have evaluated our model on images extracted from 13 video recordings of fetuses undergoing painful and non-painful stimulus and achieved an accuracy of 84.8% on the task of discriminating images of pain from those in a non-painful control group. Our results demonstrate the effectiveness of applying such methods with fetal images, and above all, show that it is possible to develop a model for automatically detecting pain in fetuses.

Palavras-chave: Deep Learning, Transfer Learning, Fetal Pain, Automatic Pain Assessment.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The International Association for the Study of Pain (IASP) defines pain as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" [Merskey and Bogduk, 1994]. The definition accompanying notes also establishes that "the inability to communicate verbally does not negate the possibility that an individual is experiencing pain and is in need of appropriate pain-relieving treatment."

As newborns are unable to self-report pain, its diagnosis is much harder when compared to adults. Thus, specialists have used non-verbal responses like facial expressions, crying sounds, and movements, alongside physiological measurements for better pain assessment. These methods have been tested and found to be reliable indicators of pain. Several observational scales have been published and verified based on them, such as the Neonatal Infant Pain Scale (NIPS) [Lawrence et al., 1993], and the Neonatal Facial Coding System (NFCS) [Grunau et al., 1998].

In the case of fetuses we have even more restrict methods of pain assessment. Fortunately, some recent studies have shown the feasibility of applying these pain scales on a fetus through the use of 4-D (four-dimensional) ultrasound images [Bernardes et al., 2018]. This process allowed the monitoring of facial expressions on a fetus while they were exposed to noxious stimuli like an anesthetic puncture.

With recent advances in Artificial Intelligence (AI), the capacity of machines to detect patterns in images has largely improved, which consequently allowed for its application in diverse scenarios. Hence, these techniques could also be useful for pain assessment by helping matching patterns of facial expressions that are common indicators of pain.

1.1 Motivation

With studies showing that fetuses beyond a certain week may also experience pain [Derbyshire, 2006; Derbyshire and Bockmann, 2020], early identification of this discomfort can be valuable in many situations.

One example is intrauterine surgery, which may be of significant benefit in the future development and survival of the fetus. Early correction, prior to birth, of congenital problems, will likely increase the odds of a healthy baby. These procedures, however, are quite invasive to the fetus and could eventually cause harm. The assessment of pain during the intrauterine life of a fetus is, therefore, a task with the potential of bringing significant improvements to fetus life quality.

Another critical topic is abortion. In the United States, a 2016 law from the state of Utah determines that women seeking abortion 20 weeks or more into pregnancy will first have to be given anesthesia or painkillers [Healy, 2016]. This procedure is intended not for them but the fetus.

This topic involves much ethical debate, as the exact week when a fetus starts experiencing pain nor if they feel pain at all is well defined [Lee et al., 2005]. At the same time, abortion is only legal until a particular week [Derbyshire, 2006]. So discussion arises not only if the fetus can or can not experience pain, but also as it may be the case that fetuses can only experience pain after weeks in which abortion is no longer possible. Evidence on the presence of pain in this scenario would be a significant contribution in such a delicate situation and may assist the decision by the doctors and the mother.

For both scenarios, the current standard for assessing pain in infants and fetuses relies on caregivers' observation of specific behaviors such as facial expressions. However, these observations are subject to bias and can be affected by several factors, such as identity, background, culture, and gender, which may lead to inconsistent assessment and treatment of pain. An impartial perspective during the pain assessment process could bring a more realistic and deterministic view on the subject. Hence, computational help would be of great use in finding evidence of pain and in effectively managing it.

1.2 Thesis Statement

We developed a learning model capable of automatically detecting the presence of pain in fetuses through the evaluation of their facial expressions in images collected from 4-D ultrasound machines. We have used modern deep learning techniques like transfer

learning and data augmentation to find the best model. Our results demonstrated the effectiveness of applying such methods to the assessment of pain in fetuses and, to the best of our knowledge, is the first work attempting to do so.

1.3 Contributions

As fetuses have been shown to respond to stimuli like anesthesia with facial expressions indicating pain, the goal of this work is to help automate the pain assessment process and to generate unbiased evidence of pain. We have developed a process capable of detecting the presence of pain from images collected from 4-D ultrasound machines.

If this system is eventually integrated into the ultrasound machine itself, it will bring many benefits, such as the monitoring of anesthetic procedures efficacy, much like what it is done in adults. As an example, if the surgeon detects that after the first anesthesia, the fetus still shows signs of pain, he will be able to make better decisions and apply another one if necessary.

This work also opens the way to explore the evolution of pain-related facial responses during fetal development. Considering that after the 20th gestational week, fetuses start to develop brain structures capable of showing signs of pain, this model would, therefore, allow for continuous monitoring of pain across time.

In summary, our main contributions are:

- We have created a systematic procedure for collecting and processing images of a fetus from videos of 4-D ultrasound machines. This procedure is also capable of detecting their facial landmarks. From this procedure, we have created a labeled database consisting of 226 images of 13 fetuses with facial expressions while in the manifestation of pain and two other control conditions. To the best of our knowledge, no such database existed.

- We have developed a learning model capable of detecting the presence of pain indicators from images of a fetus's face. We believe this is good evidence towards an unbiased pain assessment process. This novel approach has the potential to improve the pain assessment process significantly on fetuses. It would facilitate pain management by the doctors and caregivers and could even be the first indicator of discomfort or distress, leading to earlier intervention if necessary.

- We have shown that transfer learning with a network pre-trained with the face recognition task transfers well to fetus images even though the domain is different. We have achieved an accuracy of 84.8% on the classification task between images

of pain and a non-painful control group. If we look only at the images from acute pain videos, and classify them between the images before the stimulus (at rest), and after it (acute pain), we were able to achieve an AUC of 0.923.

1.4 Organization

The rest of this dissertation is structured as follows. First, Chapter 2 discusses related work in pain assessment. Chapter 3 introduces some background concepts on deep learning, necessary to further understand our work. Chapter 4 describes the fetal pain assessment study, and also introduces our dataset. Chapter 5 follows with our methodology, including our learning model. Then, Chapter 6 describes our validation process, as well as our experimental results. Chapter 7 concludes the dissertation and present future work possibilities.

Chapter 2

Related Work

In this chapter, we present the most relevant research results that guided our work, exposing the methodologies used by the authors and how they correlate to ours. Research on automatic pain assessment has a significant intersection between the medical field and applied machine learning both for fetuses and infants [Bellieni, 2012; Zamzami et al., 2016b]. Thus, we start by exploring the signs of pain in infants and fetuses, their main indicators, the available pain scales to measure them, and automatic methods of assessment. We conclude the chapter by explaining how our work is different from the mentioned ones, as we explore automatic pain assessment from a novel perspective with fetuses.

2.1 Pain

Pain is a universal form of distress present in humans and some other animals. Acute and chronic pain is very typical in the population and constitutes widespread public health problems [Goldberg and McGee, 2011]. Its prolonged presence could cause many adverse consequences, including psychological effects, which is especially true in the case of neonates and fetuses.

The study of neonatal pain appears to have begun as early as the 1870s when Dr. Flechsig proposed it was unlikely that neonates could feel pain because their neuronal myelination was not complete [Cope, 1998]. Charles Darwin's book written a couple of years later agreed with this view, as he wrote that "infant's pain expressions were related to reflexes only" [Darwin et al., 1872]. Even in the 1950s, some pediatric surgeries were still performed without analgesia and anesthesia [Cope, 1998].

It was only in the early 1980s that the first fetal surgery was performed by Dr. Michael Harrison [Harrison et al., 1982]. The fetus to be operated had a blockage in

the urinary tract that caused the kidney to dangerously extend, which is a condition known as congenital hydronephrosis. A vesicostomy was conducted to correct this issue by placing a catheter in the fetus to allow the urine to be released normally.

Further progress has been made in the years since the first operation, as advances in imaging technology and in surgery techniques allowed additional abnormalities to be treated and for less invasive forms of fetal surgical intervention to be performed.

Even though the cases in which fetal surgery is necessary are relatively rare, it has become the standard form of intervention in some abnormalities like myelomeningocele, as shown by the Management of Myelomeningocele (MOMS) trial [Adzick et al., 2011]. The study compared outcomes of in utero repair (before birth) with standard postnatal repair (after birth). The conclusion was that prenatal repair might result in better neurological function than repair deferred until after delivery.

A follow-up cohort study by the same study group has evaluated children originally enrolled in the MOMS trial, who are now in their school age (5.9 – 10.3 years old) [Houtrow et al., 2020]. They discovered that, even though there were no significant differences in adaptive behavior, motor function and quality of life were significantly better in the group with prenatal repair.

The study of fetal surgery is tightly coupled with the one of fetal pain. Fisk et al. [2001], for instance, evaluated the effect of opioid analgesia on fetal hormonal stress responses to intrauterine needling and showed that fentanyl, an opioid commonly used for pain medication and anesthesia, does attenuate the fetal stress responses.

Later on, van de Velde and Buck [2012] studied fetal reactions to painful stimuli and showed that painful interventions could have long-term effects on them. This study also concluded that adequate pain relief during potentially painful procedures is recommended, as it leads not only to better fetus well-being, but also helps with fetus immobilization, which prevents accidental fetal movements complicating these procedures.

These conclusions also affect anesthesiologists, as Devoto et al. [2017] shows that fetal pain is among their primary concerns during fetal surgery for myelomeningocele. Considering these procedures are so delicate, fetal pain assessment and management are of fundamental help, which leads research to the study of pain detection through the use of pain indicators.

2.1.1 Pain Indicators

Fetuses and infants can produce different signals of pain, which can be decoded to both identify its presence and to measure its level. These signals come from a variety

of sources, such as facial expressions, crying sounds, body movements, physiological indicators, and biological markers [Bellieni, 2012].

Even though we have this many indicators, pain identification is a challenging task as we have the manifestation of the same indicators present in similar feelings, such as anger, hunger, or stress. The recommendation to address this issue is that these indicators should be used in combination with each other [Bellieni, 2012] because most of the time, their presence alone is not sufficient.

Crying, for instance, can also be generated by hunger or anger which implies it can not be used as a sole indicator of pain. Thus, pain scales normally combine features of crying with other indicators for pain assessment. Fetuses have been shown to express a homolog of crying [Gingras, 2005], which can also be further explored for automatic pain assessment, as shown by Salekin et al. [2019b].

Physiological indicators, on the other hand, have the limitation that they are subject to variations due to underlying illness [Sweet and McGrath, 1998]. Body movements have also been pointed out to be indicators of pain, as fetuses already present withdrawn reflexes during stressful procedures [Zimmermann, 1991], but care must be taken as they can also be misleading as other factors may cause the movements.

Studies have also shown that biological markers like stress hormones (cortisol, adrenaline, and beta-endorphins) are increased in concentration in the blood in the presence of pain [Giannakoulopoulos et al., 1994]. However, the problem of these indicators is that they depend on results from laboratory tests, which makes it unfeasible to use during clinical procedures.

One of the most relevant indicators of pain, not only in adults but also in neonates and fetuses are facial expressions. As suggested by Yan et al. [2006], a great way of evaluating fetal facial expressions is through the means of 4-D sonography, and as he points out, these studies may be the key to predicting fetal brain function and well-being. Later research by Reissland et al. [2011, 2013] also suggests that, when healthy fetuses mature from 24 to 26 weeks of gestation, their capability of showing complex facial movements increases, and they were even able to observe facial expressions which resemble a face while in pain or distress.

Facial expression indicators are frequently present in pain scales. Several facial movements are usually tracked, such as brow bulge, eye squeeze, nasolabial furrow, and open mouth. The assessment of these manifestations is done by observers, which use pain scales to identify its presence and intensity of pain.

2.1.2 Pain Scales

Multidimensional neonatal pain scales were developed using the many indicators mentioned in the previous section. These scales are used by caregivers to assess pain with behavioral and physiological indicators. The most popular ones are:

- Neonatal Infant Pain Scale (NIPS), by Lawrence et al. [1993]
- Face, Legs, Activity, Crying and Consolability (FLACC), by Merkel et al. [1996]
- Neonatal Facial Coding System (NFCS), by Grunau et al. [1998]

These scales were all developed targeting neonates, as some indicator observations are not easily measured in a fetus due to the difficulties of collecting images of their faces. However, recent studies by Bernardes et al. [2018] have reported that the use of the NFCS is feasible to detect pain-related facial expressions compared with control conditions in a randomized and blinded assessment report. We further discuss this study in Chapter 4.

Nevertheless, the scales also have some limitations, as they are highly dependent on the observant's bias, require specific training for proper utilization, and are not able to monitor pain in a continuous manner. Thus developing tools that are capable of addressing this task automatically and continuously is highly compelling as they can result in a more consistent pain assessment.

2.2 Learning Models for Pain Assessment

The development of learning models to automatically assess pain has been a popular topic of research lately. As an example, for adults pain assessment, [Mauricio et al., 2019] has achieved remarkable results in identifying spatiotemporal features extracted from video sequences for pain recognition.

In terms of neonates, the first work attempting to assess pain automatically emerged in 2006 with the development of the iCOPE database [Brahnam et al., 2006]. The database consisted of 204 images from 26 infants. The images were collected in five different conditions, a resting baseline, with bodily disturbance, with an air stimulus on the nose, with friction on the external surface of the heel, and with the pain of a heel stick. The idea behind using this variety of conditions was to make sure the set of images were representative of the many possible situations, but also challenging enough to discriminate. The five conditions were later divided into two groups for classification: pain and non-pain.

Their studies considered the best scenario where the system would be able to train on a fetus and evaluate that same fetus later on. However, as permanence in the baby nursery is quite short, they also experimented with the case where this was not possible, thus developing a validation process where the classifier had to be trained beforehand and evaluated on images of a new infant, which was not in the dataset distribution previously. Given they had a small number of subjects, it was feasible to use a validation strategy known as leave-one-out, which consisted of iterating over every combination of using 25 subjects for training and 1 for testing.

At the time, neural networks were not as popular and advanced as they are today. Thus, the first attempts to automatically detect pain used traditional algorithms such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Support Vector Machines (SVM) [Brahnam et al., 2005, 2006, 2008]. In order to extract features from the images, these studies relied heavily on texture descriptors like Local Binary Patterns (LBP) and a few other variations Nanni et al. [2010]. Still, the results were satisfactory, and the experimental process they developed is similar to what is used today with more modern techniques.

Zamzami et al. [2016a], more recently, proposes a multimodal approach for automatic pain assessment, which combines a few indicators like facial expressions, body movements, and changes in vital signs for producing a pain score. This score was evaluated on videos of 18 infants recorded at the Neonatal Intensive Care Unit (NICU) of the Tampa General Hospital, and the multimodal approach performed much better than each individual indicator on its own, by achieving 95% accuracy for pain classification. A later cohort study also included crying sounds into the model and yielded an accuracy of 96.6% [Zamzmi et al., 2017].

Modern approaches in automatic pain assessment of neonates make the use of deep neural networks to achieve state of the art results. Zamzami et al. [2018], for instance, has significantly improved the results in the aforementioned iCOPE database, achieving 0.948 of AUC by combining both handcrafted features and features extracted by Convolutional Neural Networks (CNNs) using transfer learning.

Zamzmi et al. [2019], in a similar study, has compared the use of popular CNNs, such as VGG and ResNet, with a network particularly tailored for neonatal pain assessment, the N-CNN. To evaluate the networks, they have also recorded videos at the NICU of the Tampa General Hospital, where a camera was installed on a stand next to the neonate's incubator, targeting their faces. These recordings resulted in the Neonatal Pain Assessment Dataset (NPAD), which consists of 31 neonates and are available

at the study's website [1]. Their findings suggest that automatic recognition of neonatal pain using these networks is not only viable but also a more efficient alternative to the current standard of pain assessment, which relies on caregivers.

New studies, like Fotiadou et al. [2014], have also been using videos directly as inputs to the learning models. Salekin et al. [2019a], for instance, has proposed a multi-channel shared network to classify pain from videos by extracting features from facial expressions and body movements. Likewise, [Zhi et al., 2018] proposes the use of dynamic facial texture features and dynamic geometric features to extract features from video sequences and use them to classify facial expressions of infants as pain or no pain. These techniques take advantage of the spatiotemporal sequence of the frames, to create a time series of frame-level features. [Salekin et al., 2019b] also contributed with an alternative approach by using convolutional neural networks to extract features from crying sounds.

Our work has adopted a similar approach of not having only images of rest and pain, but also a third set of images from another stimulus, which in our case was a vibro-acoustic sound with the intention of causing discomfort, but no pain. As for the validation strategy, we have also adopted the leave-one-out method used by Brahnam et al. [2006]. Like Zamzami et al. [2018], we relied heavily on the use of transfer learning to train our convolutional neural networks, given we have limited data available. For the same reason, we have also used data augmentation, which was very common in many of the mentioned studies due to data limitation. To the best of our knowledge, our work is the first attempt to detect pain in fetuses automatically, and we believe many of the approaches mentioned earlier have the potential to help in this task.

[1] `https://rpal.cse.usf.edu/project_neonatal_pain`

Chapter 3

Background on Deep Learning

In the task of image classification, traditional machine learning algorithms required hand-engineered features, like filters and descriptors, which were meant to extract information from the images to be used as an input for algorithms. These algorithms would then be trained to find patterns in these features capable of distinguishing between different classes.

Neural networks, on the other hand, have the advantage of being able to learn these features directly from the data, which makes the process of feature engineering a lot simpler and achieves better results in most cases. Furthermore, the recent advances in deep neural networks have taken these capabilities to a new level. They not only win most of the competitions in the field but also achieve state of the art results in a wide range of real applications. One type of network that is responsible for these results is the Convolutional Neural Network (CNN) [Lecun et al., 1998].

3.1 Convolutional Neural Networks

Convolutional neural networks are similar to traditional neural networks. They both have an input layer which receives the data, followed by hidden layers with numerous neurons with weights and biases capable of learning the characteristics of the data, and a fully connected output layer at the end which is responsible for classification. The main difference is that CNNs assume the input has some spatial relationship, which is a pattern present in images. Thus, knowledge of where pixels are located in reference to each other is preserved. CNNs are capable of extracting and capturing patterns from the images that would not have been possible if we used traditional networks. To extract these features, the networks uses two primary operations: convolution and pooling.

Convolutions are linear mathematical operations that act as learnable filters (also called kernels) to capture patterns in the images. These filters are usually small in terms of dimension, typically 3x3 or 5x5 matrices. Each one of them convolves across the width and height of the input image and compute dot products with the pixels of the image, producing an activation map out of it. These activation maps, once learned, are able to detect features in the images, such as edges, corners, or color shifts.

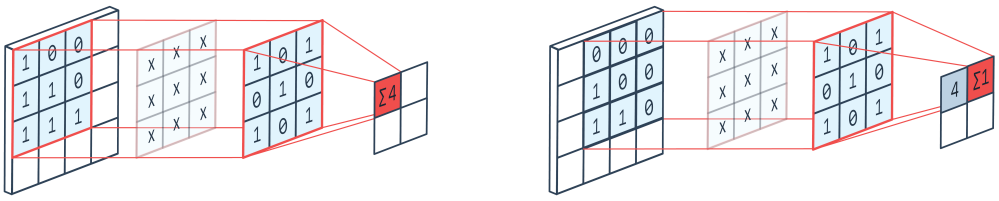


Figure 3.1: The convolution operation. Figure extracted from the Peltarion website[1].

Pooling is another mathematical operation responsible for reducing the spatial size of the convolved feature. These series of transformations reduce the dimensionality of the data and makes it possible to process images of high resolution. The most common cases of pooling are average pooling and max pooling, which are illustrated on Figure 3.2.

(a) Average Pooling (b) Max Pooling

Figure 3.2: Common pooling types. Figure extracted from the Peltarion website[2].

As multiple convolutional and pooling layers get stacked, the network becomes able to detect more complex patterns that are composed of multiple inputs of different feature extractors in the first layers. By turning this activation maps back into images, we are able to see what kinds of features they are detecting, as demonstrated by [Zeiler and Fergus, 2014].

[1] https://bit.ly/38bR3uK
[2] https://bit.ly/2Pzpkxq

3.2 Transfer Learning

Transfer Learning is a technique commonly used in machine learning when a learning model that was originally developed for one task is then reused on a second related task. It comes from the assumption that what has been learned in one setting can be used to improve optimization in another setting. The idea behind is inspired by human behavior, as sometimes we can use expertise in solving one problem to solve a new one.

Another motivation behind using transfer learning comes from the high computational cost necessary to train large deep neural networks for image classification. Since the number of parameters present in a CNN is very high, it requires a large amount of training data to tune the network for making precise predictions.

As an example, a commonly used dataset in the field for pre-training networks is a subset of ImageNet [Deng et al., 2009], which contains 1.2 million images and has 1000 labeled categories for classification. Even with today's computational power, it still requires a significant amount of hours to be trained.

In this scenario, using transfer learning trough pre-trained networks arises as a solution. In this process, the weights and biases from a network trained in another task, are reused to train a new similar task.

Another reason for using transfer learning comes from the cases where we do not have enough data to train a CNN. Celona et al. [2019] highlights this is especially true in the medical field, as the acquisition costs are elevated, and it also involves a complicated set-up for photographing, which makes it very common to have little annotated data.

3.3 Data Augmentation

Another solution that handles small datasets is data augmentation. It consists of applying transformations, such as geometric and color augmentations, for generating alternative images that derive from the original dataset.

For each input image in the dataset, a new image is generated that can be zoomed, shifted, mirrored, rotated, distorted, or have changes in its color, brightness, contrast. Hence, this technique increases the amount of data available for input.

Having a large dataset is crucial for the performance of the deep learning models, but instead of starting with a large dataset of images, a more common scenario is to have a small amount of data available from the specific domain of research. This usually happens due to the high cost of collecting data, be it in terms of human labor

or monetary resources. As mentioned in the previous section, this is mainly the case in the medical field.

Another problem of small datasets is that problems trained on them are often over-fitted to the specific data available, which means they lack the power of generalization, as the dataset is not representative of the real world. In these cases, as discussed by Perez and Wang [2017], data augmentation can act as a regularizer for preventing over-fitting and also improve performance in imbalanced class problems.

Chapter 4

Data

Fetal therapy is a promising field in pediatric medicine, and prenatal surgery has become an option for an increasing number of babies with congenital disabilities. Regardless of its popularity increase, it is still a relatively rare procedure as it affects only a small percentage of pregnancies and offers risks for both the mother and the unborn baby. The procedure is also highly sophisticated and thus requires a skilled team, assisted with advanced technological resources to perform such complex procedures. The particularities of this topic, make research in the field very challenging.

At the same time, as fetuses are protected against biological and social effects while in utero, the circumstances offer an excellent opportunity for investigation of fetal behavior free of influences from the outside world. The study of fetal pain is a great example, as it is still a topic of debate if human fetuses feel pain or not. The Fetal Pain Study Group from the University of São Paulo was formed with the purpose of helping answer these questions. In the following section, we describe one of their most recent studies, which consisted of the assessment of pain trough facial expressions in fetuses.

4.1 Fetal Pain Study

Trough the use of high definition 4-D ultrasound machines, it is possible to record and observe fetal responses to different stimuli, and by looking at their facial expressions and body movements, one could potentially assess visual pain responses during the intrauterine life. This was done by the aforementioned study group, which proved the feasibility of using a pain scale, initially developed for acute pain assessment in neonates, in fetuses [Bernardes et al., 2018].

Based on this hypothesis, the Fetal Pain Study Group conducted a novel study, which is the first attempt to assess specific pain-related facial patterns in human fetuses. They were able to evaluate facial expressions after an anesthetic injection was administered before an intrauterine surgical procedure, which was used as a model of acute pain. They were pioneer in the usage of two ultrasound machines for this purpose, as the exact moment of the anesthetic puncture was recorded to capture the reaction of the fetus and its manifestations of pain.

While one machine was used to perform the anesthesia, a second ultrasound machine was placed in the clinical room and operated by a fetal medicine specialist to monitor the fetus's face and its expressions. The spatial set-up of the room can be seen in Figure 4.1.

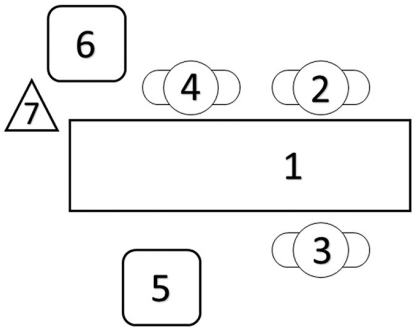


Figure 4.1: Operating room set-up for surgery and face recording. (1) Position of the mother; (2) chief surgeon who performed the puncture; (3) assistant surgeon who obtained the 4-D images; (4) surgical technologist; (5) ultrasound machine used in surgery focusing the fetal trachea/thigh; (6) ultrasound machine used for fetal face recording; and (7) an external camera. Diagram extracted from [Bernardes et al., 2018]

In order to measure and quantify pain, a second study (yet to be published) evaluated the presence or absence of pain in a larger group of 13 fetuses. Besides the anesthetic acute pain stimulus, two other scenarios were used as control conditions: resting, and responses after an acoustic stimulus of a horn, which is routinely used to assess fetal well-being. It is important to notice that for the acute pain group, all fetuses were previously diagnosed with diaphragmatic hernia and had an indication of intrauterine surgery (fetoscopic endoluminal tracheal occlusion). They were all assessed in their preoperative period.

This study was then able to refine the Neonatal Facial Coding System (NFCS) to be more suitable for the application on fetuses. As fetuses can display facial expressions unrelated to pain [Reissland et al., 2011], the scoring system should be capable of discriminating acute pain responses from those at rest and from other non-painful

stimuli that also trigger facial expressions, like the vibro-acoustic sound of a horn. After refinement, indicators unable to discriminate between painful stimuli, and the control groups were removed. Likewise, indicators that were undetectable from static images were also not considered. Additionally, one item deemed relevant for the research was added: neck deflection.

The final scale thus contained the following seven items: brown lowering, eyes squeezed shut, deepening of the nasolabial furrow, open lips, horizontal mouth stretch, vertical mouth stretch and the new item neck deflection. Each item is considered one point if present or zero if absent on a given screenshot, then the present items are summed to give an overall score, and the scale ranges from zero to seven. The study concluded that no fetus in the control groups had a score higher than four, and at the same time, in the acute pain group, no score was less than five. These results allowed researchers to determine that a "pain cut-off" exists in the new seven-item scoring system.

In summary, the study concludes that fetal humans undergoing an anesthetic puncture show facial expressions changes, which can be detected, quantified, and scored using a refined scale derived from one used in newborns. Furthermore, the features of the facial expressions present while in acute pain exposure are sufficiently discriminative from those expressed while in rest or during a sound stimulus. Hence, making it possible to establish a threshold which separates acute pain responses from non-painful responses.

4.2 Data Description

The data collected by the second study and its findings present a unique opportunity to do further experiments. To the best of our knowledge, no publicly available dataset exists with images or videos of fetuses while in acute pain exposure. This fact alone highlights the novelty and innovative aspects of the study mentioned above and our research.

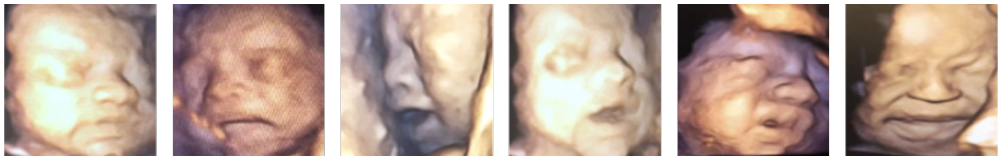
A total of 13 films were recorded from a 4-D ultrasound machine of the model Voluson E8 by General Electric, being 6 from the acute pain group, 4 from resting conditions, and 3 from the exposure to acoustic stimulation. An example image from each one of them can be seen on Figure 4.2. All the fetuses were in the third trimester of gestation, with an average of 31.1 $\pm$ 2.8 weeks. Videos from each of these three conditions were collected as follows:

- Acute Pain (**AP**): fetuses from this group were diagnosed with a diaphragmatic

hernia, which indicated intrauterine surgery (fetoscopic endoluminal tracheal occlusion). The videos were recorded in the preoperative period during the anesthetic puncture, using the setup described in the previous section. Videos from this group had two parts in it, first a baseline period defined as the first 45 seconds before the anesthesia puncture and second the 45 seconds immediately after the puncture.

- Rest (**RE**): fetuses in this group were recorded during routine ultrasound exams to assess fetal well-being. The videos lasted 45 seconds and begun after a 5 minutes period of rest for the mother. All the fetuses were considered healthy.
- Acoustic Stimulus (**AS**): fetuses from this group were exposed to a vibro-acoustic stimulation that is used to improve the efficiency of fetal heart rate testing to assess fetal well-being. Their facial expressions were recorded 45 seconds before and after the stimulus of a horn, which was applied to the maternal abdomen next to the cephalic fetal pole for approximately 4 seconds.

All mothers gave written informed consent to participate in the study and to record the behavioral reactions of the fetuses. The study was also approved by the ethics review board of the Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, under protocol number 2.649.528.



(a) Acute Pain Group

(b) Rest Group

(c) Acoustic Stimulus Group

Figure 4.2: Images of each individual fetus grouped by their conditions.

Chapter 5

Deep Learning Models for Fetal Pain Assessment

Based on the techniques mentioned in Chapter 3, we have proposed a process and a few learning models for classifying images of fetuses with facial expressions containing the presence of pain or not. In summary, our pipeline consists of sampling the videos into frames, finding the images which contain a clear fetus face, and training a Convolutional Neural Network (CNN), with the help of transfer learning, for the binary classification task of finding the presence of pain. In the following sections, we describe each of these steps.

5.1 Image Sampling

It is common to have a small number of data to work within the medical field in general, given the inherent difficulty of collecting it [Zhang et al., 2019]. In our case, especially, only a small percentage of pregnancies require intra-uterus intervention before birth, and thus fetal anesthesia is a relatively rare procedure. Thus, as seen in the previous chapter, we ended up with 13 videos available, which is a number similar to what we have seen in other studies, such as the iCOPE database.

Since we had a small number of videos, it was not possible to work with them directly. So, we brought the data to another dimension, reducing the space from videos to images by sampling them and capturing frames at a rate of every 2 seconds.

With this process, we generated a total of 508 images, but since the images were recorded from ultrasound machines, they depend on the calibration by the specialists to capture the exact section of the 3-D space where the fetus's face is clear. Because of this, it was common to find parts of the video where the face of the fetus was not visible

and showed non-distinguishable parts. As we had a significant number of images, and manual selection would be not only hard but also dependent on the observant, this became a problem. Thus, our final dataset consisted of **226** images.

To overcome this issue, we decided to use another neural network capable of detecting facial landmarks, like the nose, the mouth, and the eyes. The network we used was the Multi-task Cascaded Convolutional Networks (MTCNN) developed by Zhang et al. [2016], which is trained to identify faces in images. It worked surprisingly well in our domain, even though the images had quite different characteristics.

With this process, we were able to filter our dataset and reduce the number of images from 508 to 232, but being sure the images contained a clear face. The network also returns a confidence value of which it found the face in the image, and we have used only confidences of over 95%, which, after manual inspection, showed to be very reliable, with just six clear errors that were removed manually.

The position of the facial landmarks encountered by the network also allowed us to crop images around the fetus's face. This process is achievable after we have the coordinates of the landmarks returned through the MTCNN, which also makes face alignment possible. This helps to discard images with blurred surroundings around the fetus, which contains non-distinguishable parts. In Figure 5.1, we can see an example of a sampled image and its respective cropping.

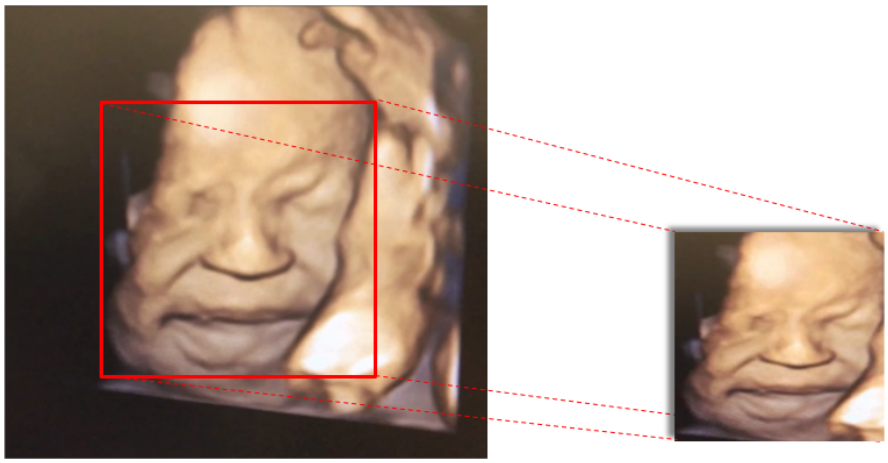


Figure 5.1: Image cropping with MTCNN.

In the videos of acute pain, as we knew precisely when the anesthetic puncture stimuli were applied, it was possible to divide the images into the two classes of pain

and non-pain. This division is relevant, as it allows us to experiment in the scenario where we can evaluate the same fetus, for both conditions. In the other two groups, this is not possible as we have only images of the non-pain class.

5.2 Data Augmentation

Even though we had increased the size of the dataset by turning the videos into images, it is still considered a relatively small dataset for deep learning models. To further augment our chances of succeeding, we have applied the use of data augmentation techniques to increase the variability of our data. The effectiveness of this technique has been demonstrated by Perez and Wang [2017] and is widely used in the field.

There is a wide variety of transformations possible for using data augmentation, and even simple techniques already work very well. We have chosen to apply the following transformations:

- Horizontal flip, which mirrors the image horizontally.
- Rotation, which applies rotations to the images up to a maximum degree.
- Zooming, which zooms into parts of the image up to a maximum level.
- Warping, which adds distortions to the image up to a maximum level.
- Lighting, which changes the brightness and the contrast of the images.

All of these methods have a probability of being applied and can be used in combination with each other. Thus for each image, given the probability, a combination of these techniques would be applied. Some examples of these different combinations within the same image are shown in Figure 5.2.

To further experiment with this process, we have compared two levels of intensity in the changes regarding their max levels of rotation, zoom, warping, and lightning. First, a weak set of transformations, which does subtle changes in the images. Later, a stronger set, which applies substantial changes to the images.

5.3 Residual Networks (ResNets)

Uncountable convolutional neural network architectures have been developed by researchers around the world with many creative modifications, designed for a range of

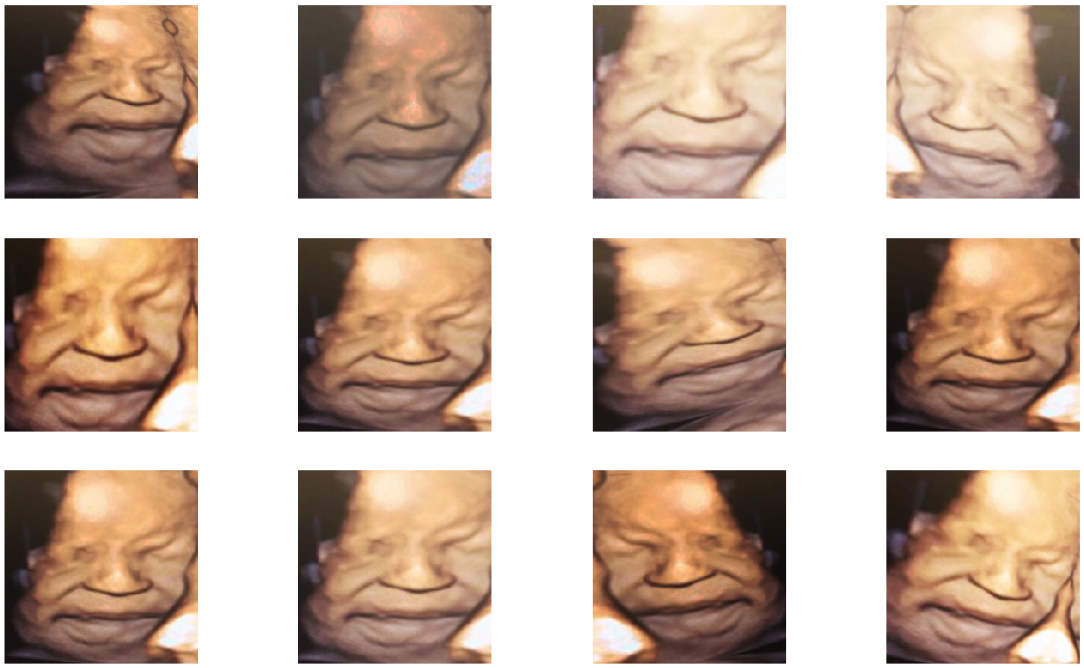


Figure 5.2: Application of different data transformations to a fetus image

applications. One type of network that is commonly used is the Deep Residual Network (ResNet) developed by Parkhi et al. [2015]. This network was the winner of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky et al., 2015], achieving an error rate of only 3.57% in the image classification task and being the first to beat human performance.

The intuition behind ResNet emerged when the researchers noticed that as they increased the depth of a regular network, its training and test errors got worse than when compared to an equivalent shallow network. This happens because of a well-known problem, the vanishing gradient. As the gradient back-propagates to the earlier layers of a network, the repeated multiplications make the gradient infinitely small, which prevents it from reaching the weights of the earlier layers.

Other techniques have been developed to deal with this problem, such as batch normalization, but despite that, deeper networks still suffer from degradation in convergence, as the errors remain higher than if it was on an equivalent shallow network.

The insight the authors had, was that when adding extra layers, if these layers are identity mappings, they become equivalent to the shallower network. Thus, the deeper network should not produce an error higher than its shallower counterpart. They achieved this behavior by injecting these identity mappings in the network trough shortcut connections, which are simply connections skipping one or more layers. The output of these layers is added to the outputs of the stacked layers, and add no extra

parameter nor computational complexity to the networks.

Another insight they had was regarding residuals, which are just the error in a result. Thus, the network should be able to learn these residuals so that the predictions are closer to the actual values. By combining these two ideas, they have created the residual learning building block, as shown in Figure 5.3 extracted from the original paper.

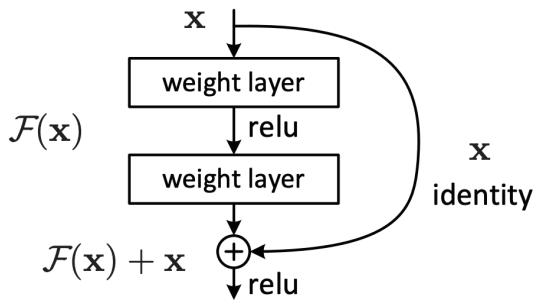


Figure 5.3: Residual Network building block. Image extracted from Parkhi et al. [2015]

During the training period, the ResNet learns the weights of its layers in a way that if the identity mapping were optimal, all the weights would be set to zero.

In the diagram, we can see that if $F(x)$ becomes zero, it means x is getting directly mapped to the actual value, and no corrections need to be made. These are the identity mappings that help the network grow deeper. On the other hand, if there is a deviation from optimal identity mapping i.e., a residual, the weights and biases of $F(x)$ will be learned to adjust for it. In other words, $F(x)$ learns how to adjust our predictions to match the actual values.

These building blocks are stacked together to arrive at a deep network architecture. Figure 5.4, also extracted from the original paper, shows a comparison between a 34-layer plain network and 34-layer ResNet. In our work, we have used 50-layer one with the intuition that a relatively larger network would yield better results.

5.4 Transfer Learning and Fine-tuning

As described in Chapter 3, training a deep neural network from scratch is a very costly task, both in terms of the amount of data necessary as well as in terms of computational power. Because of these factors, transfer learning is often used in a variety of applications as a solution for when we have a limited amount of data, which is especially the case in the medical field. The intuition is that by using pre-trained weights, we can get the benefits of networks trained on much larger datasets, often

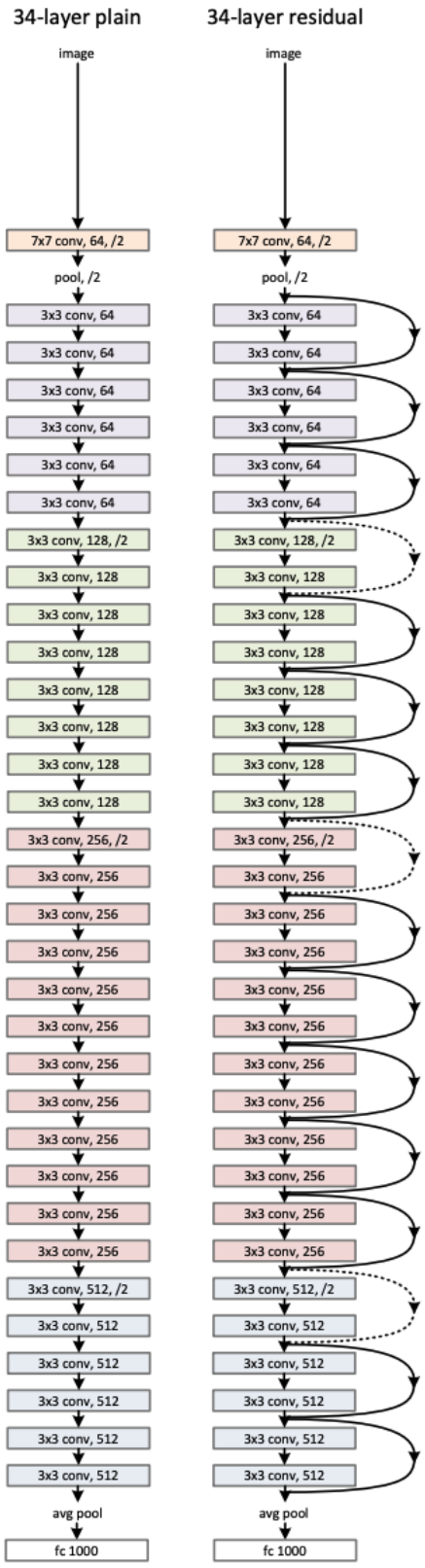


Figure 5.4: 34-layer plain network in comparison with 34-layer residual network. Image extracted from Parkhi et al. [2015]

with millions of images. Then, by changing only the last fully connected layer of the network to match the number of classes in our problem, we are able to fine-tune them with our data.

One general-purpose dataset very often used for pre-training is ImageNet [1] [Deng et al., 2009], which consists of more than 14 million images, which have been hand-annotated by the project to indicate what objects are pictured on them. This visual database was designed for use in visual object recognition software studies and had a significant impact on deep learning research. The most popular network architectures we know today, have emerged in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), an annual computer vision contest held between 2010 and 2017, which uses a subset of the ImageNet database with 1.2 million images and 1000 classes.

Convolutional neural networks are very often pre-trained on this dataset, as it contains images of many different categories, like plants, animals, cars, objects, and many others. Hence, by training on this data, the network can learn low-level features such as lines, edges, and basic shapes, but also mid-level features which build on top of the low-level ones, and can detect objects or more complex shapes. These types of features can be considered independent from the final task to some extent, as for many applications detecting basic shapes will often be necessary.

Another popular large-scale dataset used for pre-training is the VGGFace2 [2] [Cao et al., 2018], which consists of 3.3 million images of faces downloaded from Google Image Search, with variations in pose, age, illumination, ethnicity, and profession. A few example images from this dataset can be seen in Figure 5.5.

This dataset was the same used by Zamzami et al. [2018] for automatically detecting pain in infants, therefore we hypothesize it would perform well in our data too. We believe the high-level features learned by a network trained on this dataset should be able to detect similar features to the ones we aim to detect in fetuses, such as the mouth, the nose, and the eyes, to mention a few.

When using transfer learning, the features computed in the early layers of the network usually are already well trained in doing basic tasks such as recognizing basic lines, patterns, or gradients. On the other hand, the features computed in the later layers are the ones highly dependent on the specific task we are trying to predict. Thus, when we are fine-tuning the network in our domain, we have two main approaches to train the network, namely with frozen or unfrozen layers. These methods allow us to decide which specific layers of our model we want to train at a given time.

[1] `http://www.image-net.org`
[2] `http://www.robots.ox.ac.uk/~vgg/data/vgg_face2`

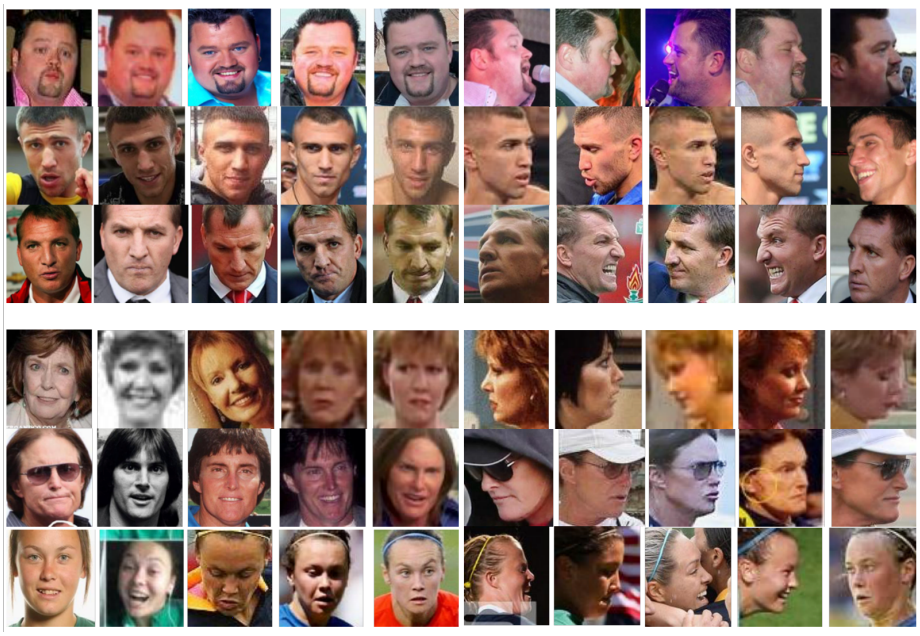


Figure 5.5: VGGFace2 example images. Images extracted from Cao et al. [2018]

In the frozen approach, all layers except the last one will not be trainable. In the unfrozen approach, however, all the layers are kept unfrozen during training, and the errors are back-propagated to the entire network during fine-tuning. Hence, even the early layers may be affected. In our experiments, we decided to test both approaches, as even though unfreezing the layers appears to be better, we were unsure if the amount of data and its noise would be enough to improve the weights of the whole network.

Chapter 6

Experimental Results

In this chapter, we present the experimental results of our research. As no previous work has attempted to develop models for automatic pain assessment in fetuses, we have no baseline to compare to. Thus, we discuss the decisions we have made along the way, which led to our best results. In particular, our experiments aim to answer the following research questions:

- **RQ1:** Is it possible to identify the presence of pain in images of fetuses from 4-D ultrasound machines? Can we create an effective learning model for automatic pain identification?
- **RQ2:** Is this model capable of discriminating images of fetuses while in acute pain exposure from those in a control group while in rest our in a non-painful sound stimulus?
- **RQ3:** Does transfer learning transfer well from a face recognition task in adults to our domain with fetuses?

6.1 Setup

Given we had relatively few data available, we chose a validation strategy that works best in this scenario. Like Celona and Manoni [2017], we used the leave-one-out method for cross-validation, but instead of leaving one image out, we leave one subject. Additionally, we make use of the images from the Acoustic Stimulus (AS) group only for training purposes, as they belong to a control group, this scenario wouldn't be evaluated in a real-life application.

Hence, we produce 10 different combinations containing training and test subsets, given that Acoustic Stimulus (AS) images are always on training. On each of these combinations, we train our networks in the training subset with images of twelve fetuses and evaluate on the test subset with images of one. All the evaluations are then averaged to assess the overall performance of the models.

In order to find the best network architecture for this particular problem, we have tested a few variations in the setup as described in Chapter 5. These variations are regarding three variables:

- Data augmentation, which could be with weak or strong transformations.
- Network training, which could be with frozen or unfrozen layers.
- Pre-training, which could be on the ImageNet or the VGGFace2 datasets.

By combining all the possibilities of these three variables, we have a total of eight experiments. Like Zamzami et al. [2018], we have chosen two types of pre-training for the CNNs, so we can compare the differences between using CNNs trained on a relatively similar dataset like VGGFace2, as opposed to CNNs trained on a general-purpose dataset like ImageNet.

Besides these variations, all the networks used Adam as a gradient descent optimization algorithm [Kingma and Ba, 2015], which uses adaptive momentum to reduce the error in the training set quickly. We have used a batch size of 8 for both training and validation, which yielded the best results after we have experimented with different sizes (4, 8, 16, 24). We have also applied some methods to prevent over-fitting like L2 for weight regularization [Ng, 2004] and dropout [Srivastava et al., 2014]. Lastly, the loss function we used was binary cross-entropy, as it shows good performance for classification problems with two classes.

The metric we used to evaluate our model during the validation process was accuracy. To calculate it for a given test set, we divide the number of images we have predicted the correct class by the total number of images available in that set.

Additionally, we have also calculated another metric for the videos of acute pain (AP). As we have 45 seconds of video before the acute pain stimulus, and 45 seconds after it, we have images from both classes in these videos: pain and non-pain. This division allows the use of a metric that considers not only the cases we are making the correct prediction but also how much of each class we are making the wrong predictions. Thus, like Zamzami et al. [2018], we have used the Area Under the Receiver Operating Characteristic Curve (AUC) to evaluate the performance of our models in the set of acute pain videos.

6.2 Results

In this section, we compare the performance of each training approach and discuss their results. Table 6.1 displays the results in terms of accuracy considering each training method, which gives us some insights about the behavior of the different models. For instance, we can see our best result came from a pre-training on VGGFace2, which confirms our hypothesis that it was better to use pre-training on a set of images similar to ours and that the features learned from these images transfer well to fetuses.

However, when looking at all the results, we can see that the overall standard deviation was reasonably high, which shows how challenging the task is when we have little data. In fact, by looking only at the dimensions of training type and transformations, a clear winner approach is not evident, as the results are very similar and one variation not always perform better than the other.

Table 6.1: Accuracy comparison considering all videos.

			Accuracy	
			Mean	Std
Training	Transforms	Network		
frozen	weak	ResNet (ImageNet)	0.786	0.231
frozen	weak	ResNet (VGGFace2)	0.821	0.162
frozen	strong	ResNet (ImageNet)	0.772	0.211
frozen	strong	ResNet (VGGFace2)	**0.848**	**0.143**
unfrozen	weak	ResNet (ImageNet)	0.804	0.214
unfrozen	weak	ResNet (VGGFace2)	0.784	0.213
unfrozen	strong	ResNet (ImageNet)	0.782	0.205
unfrozen	strong	ResNet (VGGFace2)	0.812	0.168

When we look at the accuracy reported in each test set for the best model in Table 6.2, we can see the model performs reasonably well both in the acute pain (AP) and rest (RE) groups, with an average accuracy score of 0.803 in the former and 0.917 in the latter.

Table 6.2: Accuracy per test set in the leave-one-out for the best model.

Accuracy									
1_{AP}	2_{AP}	3_{AP}	4_{AP}	5_{AP}	6_{AP}	7_{RE}	8_{RE}	9_{RE}	10_{RE}
0.917	0.824	0.583	0.875	0.850	0.769	1.000	1.000	1.000	0.667

Nonetheless, we can also take a closer look at the results from the acute pain (AP) group, from which we can measure the AUC. As we can see on Table 6.3, we

have performed much better on this group, especially considering we had images of the same fetuses on both states, pain and non-pain. This result is very promising, as it indicates our model is able to discriminate pain from rest on images of fetuses.

Table 6.3: Accuracy and AUC considering only Acute Pain videos.

			Accuracy		AUC	
			Mean	Std	Mean	Std
Training	Transforms	Network				
frozen	weak	ResNet (ImageNet)	0.710	0.233	0.849	0.173
frozen	weak	ResNet (VGGFace2)	0.768	0.155	0.885	0.150
frozen	strong	ResNet (ImageNet)	0.698	0.205	0.850	0.130
frozen	strong	ResNet (VGGFace2)	**0.802**	**0.118**	**0.923**	**0.063**
unfrozen	weak	ResNet (ImageNet)	0.684	0.198	0.813	0.163
unfrozen	weak	ResNet (VGGFace2)	0.763	0.132	0.898	0.112
unfrozen	strong	ResNet (ImageNet)	0.692	0.185	0.909	0.110
unfrozen	strong	ResNet (VGGFace2)	0.742	0.139	0.833	0.156

We can see that the best model is still the same as the one from Table 6.1. However, now we have some more insights in terms of the other two dimensions. For example, we can see that the strong transformations have a slight advantage when compared to the weak in terms of AUC. Likewise, training the network with frozen layers performs better than unfrozen in terms of accuracy, which could also be caused by the noise in the data, so the error propagates back into the first layers, causing the network to worsen its performance. Although more data would be ideal to make more conclusions.

Also, we can see the standard deviation is much lower, which can be explained not only by the fact we have fewer validations sets to consider but also because our model is performing better at predicting acute pain videos.

Table 6.4: AUC per test set in the leave-one-out for the best model, considering only Acute Pain videos.

AUC					
1_{AP}	2_{AP}	3_{AP}	4_{AP}	5_{AP}	6_{AP}
0.991	0.983	0.829	0.938	0.870	0.929

When we look at the result from each test set of acute pain (AP) from the best model in Table 6.4, we see that videos 3_{AP} and 5_{AP} have a lower AUC, which shows how much variations in the images can affect the final result when we work with a small number of subjects.

6.3 Visual Explanations

The success of convolutional neural networks came with the ever-increasing complexity of the architectures, which led to difficulties in understanding why the models make certain decisions. Some methods exist to try to overcome this issue, such as Grad-CAM [Selvaraju et al., 2017], which tries to provide visual explanations of why the model made a given decision. This method uses the gradients of the target flowing back into the final convolutional layer to produce a heat map that highlights the important regions in the image that were used for prediction.

We have attempted to use this technique to identify the parts of the image our models found that were the most relevant for classifying an image as pain. In an ideal scenario, the heat map should be stronger in the parts of the image that are indicators of pain, like the ones used by the pain scales. However, even though this did happen for some examples, as we can see in Figure 6.1, for most cases, the heat map was inconclusive. We believe this could also be an effect of the limited amount of data, thus we propose this topic gets further investigated in a future work.

Figure 6.1: Grad-CAM heat map for visual explanations.

It is important to highlight these explanations are an essential feature in an eventual application of our system in real life. As doctors and specialists look at the images produced by the model, the heat map should ideally agree with the indicators from pain scales and direct their view to regions of interest where manifestations of pain are present. We believe this feature would certainly give more robustness to the results and facilitate adoption by the users. As such, we stand out its importance and suggest further research as a future work topic in the next chapter.

6.4 Answering Our Research Questions

In this section, we aim to answer the proposed research questions from the beginning of this chapter based on the results we presented.

Regarding **RQ1**, we believe our results showed in tables 6.1 and 6.3 are a good argument to show it is viable to develop a learning model capable of effectively identifying pain. We believe an accuracy of 84.8% is a good evidence that we are in the right direction, even considering that a new experiment with more data could be necessary to validate these conclusions any further. As data is complicated to collect, we think our experimental process was able to extract significant results out of it.

As for **RQ2**, we did see our model had problems in discriminating images of acute pain from those of an acoustic stimulus, although it did so very well from images of rest. It appears the images from acoustic stimulus indeed made the task more difficult, but we believe this effect would be mitigated if we had a larger dataset available for training. Nevertheless, even if we have a false positive predicted from an acoustic stimulus image, from a precautionary perspective, it would still benefit the fetus, as this could be an indication of discomfort and could also be treated.

Finally, for **RQ3**, it does appear transfer learning with pre-training in the VGGFace2 dataset performs better than when trained on ImageNet, as it achieved our best result. We believe this comes from the fact the pre-training on face images was able to learn features in their middle to last layers related to the human face, such as the mouth, the eyes, the chin, the nose. Even though the fetus images are relatively different, these features are still present, which could explain why the model detected them and performed better.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The results of our study presented in this dissertation are promising as we believe they move us towards the ultimate goal of automatically detecting pain in fetuses. Our learning model was indeed able to discriminate images of fetuses while in acute pain exposure from those in control groups of rest and acoustic stimulus. If confirmed on a larger dataset, we believe this work has the potential to influence and improve the current practice of assessing fetal pain.

We also think our work can serve as good evidence to help answer the question of when fetuses start to feel pain, as the exact gestational week in which they start showing pain responses is still not a consensus. By providing an unbiased and automatic approach for detecting pain, we could continuously monitor a fetus across many weeks, and an absence of pain may even be a good indicator of fetal wellbeing. However, a dataset with more variable gestational age would be necessary to train such a model.

On the other hand, if we do detect pain, the question arises as to what is the cause of it, as some condition may be present since our model accused the presence of pain. Is this condition some malformation? Is it a disease? Or is it related to chronic pain? It does open a range of possibilities but also brings awareness to future problems, which in some cases could be corrected with in-uterus repairs, with many benefits for the fetus.

As for the control group of acoustic stimuli, our model found it more difficult to discriminate it from pain, which was an expected outcome. Because the group shares some common indicators with those of pain, it makes it indeed harder to predict, as shown in the original study that collected the data [Bernardes et al., 2018]. Nonetheless, we still believe that from a precautionary approach, it may be a good practice to

investigate these cases, as they can be signs of discomfort or stress, and may also be caused by some conditions.

If our model eventually gets integrated into an ultrasound machine, it would make pain detection much simpler and easier to use, allowing continuous monitoring. This would be beneficial in many situations, especially during fetal surgery procedures, as it could aid anesthesiologists to see how effective their anesthesia was and aid the doctors while they perform delicate procedures. Likewise, during routine prenatal ultrasound exams, this system could be the first to indicate pain or discomfort, which would then be further investigated.

It is also important to highlight that even though nurses and caregivers must assess the videos used as inputs by our model, we think a model trained on a range of different videos from different sources, would tend to be much more unbiased than an assessment of a single caregiver. This result is also a great benefit, as we can produce a system ideally free from bias factors such as identity, background, culture, and gender, which may lead to inconsistent assessment and treatment of pain.

7.2 Future Work

Our studies have shown that it is viable to construct a model capable of identifying the presence of pain on images of fetuses from 4-D ultrasound machines. A larger dataset is already being collected by the same fetal pain study group, which has the potential to confirm our results and produce models that are even more robust and accurate. We are also currently not able to explain why the model made such predictions or what is the main factors it considered for detecting pain. Thus, we identify as future work the following possibilities:

- Evaluate our models on larger datasets. Even though the data is quite complicated to collect and studies with fetuses and infants usually have a small number of subjects, we think it would be very beneficial to experiment with our methodology in a more extensive number of fetuses. This addition could bring more variability into the model inputs in terms of fetal positions, gestational age, gender, image quality, and many other factors, which will end up producing a better model.

- Expand our system to include chronic pain. Monitoring the same fetuses at different gestational ages has the potential to identify the presence of chronic factors. A model that evaluates not only acute pain but also chronic pain could,

therefore, help in this scenario, as it may lead to further investigation of what is causing the chronic pain and maybe be the first indicator that an intervention may be necessary.

- Include other types of features. As it is the case with pain scales, the combination of indicators is what tends to work best. Thus one could construct a model that takes as inputs not only images of the face and facial expressions but also other indicators such as sounds, body movements, physiological indicators, and biological markers. These new factors could help produce more robust models, and maybe identify conditions not visible through facial expressions only.

- Produce explainable models. Given a single fetus where the presence of pain has been identified, one should be able to visualize what are the most relevant features that the model analyzed to output its prediction. By making the decision of the models more transparent, one could point to the exact locations where the pain was present, which will lead to a better understating by fetal pain specialists. In an ideal scenario, these visual explanations should match the individual indicators present on the pain scale.

In summary, our main interests as future works are to help medical experts to understand the output of the models better and be more effective on pain assessment and management, which will eventually lead into improving overall fetuses life quality and well-being.

Bibliography

Adzick, N. S., Thom, E. A., Spong, C. Y., Brock, J. W., Burrows, P. K., Johnson, M. P., Howell, L. J., Farrell, J. A., Dabrowiak, M. E., Sutton, L. N., Gupta, N., Tulipan, N. B., D'Alton, M. E., and Farmer, D. L. (2011). A randomized trial of prenatal versus postnatal repair of myelomeningocele. *New England Journal of Medicine*, 364(11):993--1004.

Bellieni, C. V. (2012). Pain assessment in human fetus and infants. *The AAPS Journal*, 14(3):456--461.

Bernardes, L. S., Ottolia, J. F., Cecchini, M., de Amorim Filho, A. G., Teixeira, M. J., Francisco, R. P. V., de Andrade, D. C., de Estudo da Dor Fetal, G., et al. (2018). On the feasibility of accessing acute pain–related facial expressions in the human fetus and its potential implications: a case report. *Pain Reports*, 3(5).

Brahnam, S., Chuang, C., Shih, F. Y., and Slack, M. R. (2005). SVM classification of neonatal facial images of pain. In *Fuzzy Logic and Applications, 6th International Workshop, WILF 2005, Crema, Italy, September 15-17, 2005, Revised Selected Papers*, pages 121--128.

Brahnam, S., Chuang, C.-F., Shih, F. Y., and Slack, M. R. (2006). Machine recognition and representation of neonatal facial displays of acute pain. *Artificial Intelligence in Medicine*, 36(3):211--222.

Brahnam, S., Nanni, L., and Sexton, R. S. (2008). Neonatal facial pain detection using NNSOA and LSVM. In *Proceedings of the 2008 International Conference on Image Processing, Computer Vision, & Pattern Recognition, IPCV 2008, July 14-17, 2008, Las Vegas Nevada, USA, 2 Volumes*, pages 352--357.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE.

Celona, L., Brahnam, S., and Bianco, S. (2019). Getting the most of few data for neonatal pain assessment. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2019, Trento, Italy, 20-23 May 2019*, pages 298--301.

Celona, L. and Manoni, L. (2017). Neonatal facial pain assessment combining handcrafted and deep features. In *New Trends in Image Analysis and Processing - ICIAP 2017 - ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers*, pages 197--204.

Cope, D. K. (1998). Neonatal pain: the evolution of an idea. *The American Association of Anesthesiologists Newsletter*, pages 6--8.

Darwin, C., Murray, J., Duchenne, G., and Rejlander, O. (1872). *The Expression of the Emotions in Man and Animals*. Marilee E. Thomas and Robert C. Thomas Science and Related Subjects Collection. John Murray.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248--255.

Derbyshire, S. W. and Bockmann, J. C. (2020). Reconsidering fetal pain. *Journal of Medical Ethics*, 46(1):3--6.

Derbyshire, S. W. G. (2006). Can fetuses feel pain? *BMJ*, 332(7546):909--912.

Devoto, J. C., Alcalde, J. L., Otayza, F., and Sepulveda, W. (2017). Anesthesia for myelomeningocele surgery in fetus. *Child's Nervous System*, 33(7):1169--1175.

Fisk, N. M., Gitau, R., Teixeira, J. M., Giannakoulopoulos, X., Cameron, A. D., and Glover, V. A. (2001). Effect of direct fetal opioid analgesia on fetal hormonal and hemodynamic stress response to intrauterine needling. *Anesthesiology*, 95(4):828--835.

Fotiadou, E., Zinger, S., a Ten, W. E. T., Oetomo, S. B., and de With, P. H. N. (2014). Video-based facial discomfort analysis for infants. In Said, A., Guleryuz, O. G., and Stevenson, R. L., editors, *Visual Information Processing and Communication V*. SPIE.

Giannakoulopoulos, X., Glover, V., Sepulveda, W., Kourtis, P., and Fisk, N. M. (1994). Fetal plasma cortisol and β-endorphin response to intrauterine needling. *The Lancet*, 344(8915):77--81.

Gingras, J. L. (2005). Fetal homologue of infant crying. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 90(5):F415--F418.

Goldberg, D. S. and McGee, S. J. (2011). Pain as a global public health priority. *BMC Public Health*, 11(1).

Grunau, R. E., Oberlander, T., Holsti, L., and Whitfield, M. F. (1998). Bedside application of the neonatal facial coding system in pain assessment of premature infants. *Pain*, 76(3):277--286.

Harrison, M. R., Golbus, M. S., Filly, R. A., Callen, P. W., Katz, M., de Lorimier, A. A., Rosen, M., and Jonsen, A. R. (1982). Fetal surgery for congenital hydronephrosis. *New England Journal of Medicine*, 306(10):591--593.

Healy, J. (2016). When can fetuses feel pain? utah abortion law and doctors are at odds. `https://www.nytimes.com/2016/05/05/us/utah-abortion-law-fetal-anesthesia.html`. Accessed: 01/14/2020.

Houtrow, A. J., Thom, E. A., Fletcher, J. M., Burrows, P. K., Adzick, N. S., Thomas, N. H., Brock, J. W., Cooper, T., Lee, H., Bilaniuk, L., Glenn, O. A., Pruthi, S., MacPherson, C., Farmer, D. L., Johnson, M. P., Howell, L. J., Gupta, N., and Walker, W. O. (2020). Prenatal repair of myelomeningocele and school-age functional outcomes. *Pediatrics*, 145(2):e20191544.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lawrence, J., Alcock, D., McGrath, P., Kay, J., MacMurray, S., and Dulberg, C. (1993). The development of a tool to assess neonatal pain. *Neonatal network : NN*, 12(6):59—66. ISSN 0730-0832.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278--2324.

Lee, S. J., Ralston, H. J. P., Drey, E. A., Partridge, J. C., and Rosen, M. A. (2005). Fetal pain: A systematic multidisciplinary review of the evidence. *JAMA: Journal of the American Medical Association*, 294(8):947–954. ISSN 0098-7484.

Mauricio, A., Cappabianco, F. A. M., Veloso, A., and Cámara, G. (2019). A sequential approach for pain recognition based on facial representations. In *Computer Vision Systems, 12th International Conference, ICVS 2019, Thessaloniki, Greece, September 23-25, 2019, Proceedings*, pages 295--304.

Merkel, S., Voepel-Lewis, T., Shayevitz, J., and Malviya, S. (1996). The flacc: A behavioral scale for scoring postoperative pain in young children. *Pediatric nursing*, 23:293–7.

Merskey, H. and Bogduk, N. (1994). Classification of chronic pain; description of chronic pain syndromes and definitions of pain terms. *Task force on taxonomy of the International Association for the Study of Pain*, pages 41--43.

Nanni, L., Brahnam, S., and Lumini, A. (2010). A local approach based on a local binary patterns variant texture descriptor for classifying pain states. *Expert Systems with Applications*, 37(12):7888--7894.

Ng, A. Y. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Twenty-first international conference on Machine learning - ICML '04*. ACM Press.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 41.1--41.12.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621.

Reissland, N., Francis, B., and Mason, J. (2013). Can healthy fetuses show facial expressions of "pain" or "distress"? *PLoS ONE*, 8(6):e65530.

Reissland, N., Francis, B., Mason, J., and Lincoln, K. (2011). Do facial expressions develop before birth? *PLoS ONE*, 6(8):e24081.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Salekin, M. S., Zamzami, G., Goldgof, D. B., Kasturi, R., Ho, T., and Sun, Y. (2019a). Multi-channel neural network for assessing neonatal pain from videos. In *2019 IEEE*

International Conference on Systems, Man and Cybernetics, SMC 2019, Bari, Italy, October 6-9, 2019, pages 1551--1556.

Salekin, M. S., Zamzami, G., Paul, R., Goldgof, D. B., Kasturi, R., Ho, T., and Sun, Y. (2019b). Harnessing the power of deep learning methods in healthcare: Neonatal pain assessment from crying sound. *CoRR*, abs/1909.02543.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618--626.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929--1958.

Sweet, S. D. and McGrath, P. J. (1998). Physiological measures of pain. *Progress in Pain Research and Management*, 10:59--82.

van de Velde, M. and Buck, F. D. (2012). Fetal and maternal analgesia/anesthesia for fetal procedures. *Fetal Diagnosis and Therapy*, 31(4):201--209.

Yan, F., Dai, S.-Y., Akther, N., Kuno, A., Yanagihara, T., and Hata, T. (2006). Four-dimensional sonographic assessment of fetal facial expression early in the third trimester. *International Journal of Gynecology & Obstetrics*, 94(2):108--113.

Zamzami, G., Goldgof, D. B., Kasturi, R., and Sun, Y. (2018). Neonatal pain expression recognition using transfer learning. *CoRR*, abs/1807.01631.

Zamzami, G., Pai, C., Goldgof, D. B., Kasturi, R., Ashmeade, T., and Sun, Y. (2016a). An approach for automated multimodal analysis of infants' pain. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 4148--4153.

Zamzami, G., Pai, C., Goldgof, D. B., Kasturi, R., Sun, Y., and Ashmeade, T. (2016b). Machine-based multimodal pain assessment tool for infants: A review. *CoRR*, abs/1607.00331.

Zamzmi, G., Pai, C.-Y., Goldgof, D., Kasturi, R., Sun, Y., and Ashmeade, T. (2017). Automated pain assessment in neonates. In *Image Analysis*, pages 350--361. Springer International Publishing.

Zamzmi, G., Paul, R., Salekin, M. S., Goldgof, D., Kasturi, R., Ho, T., and Sun, Y. (2019). Convolutional neural networks for neonatal pain assessment. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(3):192--200.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818--833.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878.

Zhang, P., Zhong, Y., Deng, Y., Tang, X., and Li, X. (2019). A survey on deep learning of small sample in biomedical image analysis. *CoRR*, abs/1908.00473.

Zhi, R., Zamzmi, G., Goldgof, D., Ashmeade, T., and Sun, Y. (2018). Infants' pain recognition based on facial expression: Dynamic hybrid descriptions. *IEICE Transactions on Information and Systems*, E101.D(7):1860--1869.

Zimmermann, M. (1991). Pain in the fetus: neurobiological, psychophysiological and behavioral aspects. *Der Schmerz*, 5(3):122--130.