Régularisation entropique des vecteurs de caractéristiques d'un réseau de neurones pour une meilleur transférabilité

Raphael BAENA Lucas DRUMETZ Vincent GRIPON
IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, France

Résumé – Cet article étudie la classification de données à partir d'un apprentissage sur une approximation de leurs étiquettes. Par exemple, un problème de régression peut être discrétisé pour obtenir un problème de classification plus simple à résoudre. Un problème rencontré avec l'utilisation de l'entropie croisée comme critère d'entraînement dans cette situation est le surapprentissage des caractéristiques focalisant trop sur l'étiquetage grossier. Pour résoudre ce problème, nous introduisons une régularisation entropique appliquée à l'espace de caractéristiques du modèle. Ces caractéristiques peuvent ensuite être utilisées pour résoudre une tâche raffinée avec une erreur moindre. La validité de notre méthode est soutenue à la fois par une analyse théorique et des expériences.

Abstract – This article discusses the problem of classification when learning on an approximation of the labels. For example, a regression problem can be discretized to obtain a classification problem that is easier to solve. One problem encountered when training with the cross-entropy is the overfitting of the features that occurs with coarse labeling. To address this problem, we introduce an entropic regularization applied to the feature space to increase features' diversity. These features can be later used to solve a refined task with a smaller error. The validity of our method is supported by both theoretical analysis and experiments.

1 Introduction

Résoudre des problèmes de classification ou de régression peut s'effectuer en apprenant directement sur des étiquettes détaillées, appelées étiquettes raffinées dans le reste de cet article, ou sur des étiquettes grossières qui sont des approximations [20, 15]. Ce dernier cas peut être préférable lorsque les étiquettes raffinées sont difficiles à obtenir par exemple. Par ailleurs, certains problèmes de classification peuvent être considérés comme la discrétisation d'un problème de régression sous-jacent plus complexe [6].

Dans cet article, nous cherchons à montrer qu'à partir d'un apprentissage sur des étiquettes approximatives (grossières), il est possible de retrouver partiellement de l'information sur les étiquettes raffinées. Ce phénomène met en évidence la capacité des algorithmes d'apprentissage à généraliser au-delà d'une tâche spécifique et à aborder des problèmes plus subtils. Cette capacité n'est cependant pas toujours garantie. En effet, à mesure que le modèle s'entraîne sur des étiquettes grossières, sa généralisation sur les étiquettes raffinées peut se détériorer en raison d'un surapprentissage. Ce phénomène est souvent observé dans le cadre de l'apprentissage par transfert, en particulier avec les problèmes avec peu d'exemples, où l'arrêt précoce de l'entraînement peut conduire à de meilleures performances. Une explication possible est proposée dans [22] où les auteurs décrivent apprentissage en deux phases ; la deuxième phase où l'information mutuelle entre l'espace des caractéristiques et l'entrée est réduite tandis que l'accent est mis sur la tâche d'entraînement.

En nous appuyant sur cette observation, nous avons réalisé une expérience pour montrer l'impact d'un apprentissage sur des étiquettes grossières sur la prédiction des étiquettes raffinées. L'expérience est conduite sur un jeu de données d'estimation d'âges [20] où la tâche consiste à prédire l'âge d'une personne à partir de la photo de son visage. L'apprentissage est réalisé sur des plages d'âges qui composent ainsi les étiquettes grossières. En parallèle, nous essayons de prédire les âges exacts (étiquettes raffinées). Nos résultats représentés sur la Figure 1 ont montré une forte relation entre l'entropie de l'espace des caractéristiques et l'erreur quadratique moyenne de la tâche de régression. Dans une première phase, l'erreur quadratique moyenne de régression diminue avec le taux d'erreur de classification. Cependant, dans une deuxième phase, plus longue, l'erreur quadratique moyenne de régression atteint un minimum avant d'augmenter tandis que le taux d'erreur de classification reste stable, illustrant parfaitement le comportement décrit dans [22].

Sur la base de ces observations, nous introduisons dans cet article une régularisation entropique de l'espace de caractéristiques d'architectures profondes. En effet, sur plusieurs expériences, les modèles entraînés avec cette régularisation montrent de meilleures capacités de généralisation sur des tâches plus subtiles que celle ayant servi d'entraînement. Remarquons aussi qu'il n'est plus nécessaire de compter sur un arrêt précoce de l'entraînement, comme montré sur la Figure 1. Nous mettons à disposition notre code `https://github.com/raphael-baena/FIERCE-repo`.

2 État de l'art

2.1 Apprentissage avec l'entropie croisée

Dans ce papier nous considérons uniquement des architectures profondes reconnues pour atteindre l'état de l'art dans le domaine de la vision. On peut écrire une architecture profonde comme une fonction $f_{\boldsymbol{\theta}} : \mathbf{x} \mapsto \mathbf{y}$, où $\mathbf{x}$ est l'entrée et $\mathbf{y}$ la prédiction de l'étiquette associée. Cette fonction est généralement obtenue par la composition de fonctions plus simples appelées couches [5], donnant lieu à des architectures composites. Dans le contexte de l'apprentissage supervisé, il est

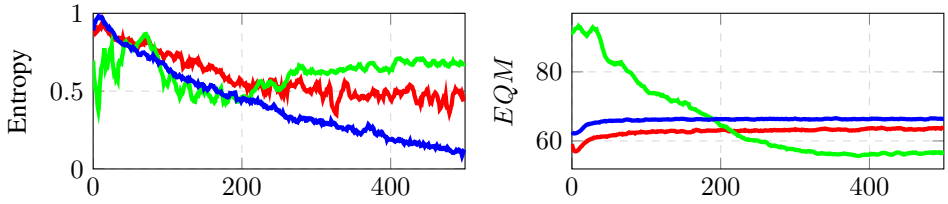

FIGURE 1 : Évolution de l'entropie des caractéristiques et l'erreur quadratique moyenne (EQM) sur le jeu d'estimation d'âges [20] : Entropie Croisée (rouge), FIERCE (méthode proposée) (vert), Lissage d'Etiquettes (bleu). Les modèles entraînées avec l'entropie croisée ou le lissage d'étiquettes atteignent une EQM minimum avant de se stabiliser à une valeur plus grande. A l'opposée, FIERCE atteint une EQM plus faible maintenue durant l'apprentissage. Cette capacité semble reliée à l'entropie des caractéristiques.

d'usage d'appeler *caractéristiques* le vecteur $\mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x})$ obtenu en sortie de l'avant-dernière couche de l'architecture. La prédiction des étiquettes résulte classiquement de l'application d'un classificateur linéaire sur les caractéristiques.

Les paramètres de l'architecture profonde sont appris sur un ensemble d'entraînement $\mathcal{D}_{entrainement}$ avec pour objectif de généraliser correctement sur d'autres données. L'entraînement repose sur la minimisation d'un critère $\mathcal{L}$. Le choix du critère a donc un impact direct sur les performances de l'architecture [13], l'espace des caractéristiques et la distribution de sortie [18, 10, 7, 16]. Parmi les différents critères utilisés en pratique, l'entropie croisée est particulièrement populaire dans le cadre de problèmes de classification [5].

Cependant, la minimisation de l'entropie croisée avec une descente de gradient stochastique (ou ses variantes) a tendance à éliminer les caractéristiques considérées comme peu informatives pour la tâche considérée. Les auteurs de [22] ont observé un phénomène d'apprentissage en deux phases successives. Lors de la première phase, l'information mutuelle $I(\mathbf{y}, \mathbf{r})$ entre les caractéristiques $\mathbf{r}$ et la sortie $\mathbf{y}$ augmente. Au cours de la deuxième phase, beaucoup plus longue, l'information mutuelle $I(\mathbf{x}, \mathbf{r})$ entre l'entrée et les caractéristiques diminue. En d'autres termes, l'architecture trouve d'abord les caractéristiques sur lesquelles elle peut prédire les étiquettes, avant d'exceller sur la tâche d'entraînement en compressant les caractéristiques pour de ne conserver que les plus pertinentes.

De même, dans [7], les auteurs ont montré que plus l'erreur de classification est réduite, plus la confiance en la prédiction est surestimée. Une méthode pour amoindrir ce phénomène consiste à ajouter une régularisation $\mathcal{R}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$ au critère d'entraînement [7]. Les régularisations comprennent une large gamme de techniques agissant sur les paramètres de l'architecture, comme la normalisation des échantillons [12], la pénalisation des poids [8] ou encore le *dropout* [11].

2.2 Régularisation entropique sur la sortie

Certaines régularisations s'appliquent directement sur la distribution de sortie, par exemple : la pénalisation de la confiance [18], le lissage des étiquettes [23] ou la distillation [10]. De nombreux auteurs ont montré qu'en réalité ces techniques reposent sur une forme de régularisation entropique appliquée à la distribution de sortie [18, 3, 16].

Cette régularisation est utilisée dans le cadre de l'apprentissage supervisé pour obtenir des distributions de sortie plus lisses [18, 3]. Elle impacte également la géométrie de l'espace des caractéristiques [21, 17]. Cependant, les bénéfices de ces impacts ne sont pas évidents. En effet, les auteurs de [21] ont montré que le lissage des étiquettes peut supprimer la spécificité des caractéristiques des individus au sein d'une même classe. Cette perte d'information a suscité une controverse sur l'utilisation du lissage des étiquettes dans le cadre de la distillation [17]. Plus récemment, les auteurs de [14] ont montré que les critères d'entraînement conduisant aux meilleures performances ne sont pas nécessairement synonymes de bonnes performances en transfert. Nous montrons un résultat similaire dans nos expériences.

Contrairement à ces travaux, nos proposons une régularisation entropique sur l'espace des caractéristiques (au lieu de la sortie) et nous motivons cette proposition dans la section théorique et expérimentale.

3 Impact de la régularisation sur les caractéristiques

Dans le cadre de la classification, une architecture profonde transforme d'abord les entrées en caractéristiques (idéalement linéairement séparables). Ces caractéristiques sont ensuite classifiées avec une régression logistique. Pour simplifier les équations, nous considérons un cadre bayésien. Nous donnons les calculs des équations suivantes dans [1]. Notons y la classe d'un échantillon, i.e, $y = \text{argmax}_i \mathbf{y}[i]$. La classification est donnée par $p(y = y_i|\mathbf{r})$ où $\mathbf{r}$ sont les caractéristiques par rapport à $\mathbf{x}$: $p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x})$ inféré par le réseau.

Entropie croisée : Avec une approche bayésienne, on peut calculer le gradient par rapport aux paramètres et obtenir :

$$\mathbb{E}_{\mathbf{x}}\left[-\int \nabla_{\boldsymbol{\theta}}(p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}))\log(p(y=y_i|\mathbf{r})d\mathbf{r}\right]. \qquad (1)$$

Notons que $\nabla_{\boldsymbol{\theta}}(p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}))$ est pondéré par $\log(p(y = y_i|\mathbf{r}))$, ce qui signifie que les caractéristiques les plus sélectives $\mathbf{r}$ pour la classe considérée ont plus de chances d'être échantillonnées.

Lissage d'étiquettes : Un calcul similaire est conduit avec le lissage d'étiquettes uniforme avec pour facteur $\sigma < 0,5$:

$$\mathbb{E}_{\mathbf{x}}\left[-\int \nabla_{\boldsymbol{\theta}}(p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}))[(1-\sigma)\log(p(y=y_i|\mathbf{r}) + \frac{\sigma}{c-1}\log(\prod_{j,j\neq i} p(y=y_j|\mathbf{r}))]d\mathbf{r}\right].$$

Le lissage d'étiquettes conduit à des termes supplémentaires $\frac{\sigma}{c-1}\log(\prod_{j,j\neq i} p(y = y_j|\mathbf{r}))$. Ces termes encouragent les caractéristiques les plus discriminantes associées aux autres classes j.

Par conséquent, le lissage d'étiquettes favorise certes la diversité, mais uniquement parmi les caractéristiques les plus sélectives. Celles menant à une précision pertinente, mais moindre ($p(y = y_i|\mathbf{r}) > 0,5$) ne seront donc pas encouragées. Ce résultat est cohérent avec [21, 17].

3.1 Régularisation entropique des caractéristiques pour promouvoir la diversité

Nous étudions l'utilisation de l'entropie des caractéristiques en tant que terme de régularisation. Cette régularisation est motivée après plusieurs expériences montrant la relation entre l'entropie de l'espace de caractéristiques et l'erreur sur la tâche de transfert. Nous montrons que la régularisation d'entropie sur les caractéristiques favorise la diversité. L'idée est d'ajouter l'entropie de l'espace de caractéristiques en tant que terme de régularisation :

$$\mathcal{L}^{Bayesian}_{Entropy}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \mathcal{L}_{CE} - \lambda H_{\boldsymbol{\theta}}(\mathbf{r}). \tag{2}$$

où $H_{\boldsymbol{\theta}}(\mathbf{r})$ est l'entropie estimée pour chaque lot de données et $\lambda > 0$ est un hyperparamètre. Le gradient de la régularisation entropique des caractéristiques est :

$$-\mathbb{E}_{\mathbf{x}}\left[\int \nabla_{\boldsymbol{\theta}} \left[-\lambda(\log(p_{\boldsymbol{\theta}}(\mathbf{r})) + 1)\right] d\mathbf{r}\right]. \tag{3}$$

Remarquons que si une caractéristique est très probable, i.e. $p(\mathbf{r})$ grand, le gradient sera pénalisé. En ajustant λ, les caractéristiques moins discriminantes sont conservées tant qu'elles restent pertinentes pour la tâche d'entraînement.

4 Expériences

Dans cette partie, nous montrons d'abord que l'entropie des caractéristiques est reliée à leur capacité de retrouver de l'information sur les étiquettes raffinées. Dans [1] nous fournissons plus de détails, ainsi que des expériences sur des problèmes avec peu d'exemples (few-shot) et sur la sensibilité de notre méthode aux hyperparamètres.

Implémentation de la régularisation : notre régularisation consiste à soustraire l'entropie des caractéristiques $\tilde{H}_{\boldsymbol{\theta}}(\tilde{\mathbf{r}})$ à l'entropie croisée. Malheureusement, il n'existe pas de moyen direct pour estimer de manière différentiable l'entropie de l'espace des caractéristiques. Nous proposons une manière d'approximer cette quantité de manière différentiable avec une loi catégorique (voir détails dans [1]).

4.1 Relation entre l'entropie de l'espace de caractéristiques et la capacité de transfert

En considérant un problème de régression transformé en un problème de classification, nous montrons expérimentalement que l'entropie de l'espace de caractéristiques est reliée à sa capacité de régression mesurée par l'Erreur Quadratique Moyenne (EQM).

Nous considérons un jeu de données d'estimation d'âges [20] composé d'images de visages. Le but est de prédire les âges correspondant aux différents visages. Ce jeu de données peut être transformé en un problème de classification binaire avec les étiquettes grossières suivantes $\mathbf{y} = \mathbf{1}_{age<36}$ où 36 est l'âge médian. Nous utilisons un Resnet-18 [9], excepté qu'à l'avant-dernière couche, nous moyennons les caractéristiques pour obtenir un espace de dimension 1. Nous espérons que cet espace préserve dans une certaine mesure la relation d'ordre pré-existante entre les âges. L'architecture est entraînée avec une descente de gradient stochastique (SGD).

Pour estimer les âges, nous calculons les caractéristiques de 10 000 échantillons. Ensuite, nous ordonnons les caractéristiques selon leur valeur. En supposant que la distribution de probabilité des âges est connue, nous estimons les âges à partir de la distribution des caractéristiques à l'aide d'un transport optimal 1-d [19].

Nous présentons l'évolution de l'EQM et de l'entropie des caractéristiques sur la figure 1. Nous observons que l'utilisation du lissage d'étiquettes ou de l'entropie croisée (sans régularisation) conduit à deux phases. L'erreur quadratique moyenne diminue tandis que la précision s'améliore. Puis, une deuxième phase beaucoup plus longue s'initie (vers l'époque 50) avec une erreur quadratique moyenne qui augmente tandis que l'erreur de classification reste stable. De manière similaire, l'entropie des caractéristiques atteint un maximum avant de diminuer ; cette diminution est négativement corrélée avec l'erreur quadratique moyenne de régression. Ces phases sont semblables à l'évolution de l'information mutuelle $I(\mathbf{r}, \mathbf{x})$ décrite dans [22], et motivent fortement notre méthode.

Nous observons que notre méthode fournit l'erreur quadratique la plus faible et démontre son efficacité à accroître et maintenir l'entropie des caractéristiques. Dans [1], nous montrons comment l'erreur quadratique et l'entropie se comportent en faisant varier les différents hyperparamètres de notre méthode. Nous remarquons que sur ce jeu de données le lissage d'étiquettes dégrade l'erreur de régression. L'EQM finale est supérieure avec l'entropie croisée et, par ailleurs, l'entropie des caractéristiques continue de diminuer. Dans [1], nous illustrons l'espace des caractéristiques de chaque critère. On peut y observer clairement que le lissage d'étiquettes augmente la zone d'incertitude de prédiction sans pour autant fournir d'informations pertinentes sur les étiquettes raffinées.

4.2 Distribution de sortie et transférabilité

Les techniques de régularisation entropique sur la distribution de sortie ont pour effet de lisser cette distribution [18, 23]. Au lieu d'avoir une sortie avec des pics aux valeurs 0 et 1, le lissage de la distribution permet d'avoir une plus grande gamme de valeurs. Une question naturelle est alors de savoir si cette gamme fournit des informations sur les étiquettes raffinées.

Il est possible que les valeurs de sortie donnent des informations sur les étiquettes raffinées ; par exemple, le degré de similitude entre chaque classe. Pour évaluer la validité de cette interprétation, nous considérons un ensemble de données où l'incertitude de sortie peut être facilement interprétée comme des étiquettes raffinées. À cet effet, nous prenons un jeu de données hyperspectrales (télédétection) [2]. Nous détaillons ce jeu de données dans [1]. Deux problèmes importants et interconnectés de l'imagerie hyperspectrale sont la classification sémantique supervisée des pixels et le démélange spectral [4]. En classification, on essaie d'attribuer une classe à chaque pixel. Le démélange peut être considéré comme un raffinement de la classification. Le but est alors de prédire la proportion de chaque matériau (appelée abondance) dans chaque pixel.

TABLE 1 : $EQMs$ calculées sur le jeu de données hyperspectral.

	CE	LS	FIERCE
EQM_{br}	0.186 ± 0.00	$\mathbf{0.066} \pm 0.10$	0.130 ± 0.01
EQM_{tr}	0.177 ± 0.3	0.02 ± 0.08	$\mathbf{0.006} \pm 0.00$

Nous transformons le problème de régression du démélange (étiquettes raffinées) en un problème de classification en utilisant les étiquettes grossières suivantes $\mathbf{y} = (\arg\max_i \mathbf{z}_i)$ où $\mathbf{z}_i$ est la proportion du matériau $i : \mathbf{z} \in [0, 1]^m, \sum_i (\mathbf{z}i) = 1$. Comme $\mathbf{y}$ est une proportion, on peut interpréter la sortie du réseau (après un *softmax*) comme les proportions de chaque matériau. Les performances EQM_{br} de cette prédiction sont évaluées sur l'image entière. Pour ce jeu de données, nous utilisons un réseau de neurones avec 2 couches cachées.

Nous calculons une autre mesure pour évaluer si les caractéristiques apprises peuvent être réutilisées afin d'effectuer la tâche de régression. Pour cela, nous figeons l'espace de caractéristiques et nous entraînons un régresseur logistique à résoudre le problème de démélange (etiquettes raffinées). Les performances de ce régresseur sont évaluées avec son erreur quadratique moyenne : EQM_{tr}.

Comme indiqué dans le Tableau 1, le lissage d'étiquette et notre méthode FIERCE ont l'EQM la plus faible lorsque la sortie est directement interprétée comme les étiquettes raffinées. Cependant, notons que FIERCE présente la plus grande capacité de transfert (EQM_{tr} la plus faible) : les caractéristiques apprises permettent de récupérer davantage d'informations sur les étiquettes raffinées en comparaison avec les autres critères. En considérant la différence entre les deux métriques (EQM brute et EQM de transfert), nous soupçonnons que la distribution de sortie n'est pas un indicateur pertinent de la capacité de transfert. Une discussion plus approfondie et des métriques supplémentaires sont fournies dans [1].

5 Conclusion

Dans cet article, nous avons introduit une nouvelle forme de régularisation entropique qui s'applique à l'espace des caractéristiques d'une architecture d'apprentissage profond. Son objectif est de fournir des caractéristiques plus diverses pouvant être réutilisées afin de résoudre d'autres tâches. À cette fin, l'entropie des caractéristiques est encouragée, empêchant le modèle de supprimer les caractéristiques même si elles ne sont pas les plus discriminantes.

Soutenus par des expériences et un développement théorique, nous avons démontré la capacité de notre méthode à prévenir la perte d'information causée par l'utilisation de l'entropie croisée ou du lissage d'étiquettes, conduisant ainsi à de meilleures performances sur des problèmes plus subtils dérivées de la tâche d'entraînement.

Références

[1] Raphael BAENA et al. "Preserving Fine-Grain Feature Information in Classification via Entropic Regularization". In : *arXiv preprint arXiv :2208.03684* (2022).

[2] Lucas DRUMETZ et al. "Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability". In : *IEEE Transactions on Image Processing* (2016).

[3] Abhimanyu DUBEY et al. "Regularizing Prediction Entropy Enhances Deep Learning with Limited Data". In : *Proceedings of the Neural Information Processing Systems (NIPS)*. 2017.

[4] Pedram GHAMISI et al. "Advances in hyperspectral image and signal processing : A comprehensive overview of the state of the art". In : *IEEE Geoscience and Remote Sensing Magazine* (2017).

[5] Ian GOODFELLOW et al. *Deep learning*. MIT press, 2016.

[6] Yves GRANDVALET et al. "Semi-supervised learning by entropy minimization". In : *Advances in neural information processing systems* 17 (2004).

[7] Chuan GUO et al. "On calibration of modern neural networks". In : *ICML*. 2017.

[8] Stephen HANSON et al. "Comparing biases for minimal network construction with back-propagation". In : *Advances in neural information processing systems* 1 (1988).

[9] Kaiming HE et al. "Deep residual learning for image recognition". In : *CVPR*. 2016, p. 770-778.

[10] Geoffrey HINTON et al. "Distilling the Knowledge in a Neural Network". In : *NIPS Deep Learning and Representation Learning Workshop*. 2015.

[11] Geoffrey E HINTON et al. "Improving neural networks by preventing co-adaptation of feature detectors". In : *arXiv preprint arXiv :1207.0580* (2012).

[12] Sergey IOFFE et al. "Batch normalization : Accelerating deep network training by reducing internal covariate shift". In : *ICML*. 2015.

[13] Katarzyna JANOCHA et al. "On loss functions for deep neural networks in classification". In : *arXiv preprint arXiv :1702.05659* (2017).

[14] Simon KORNBLITH et al. "Why do better loss functions lead to less transferable features ?" In : *Advances in Neural Information Processing Systems* 34 (2021).

[15] Xin LIU et al. "Agenet : Deeply learned regressor and classifier for robust apparent age estimation". In : *Proceedings of the IEEE ICCV Workshops*. 2015.

[16] Clara MEISTER et al. "Generalized Entropy Regularization or : There's Nothing Special about Label Smoothing". In : *arXiv preprint arXiv :2005.00820* (2020).

[17] Rafael MÜLLER et al. "When does label smoothing help ?" In : *Advances in neural information processing systems* 32 (2019).

[18] Gabriel PEREYRA et al. "Regularizing neural networks by penalizing confident output distributions". In : *ICLR* (2017).

[19] Gabriel PEYRÉ et al. "Computational optimal transport : With applications to data science". In : *Foundations and Trends® in Machine Learning* (2019).

[20] Rasmus ROTHE et al. "Deep expectation of real and apparent age from a single image without facial landmarks". In : *International Journal of Computer Vision* 126.2 (2018), p. 144-157.

[21] Zhiqiang SHEN et al. "Is label smoothing truly incompatible with knowledge distillation : An empirical study". In : *ICLR* (2021).

[22] Ravid SHWARTZ-ZIV et al. "Opening the black box of deep neural networks via information". In : *arXiv preprint arXiv :1703.00810* (2017).

[23] Christian SZEGEDY et al. "Rethinking the inception architecture for computer vision". In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.