

# ADVERSARIAL DETECTION OF COUNTERFEITED PRINTABLE GRAPHICAL CODES: TOWARDS "ADVERSARIAL GAMES" IN PHYSICAL WORLD

Olga Taran, Slavi Bonev, Taras Holotyak and Slava Voloshynovskiy

University of Geneva, Department of Computer Science, Geneva, Switzerland

## ABSTRACT

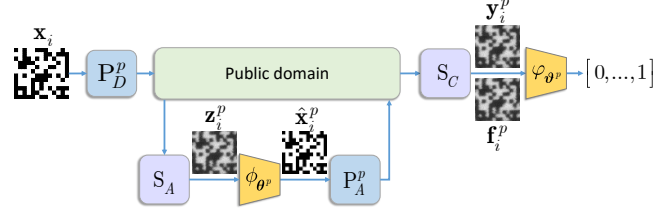
This paper addresses a problem of anti-counterfeiting of physical objects and aims at investigating a possibility of counterfeited printable graphical code detection from a machine learning perspectives. We investigate a fake generation via two different deep regeneration models and study the authentication capacity of several discriminators on the data set of real printed graphical codes where different printing and scanning qualities are taken into account. The obtained experimental results provide a new insight on scenarios, where the printable graphical codes can be accurately cloned and could not be distinguished.

**Index Terms**— Printable graphical codes, clonability attacks, adversarial discriminators, machine learning.

## 1. INTRODUCTION

Counterfeiting of physical objects is a very important problem for modern economies. The counterfeited products can be danger for life, like for example pharmaceutical, life-care products, etc., lead to market loss and damage of brands' reputation. The World Customs Organization in 2005 claimed that nearly 25% of pharmaceutical products in developing countries are forgeries [1].

Nowadays, there exist a lot of different techniques to protect the original products against falsification. However, it is crucial to guarantee non-clonability of these protection elements. One well known technology that is claimed to be robust to the clonability attacks is based on *Physical Unclonable Functions (PUFs)*. The main idea behind *PUFs* consists in the natural randomness of microstructure of physical objects [2, 3]. The *PUFs* are considered as non-clonable digital signature of substrate. This direction is gaining more popularity with increasing the abilities of modern mobile phones. However, in general, the verification process is quite sensitive to the illumination and may require to use a special equipment, like for example a portable microscope or attachable lenses. Another quite popular nowadays technology is based on a so called *anti-copying patterns*. The *anti-copying patterns* consist of a high-density black and white texture pat-



**Fig. 1:** A life cycle of the PGC:  $\mathbf{x}_i \in \{0, 1\}^{n \times m}$  is an original code;  $P_D^p$  and  $P_A^p$  are the defender's and attacker's printing processes, correspondingly;  $S_A$  and  $S_C$  are the attacker's and customer's digitization procedure respectively;  $\mathbf{z}_i^p \in \mathbb{R}^{n \times m}$  original codes digitized by the attacker;  $\phi_{\theta^p}$  is an attacker's re-generation model;  $\hat{\mathbf{x}}_i^p \in \{0, 1\}^{n \times m}$  is an attacker's estimation of the  $\mathbf{x}_i$ ;  $\mathbf{y}_i^p$  and  $\mathbf{f}_i^p \in \mathbb{R}^{n \times m}$  are original and fake codes from the public domain;  $\phi_{\theta^p}$  is a discrimination model.

terns. The verification process usually doesn't require any special equipment. Quite often, *anti-copying patterns* are injected into the traditional 2D codes, like for example Quick Response (QR) codes [4] or DataMatrix codes [5]. Obtained in such a way codes referred to as *Printable Graphical Codes (PGC)* are claimed to be unclonable under hand-crafted attack [6, 7]. The robustness of these codes to the machine learning based attacks is reported in [8, 9, 10]. In contrast to a common belief about the non-clonability of PGCs, [10] demonstrated a possibility to estimate digital codes from their printed counterparts in certain cases by using the deep neural networks (DNN) composed of fully connected layers. Extending [10], this paper aims at investigating the possibilities of the convolutional DNN for cloning. In contrast to [10], where only the non-DNN based discriminators were studied, this work aims at investigating the robustness of the DNN based discriminators to the clonability attacks.

The main contributions of this paper are:

- We investigate the impact of the used scanner on the quality of the produced clones on two new data sets.
- We investigate the possibilities of the attacker to clone modern PGC with a high quality by using convolutional DNN and we compare the obtained results with those in [10].
- We compare the robustness of the non-DNN and DNN based adversarial discriminators to the clonability attacks.

S. Voloshynovskiy is a corresponding author. The research was supported by the SNF project No. 200021\_182063.


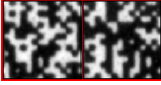
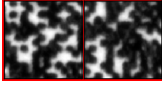
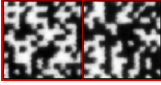
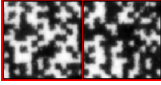

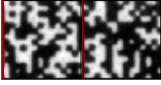
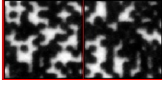
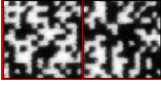
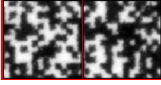
Data set	Original	Laser printers		Inkjet printers	
		SA	LX	CA	HP
DP1E					
DP1C					

Table 1: Examples of the codes from DP1E and DP1C data sets.

## 2. PROBLEM FORMULATION

A life cycle of the PGC is shown in Fig.1. It starts from printing  $P_D^p$  of the original binary codes  $\{\mathbf{x}_i\}_{i=1}^M$  by a defender (manufacturer) with a printing technology  $p = \{1, \dots, P\}$ . The printed codes are going to the public domain. The attacker aims at obtaining an accurate estimation of the original binary codes  $\{\hat{\mathbf{x}}_i^p\}_{i=1}^{K^p}$ ,  $K^p \leq M$ . For this he digitizes the original printed codes using a high resolution scanner  $S_A$  and then the obtained codes  $\{\mathbf{z}_i^p\}_{i=1}^{K^p}$  are processed via a deep mapper  $\phi_{\theta^p}$ . The obtained estimations  $\{\hat{\mathbf{x}}_i^p\}_{i=1}^{K^p}$  are printed using corresponding printing technology  $p$  (the question of the printing technology estimation is out of scope of this paper). The produced fakes are distributed in the public domain. A customer verification consists of the digitization  $S_C$  of the codes from the public domain (using some scanner or modern mobile phones) and their authentication through a discriminator  $\varphi_{\theta^p}$  that can produce either hard decision 0 or 1 (fake / authentic) or soft one ranging from 0 to 1.

In this paper we consider the worst case assuming that, besides the publicly available printed codes, the attacker has an access to the corresponding original binary codes  $\{\mathbf{x}_i\}_{i=1}^{K^p}$  to fully explore the power of training on the side of attacker. For the given pairs of original and scanned codes  $\{\mathbf{x}_i, \mathbf{z}_i^p\}_{i=1}^{K^p}$ , the attacker training procedure can be generalized as:

$$\hat{\theta}^p = \arg \min_{\theta^p} \sum_{i=1}^{K^p} \mathcal{L}^A(\mathbf{x}_i, \phi_{\theta^p}(\mathbf{z}_i^p)) + \lambda \Omega_{\theta^p}(\theta^p), \quad (1)$$

that consists in the optimization of the parameters  $\theta^p$  of the trained model  $\phi_{\theta^p}$  for the chosen printer  $p$  with respect to the given loss function  $\mathcal{L}^A(\cdot)$  and the regularization  $\Omega_{\theta^p}(\cdot)$  if any.

At the test stage, the scanned samples  $\{\mathbf{z}_i^p\}_{i=1}^{N^p}$  are passed through the pre-trained model  $\phi_{\theta^p}$ . The estimated codes  $\{\hat{\mathbf{x}}_i^p\}_{i=1}^{N^p}$  printed on the corresponding equipment are introduced to the public domain. The fake codes digitized by the customer scanner technology  $S_C$  we will denote as  $\{\mathbf{f}_i^p\}_{i=1}^{N^p}$ .

The defender provides a discriminator model  $\varphi_{\theta^p}$  to the customers. In [10] the authors considered a situation when the defender does not have an information advantage over the attacker. In this work we assume that, besides an access to the original  $\{\mathbf{x}_i\}_{i=1}^M$  and printed  $\{\mathbf{y}_i^p\}_{i=1}^M$  codes, the defender has an access to the fake codes  $\{\mathbf{f}_i^p\}_{i=1}^{N^p}$ ,  $N^p \leq K^p$ , repro-

duced by the attacker with the printing technologies  $p$  similar to those used by the defender. The discriminator model  $\varphi_{\theta^p}$  aims at distinguishing the authentic codes  $\mathbf{y}_i^p$  from the fakes  $\mathbf{f}_i^p$ . The training procedure can be generalized as:

$$\hat{\theta}^p = \arg \min_{\theta^p} \sum_{i=1}^{N^p} \alpha_1 \mathcal{L}_1(l(\mathbf{y}_i^p), \varphi_{\theta^p}(\mathbf{y}_i^p)) + \alpha_2 \mathcal{L}_2(l(\mathbf{f}_i^p), \varphi_{\theta^p}(\mathbf{f}_i^p)) + \beta \Omega_{\theta^p}(\theta^p), \quad (2)$$

where  $\varphi_{\theta^p}(\cdot)$  is the trained discriminator model with the parameters  $\theta^p$ .  $\mathcal{L}_1(\cdot)$  and  $\mathcal{L}_2(\cdot)$  are loss functions for the original and fake codes, respectively. For the classical DNN classifier,  $\mathcal{L}_1(\cdot)$  and  $\mathcal{L}_2(\cdot)$  are identical. In case the more advanced models, like for example GAN,  $\mathcal{L}_1(\cdot)$  and  $\mathcal{L}_2(\cdot)$  might differ. Usually,  $l(\cdot)$  produces code label but it might depend on the chosen loss functions.  $\Omega_{\theta^p}$  denotes a regularizer on the model parameters, if any.  $\alpha_1$ ,  $\alpha_2$  and  $\beta$  are the constants.

At the test stage, the sample  $\mathbf{y}_i^p$  or  $\mathbf{f}_i^p$  is passed through the discriminator  $\varphi_{\theta^p}$  that makes either hard decision 0 or 1 (fake / authentic) or soft one ranging from 0 to 1.

## 3. DATA SETS DETAILS

### 3.1. PGC data sets

In our work, we do not target to investigate the clonability of some particular PGC. We aim at demonstrating a general approach applicable to the majority PGC designed with identical modulation principles. In this respect, we use the *DataMatrix* code that is an open international ISO/IEC standard [5]. To evaluate the clonability aspects of the PGC codes based on the *DataMatrix* modulation and to investigate the influence of printing and scanning technologies we create two data sets: **DP1E** and **DP1C**. For both data sets we use 4 digital printers: 2 inkjet printers HP OfficeJet Pro 8210 (*HP*) and Canon PIXMA iP7200 (*CA*) and 2 laser printers Lexmark CS310 (*LX*) and Samsung Xpress 430 (*SA*). As a scanner, for **DP1C** we use Canon CanoScan 9000F at 1200 ppi and Epson Perfection V850 at 1200 ppi for **DP1E**. Each data set consists of 384 original codes  $\mathbf{x}_i \in \{0, 1\}^{384 \times 384}$  generated based on the *DataMatrix* standard and printed and scanned codes

Train on \ Test on	SA	LX	CA	HP
<i>LinearBN</i>				
SA	<b>0.14</b>	0.97	0.12	0.32
LX	0.30	<b>0.36</b>	0.19	0.37
CA	0.25	1.09	<b>0.12</b>	0.34
HP	0.35	1.67	0.17	0.40
<i>ConvBN</i>				
SA	<b>0.10</b>	0.60	0.17	0.34
LX	0.52	<b>0.13</b>	0.44	0.84
CA	0.24	1.06	<b>0.10</b>	0.31
HP	0.19	1.64	0.10	<b>0.18</b>

**Table 2: DP1C:** regeneration error (normalized *Hamming distance*) between the original  $\mathbf{x}_i$  and regenerated  $\hat{\mathbf{x}}_i$  codes.

$\mathbf{y}_i^p \in \mathbb{R}^{384 \times 384}$ ,  $p = \{\text{SA, LX, CA, HP}\}$ . The visualisation of the sub-blocks of size  $84 \times 84$  from several codes for both data sets are given in the Table 1.

## 4. TRAINING DETAILS <sup>1</sup>

### 4.1. Attacker’s re-generation model

For  $\phi_{\theta^p}$  training, the pairs  $\{\mathbf{x}_i, \mathbf{y}_i^p\}_{i=1}^{384}$  were split into *training* (100 images), *validation* (50 images) and *test* (234 images) sub-sets. In each sub-set the codes were split into non-overlapping blocks of size  $24 \times 24$ . The final size of the *training* set is 25 600 sub-images, the *validation* set contains 12 800 sub-images and the *test* set consists of 59 904 sub-images.

As a re-generation model  $\phi_{\theta^p}$  we use two types of DNN architectures based on a ”bottleneck” principle [11]: (1) *LinearBN*: linear model with the input size equals to 576 and 6 hidden layers of size 256, 128, 36 128, 256, 576 (similar to *BN* in [10] and used as a base line); (2) *ConvBN*: convolutional model with the input size equals to  $24 \times 24$  and 7 hidden layers:  $22 \times 22 \times 64$ ,  $10 \times 10 \times 32$ ,  $8 \times 8 \times 16$ ,  $4 \times 4 \times 8$ ,  $8 \times 8 \times 16$ ,  $12 \times 12 \times 32$  and  $24 \times 24$ . For both models as activation functions we use *ReLU* and *Tanh*. In the *LinearBN* model after each layer we apply 1D batch normalization and 2D batch normalization for the *ConvBN*. As the loss function  $\mathcal{L}^A(\cdot)$  the *MSE* is used. The training procedure is blind in the sense that we don’t use any information about the principles of the *DataMatrix* code generation.

### 4.2. Defender’s discrimination models

For  $\varphi_{\theta^p}$  training the triplets  $\{\mathbf{x}_i, \mathbf{y}_i^p, \mathbf{f}_i^p\}_{i=1}^{384}$  were split into *training* (100 images), *validation* (50 images) and *test* (234 images) sub-sets. In each sub-set the codes were split into non-overlapping blocks of size  $64 \times 64$ . The final size of the *training* set is 3 600 sub-images, the *validation* set contains 1 800 sub-images and the *test* set consists of 8 424 sub-images.

Train on \ Test on	SA	LX	CA	HP
<i>NN LinearBN</i>				
SA	<b>0.14</b>	1.47	0.20	0.56
LX	0.35	<b>0.74</b>	0.33	0.70
CA	0.17	1.21	<b>0.13</b>	0.50
HP	0.16	1.20	0.15	<b>0.42</b>
<i>NN ConvBN</i>				
SA	<b>0.12</b>	1.16	0.24	0.62
LX	0.53	<b>0.23</b>	0.78	1.25
CA	0.15	0.79	<b>0.09</b>	0.41
HP	0.12	0.81	0.09	<b>0.17</b>

**Table 3: DP1E:** regeneration error (normalized *Hamming distance*) between the original  $\mathbf{X}$  and regenerated  $\hat{\mathbf{X}}$  codes.

As a base line discrimination model we use (a) normalized *Hamming distance* between the original binary codes and scanned codes binarized via function  $T_{\theta^p}(\cdot)$  with an optimal threshold  $\theta^p$  estimated on the training sub-set; (b) *Person correlation* between the original binary codes and scanned codes. Also we investigate two DNN-based discrimination models:

1. *DNN*-based 2 class *classifier* with 5 convolutional layers:  $32 \times 32 \times 218$ ,  $16 \times 16 \times 256$ ,  $8 \times 8 \times 512$ ,  $4 \times 4 \times 1024$ ,  $1 \times 1$ . We use 2D batch normalization after each layer except the first and the last ones. As activation functions we use *LeakyRelu* and *Sigmoid* for the last layer. As a loss one *Binary Cross Entropy* is used. The training procedure formulated by the equation (2) can be simplified as:

$$\hat{\theta}^p = \arg \min_{\theta^p} - \frac{1}{M} \sum_{i=1}^M l(\mathbf{b}_i^p) \log(\varphi_{\theta^p}(\mathbf{b}_i^p)) + (1 - l(\mathbf{b}_i^p)) \log(1 - \varphi_{\theta^p}(\mathbf{b}_i^p)), \quad (3)$$

where  $M = M_y + M_f$  is a size of training sub-set,  $\mathbf{b}_i^p$  corresponds to the  $\mathbf{y}_i^p$  or  $\mathbf{f}_i^p$ ,  $l(\cdot)$  produces the corresponding code label.

2. *GAN*-like discriminator with a model’s architecture identical to that used for *DNN*-based *classifier* discriminator. Two *Binary Cross Entropy* functions are used for training:

$$\hat{\theta}^p = \arg \min_{\theta^p} - \left( \frac{1}{M_y} \sum_{i=1}^{M_y} \log(\varphi_{\theta^p}(\mathbf{y}_i^p)) + \frac{1}{M_f} \sum_{i=1}^{M_f} \log(1 - \varphi_{\theta^p}(\mathbf{f}_i^p)) \right), \quad (4)$$

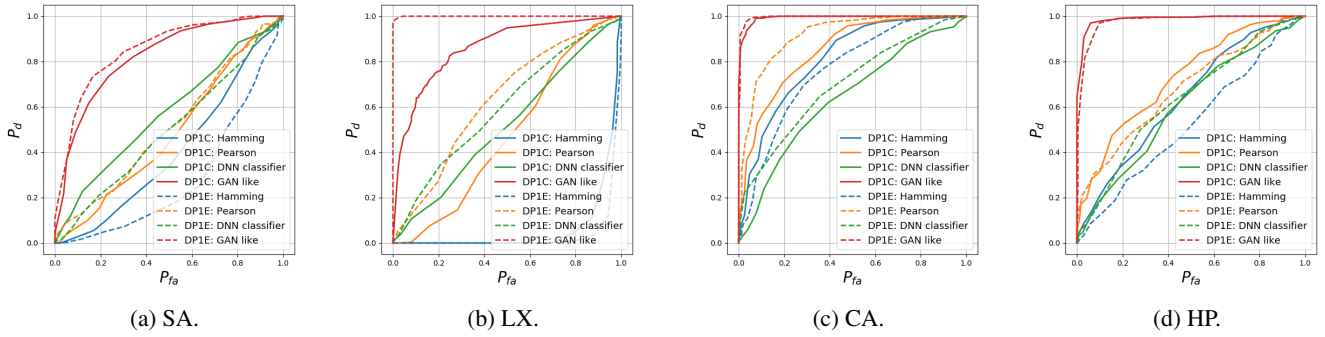
where  $l(\mathbf{y}_i^p) = 1$  and  $l(\mathbf{f}_i^p) = 0$ .

## 5. RESULTS AND DISCUSSION

### 5.1. Regeneration results

The first set of our experiments aims at investigating the regeneration accuracy of the attacker’s models. We train both

<sup>1</sup><https://github.com/tarano/adversarial-detection-of-counterfeited-pgc>



**Fig. 2:** The ROC curves.  $P_d$  denotes to the probability of the correct detection and  $P_{fa}$  is the probability of false acceptance.

*LinearNN* and *ConvNN* regeneration models on the **DP1C** and **DP1E** data sets and perform the cross-printer test. As a regeneration accuracy measure we use normalized *Hamming distance* between the original  $\mathbf{x}_i$  and regenerated codes  $\hat{\mathbf{x}}_i$ . The obtained results are given in the Tables 2 - 3. For both data sets, the *ConvNN* gives smaller regeneration error. The cross-printer test shows that the training and test for the same printers are preferable and give smaller regeneration error.

From the Table 1 it can be seen that the *LX* printer has the higher dot gain in contrast to the other used printers. Due to the difference in the illumination between Epson and Cannon scanners the regeneration error in **DP1E** data set is approximately in 2 times bigger than in **DP1C** data set.

To produce the fake codes  $\mathbf{f}_i^p$  we print the estimated binary codes  $\hat{\mathbf{x}}_i^p$  regenerated via *ConvNN* model (due to the smaller *Hamming distance* error) on the corresponding printers.

## 5.2. Authentication results

To evaluate the authentication efficiency of the defender’s discrimination models we compute the ROC curves similarly to [10]. The obtained ROCs are shown in the Figure 2. It is easy to see that the discriminator based on the normalized *Hamming distance* (blue curves) doesn’t provide an efficient authentication. The obtained curves are very close to the diagonal which means that the authentication is similar to a random guess. In case of the *LX* the  $P_{fa}$  is very high that can be explained by the big amount of errors even in the original printed codes due to the big printing dot gain. For the discriminator based on the *Pearson correlation* one can observe similar behaviour. In case of the discriminator trained as a classical 2 class DNN classifier the situation doesn’t improve.

The results obtained for the *GAN-like* discriminator show completely different picture. In case of the *CA* and *HP* the authentication ROCs show  $P_{fa}$  around 0.1 - 0.15 for the  $P_d$  close to 1. In case of the *SA* the obtained ROCs are not so ideal but much better than the previous results. In case of the *LX* one can observe a big difference between the ROCs obtained for **DP1E** and **DP1C**. This can be explained by the fact that in

the **DP1E**, due to the brighter illumination, the regeneration error is bigger and the produced fakes have bigger amount of errors. The recognition of such fakes is more efficient.

In summary, it should be noted that the attacker benefits from the codes printed with intermediate printing quality where the accurate regeneration is still possible but after printing the amount of errors are approximately on the same level as in the original codes due to the small printing imperfections and artefacts. The defender in opposite benefits from the either perfect printing quality or quite bad one. However, both these cases can be dangerous for the defender. In the first case, the attacker can estimate the original codes without any mistakes and the defender would not be able to detect the fakes under the assumption that the printing-scanning quality of the attacker no less than the defender’s one. In the second case, the amount of errors even in the original codes might be too high to distinguish the fakes.

## 6. CONCLUSIONS

In this paper, we investigate the robustness to the machine learning based copy attacks of the modern printable codes using *DataMatrix* modulation typical for many *PGC* designs. We test proposed framework under assumption that the defender has an advantage over the attacker by having an access to the fake codes. We empirically proved a possibility to accurate estimate the printed codes for high quality printers even from relatively small training data sets. We demonstrate that the high quality fakes is not the end of the story. The defender can benefit from the achievements of the modern machine learning and can use them to train more efficient discriminator.

For future work we aim at investigating the possibilities of the modern mobile phones for the detection of fake codes and to compare the abilities of machine learning approaches versus hand-crafted attacks. The impact of the number of training examples and training from the original digital templates are also among our future priorities.

## 7. REFERENCES

- [1] WCO, “Global congress addresses international counterfeits threat immediate action required to combat threat to finance/health, 2005,” .
- [2] Sviatoslav Voloshynovskiy, Patrick Bas, and Taras Holotyak, “Physical object authentication: detection-theoretic comparison of natural and artificial randomness,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, Shanghai, China, March,20-25 2016.
- [3] Sviatoslav Voloshynovskiy, Maurits Diephuis, Fokko Beekhof, Oleksiy Koval, and Bruno Keel, “Towards reproducible results in authentication based on physical non-cloneable functions: The forensic authentication microstructure optical set (famos),” in *Proceedings of IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, December 2–5 2012.
- [4] *ISO/IEC 18004: Information Technology - Automatic identification and data capture techniques-Bar Code Symbology-QR Code. 2000, 2000.*
- [5] *ISO/IEC 16022: Information technology - Automatic identification and data capture techniques - Data Matrix bar code symbology specification, 2006.*
- [6] Justin Picard, “Digital authentication with copy-detection patterns,” in *Optical Security and Counterfeit Deterrence Techniques V*. International Society for Optics and Photonics, 2004, vol. 5310, pp. 176–184.
- [7] Bernhard Wurnitzer and Slavtcho Bonev, “Matrix print data storage and method for encoding the data,” Aug. 21 2008, US Patent App. 11/572,591.
- [8] Cléo Baras and François Cayre, “Towards a realistic channel model for security analysis of authentication using graphical codes,” in *Information Forensics and Security (WIFS), 2013 IEEE International Workshop on*. IEEE, 2013, pp. 115–119.
- [9] Anh Thu Phan Ho, Bao An Mai Hoang, Wadih Sawaya, and Patrick Bas, “Document authentication using graphical codes: impacts of the channel model,” in *Proceedings of the first ACM workshop on Information hiding and multimedia security*. ACM, 2013, pp. 87–94.
- [10] Olga Taran, Slavi Bonev, and Slava Voloshynovskiy, “Clonability of anti-counterfeiting printable graphical codes: a machine learning approach,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019.
- [11] Frantisek Grézl, Martin Karafiát, Stanislav Kontár, and Jan Cernocky, “Probabilistic and bottle-neck features for lvsr of meetings,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07*. IEEE, 2007, vol. 4, pp. IV–757.