UNITOR @ DANKMEMES: Combining Convolutional Models and Transformer-based architectures for accurate MEME management

Claudia Breazzano and **Edoardo Rubino** and **Danilo Croce** and **Roberto Basili**
University of Roma, Tor Vergata
Via del Politecnico 1, Rome, 00133, Italy
claudiabreazzano@outlook.it, edoardo.ru94@libero.it
{croce,basili}@info.uniroma2.it

Abstract

This paper describes the UNITOR system that participated to the "multimoDal Artefacts recogNition Knowledge for MEMES" (DANKMEMES) task within the context of EVALITA 2020. UNITOR implements a neural model which combines a Deep Convolutional Neural Network to encode visual information of input images and a Transformer-based architecture to encode the meaning of the attached texts. UNITOR ranked first in all subtasks, clearly confirming the robustness of the investigated neural architectures and suggesting the beneficial impact of the proposed combination strategy.

1 Introduction

In Social networks, the ways to express opinions evolved from simply writing a post to publishing more complex contents, e.g., the composition of images and texts. These multi-modal objects, if adhering to some specific social conventions and visual specifications, are called MEMEs. In particular, a MEME is a multi-modal artifact, manipulated by users, who combines intertextual elements to convey a message. Characterized by a visual format that includes images, text, or a combination of them, MEMEs combine references to current events or related situations and pop-cultural references to music, comics and films (Ross and Rivers, 2017). In this context, the multimoDal Artefacts recogNition Knowledge for MEMES (DANKMEMES) task is the first EVALITA (Basile et al., 2020) task for MEMEs recognition and hate speech/event identification in MEMEs (Miliani et al., 2020). This task is divided into three subtasks: in MEME Detection, system is required to determine whether an image is a MEME, according to the definition of (Shifman, 2013); in Hate Speech Identification the aim is to recognize if a MEME expresses an offensive message; finally, in Event Clustering the aim is to cluster MEMEs according to their referring topics.

In this work, we present the UNITOR system participating in all three subtasks. Since MEMEs convey their content through the multimodal combination of an image and a text, UNITOR implements a neural network which combines state-of-the-art architectures for Computer Vision and Natural Language Processing. In particular, Deep Convolutional Neural Networks, such as (He et al., 2016; Tan and Le, 2019) are used to encode visual information into dense embeddings and Transformer-based architectures, such as (Devlin et al., 2019; Liu et al., 2019) encode the meaning of the added overlaid captions. UNITOR then stacks a multi-layered network in order to effectively combine the evidences captured by both encoders, in the final classification.

The UNITOR system ranked first in each subtask, clearly confirming the robustness of the investigated neural architectures and suggesting the beneficial contribution of the proposed combination strategy. In the rest of the paper, in Section 2 the UNITOR system is described while Section 3 reports the experimental results.

2 UNITOR Description

CNNs for Image classification. Recent years demonstrated that Convolutional Neural Networks (CNNs) are able to achieve state-of-the-art results in image processing (Jiao and Zhao, 2019), by implementing deep and complex stackings of Convolutional layers, which capture different aspects of input images at different levels of the networks.

Among the investigated architectures, we first considered ResNET (He et al., 2016): this network is the first introducing Residual Learning to define very deep and effective CNNs. Several ResNET architectures are defined by stacking 50, 101, 152 up to 1001 layers of convolu-

tion layers and skip connectors: as a result, deeper networks achieved significant improvements of previous state-of-art in a wide plethora of image processing tasks. Moreover, we investigated the recently proposed EfficientNet (Tan and Le, 2019): unlike ResNET, this is not a real architecture, but it provides an automatic methodology to improve the performance of an existing CNN (such as ResNET) by tuning its depth, width and resolution dimensions. The adoption of this methodology led to the definition of 8 CNNs (namely EfficientNET-B0, EfficientNET-B1 up to EfficientNET-B7), each characterized by an increasing depth and width. They achieve impressive results by efficiently balancing the number of the parameters of the network. The tuning process of (Tan and Le, 2019) demonstrated that a network such as EfficientNet-B3 achieves higher accuracy than ResNeXt101 (Xie et al., 2016) in using 18x fewer neural operations. Regardless of the adopted networks, these are already trained in a classification task involving the recognition of thousands of object types in several millions of images, i.e. in the ImageNet dataset (Deng et al., 2009). This pre-training step enables the network to recognize many "basic entities" (such as people or animals) before being applied to a new task, e.g., MEME Detection. The customization to a new task is obtained just by replacing the last classification layer with a new one (sized based on the number of targeted classes) and by fine-tuning the entire architecture. It is worth noticing that, once the architecture is fine-tuned on the new down-stream task, it can be also used as an *Image Encoder*: the embeddings generated on the layer previous the classification one can be used as low-dimensional representations of input images. Most importantly, these embeddings are correlated with the down-stream task, as they are expected to lay in linearly separable sub-spaces (Goodfellow et al., 2016), where the final classifier is applied. In UNITOR these vectors are used to combine visual information with other evidences: in practice, they will be used in combination with the embeddings produced from the Transformer-based architectures (applied to texts) before being used in input to the final classifier.

Transformer-based Architectures for text classification. A MEME is a combination of visual information and the overlaid caption. In this work, we thus also investigated classifiers based on the text made available via OCR to the participants by the DANKMEME organizers. In particular, we adopt the approach proposed in (Devlin et al., 2019), namely Bidirectional Encoder Representations from Transformers (BERT). It provides an effective way to pre-train a neural network over large-scale collections of raw texts, and apply it to a large variety of supervised NLP tasks, here text classification. The building block of BERT is the Transformer element (Vaswani et al., 2017), an attention-based mechanism that learns contextual relations between words in a text. The pre-training stage is based on two auxiliary tasks, whose aim is the acquisition of an expressive and robust language and text model: the *Masked Language* model acquires a meaningful and context-sensitive representation of words, while the *Next Sentence Prediction* task captures discourse level information. In particular, this last task operates on text-pairs to capture relational information between them, e.g. between the consecutive sentences in a text. The straightforward application of BERT has shown better results than previous state-of-the-art models on a wide spectrum of natural language processing tasks. In (Liu et al., 2019) RoBERTa is proposed as a variant of BERT which modifies some key hyperparameters, including removing the next-sentence pre-training objective, and training on more data, with much larger mini-batches and learning rates. This allows RoBERTa to improve on the masked language modelling objective compared with BERT and leads to better down-stream task performances. We adopt here the fine-tuning process for sequence classification, where sequences correspond to texts extracted from images. The special token `[CLS]` is added as a first element of each input sentence, so that BERT associates it a specific embedding. This dense vector represents the entire sentence and is used in input to a linear classifier customized for the target classification task: in MEME Detection and Hate Speech Identification, two classes are considered, while in Event Clustering five classes reflect the target topics. During training, all the network parameters are fine-tuned. BERT and RoBERTa are pre-trained over text in English, and they are able to capture language models specific for this language. In order to apply these architectures in Italian, we investigate several alternative models, pre-trained using document collections in Italian or in multi-

ple languages. Among these models, AlBERTo (Polignano et al., 2019) is a BERT-based model pre-trained over the Twita corpus (Basile and Nissim, 2013) (made of millions of Italian tweets) while GilBERTo[1] and UmBERTo[2] are RoBERTa-based models pre-trained over the OSCAR corpus and the Italian version of Wikipedia, respectively. Among the multi-lingual models, we investigate multilingual BERT (mBERT) (Pires et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) which extends the corresponding pre-training over texts in more than 100 languages.

Regardless of the adopted Transformer-based architecture, we also investigated the adoption of additional annotated material to support the training of complex networks over very short texts extracted from MEMEs. In particular, in Hate Speech Identification, we used an external dataset which addressed the same task, but within a different source. We thus adopted a dataset made available within the Hate Speech Detection (HaSpeeDe) task (Bosco et al., 2018) which involves the automatic recognition of hateful contents in Twitter (HaSpeeDe-TW) and Facebook posts (HaSpeeDe-FB). Each investigated architecture is trained for few epochs only over on the HaSpeeDe dataset before the real training is applied to the DANKMEMES material. In this way, the neural model, which is not specifically pre-trained to detect hate speech, is expected to improve its "expertise" in handling such a phenomenon (even though using material derived from a different source) before being specialized on the final DANKMEMES task[3].

We trained UmBERTo both on HaSpeeDe-TW and on HaSpeeDe-FB and on the merging of these, too. Initial experiments suggested that a higher accuracy can be achieved only considering the material from Facebook (HaSpeeDe-FB). We suppose this is mainly due to the fact that messages from HaSpeeDe-FB and DANKMEMES share similar political topics. As for a CNN, once the Transformer-based architecture is fine-tuned on the new task, it can be used as text encoder, by removing the final linear classifier and selecting the embedding associated to the `[CLS]` token. These vectors will be used in UNITOR in combination with the embeddings derived from the CNN architecture, as described hereafter.

Combining visual and semantic evidences. UNITOR adopts an approach similar to the Feature Concatenation Model (FCM) already seen in (Oriol et al., 2019; Gomez et al., 2020) to combine visual and textual information. For each subtask, the specific CNN achieving best results on the development set is selected, among the investigated ones. The same happens for the Transformer-based architectures. When the "best" architectures are selected and fine-tuned for visual and textual analysis, these are used to encode the entire dataset. It allows training a new classifier which accounts on the evidences from both aspects. In UNITOR these encodings are concatenated, so that the final classifier is a Multi-layered Perceptron[4]. Only this final classifier is fine-tuned, as the remaining parameters are supposed to be already optimized for the task. Future work will consider the fine-tuning of all the parameters of this combined network, here ignored for the (too) high computational cost required from this more elegant approach. It must be said that other information is available in the competition: for example, each MEME was supported with its publication date or the list of politicians appearing in the picture. We investigated the manual definition of feature vectors to be added in the concatenation described above. Unfortunately, these vectors did not provide any significant impact during our experiments, so we only relied on visual and textual information. We suppose this additional information it is too sparse (given the dataset size) to provide any valuable evidence.

Modelling Event Clustering as a Classification task. While Event Clustering may suggest a straightforward application of unsupervised algorithms, we adopted a supervised setting, by imposing the hypothesis that train and test datasets share the same topics. We modelled this subtask as a classification problem, where each MEME is to be assigned to one of the five classes reflecting the underlying topic. UNITOR implements two different approaches. In a first model, the same setting adopted in the other subtasks is used: a CNN and a Transformer-based are optimized on the Task 3 and used as encoder to train the final

[1] `https://huggingface.co/idb-ita/gilberto-uncased-from-camembert`

[2] `https://huggingface.co/Musixmatch/umberto-wikipedia-uncased-v1`

[3] An alternative approach consists in adding the messages from HaSpeeDe to the training set: this approach led to lower results, not reported here due to lack of space.

[4] We investigated also more complex combinations, such as the weighed sum, or point-wise product of embeddings, but lower results were obtained.

MLP classifier. Unfortunately, most of the texts are really short to be valuable in the final classification. We thus adopted a second model which is inspired by the capability of BERT-based models to effectively operate over text pairs, achieving state-of-the-art results in tasks such as in Textual Entailment and in Natural Language Inference tasks (Devlin et al., 2019). In this second setting, each input MEME generates five pairs (one for each topic) which are in the form ⟨*topic definition*, *text*⟩. Let us consider the example *"ma come chi sono? presidé só io senza fotoscioppe!"*, associated to the topic #2, defined[5] as "*L'inizio delle consultazioni con i partiti politici e il discorso al Senato di Conte*". It generates new inputs in the form "`[CLS]` *ma come chi . . . fotosciopppe!* `[SEP]` *L'inizio delle . . . Senato di Conte.* `[SEP]`" which defines sentence pairs in BERT-like architectures. The same approach is applied with respect to each topic. In other words, the original classification problem over five classes is mapped to a binary classification one: each pair is a positive example when the text is associated to the correct topic, negative otherwise. In this way, we expected to detect a possible "semantic connection" between the extracted text and the paired (correct topic) description. At classification time, for each MEME, five new examples are derived (one per topic) and classified. The one generated by the topic receiving the highest softmax score is selected as output.

3 Experimental evaluation and results

UNITOR participated to all subtasks within DANKMEMES. For parameter tuning, we adopted a 10-cross fold validation, so that the training material is divided in 10 folds, each split according to 90%-10% proportion. The model is trained using a standard Cross-entropy Loss and an ADAM optimizer initialized with a learning rate set to $2 \cdot 10^{-5}$. We trained the model for 5 epochs, using a batch size of 32 elements. When combining the networks, the number of hidden layers in the MLP classifier is tuned between 1 and 3. At test time, for each task, an Ensemble of such classifiers is used: each image is in fact classified using all 10 models trained in the different folds and the label suggested by the highest number of classifiers is selected. UNITOR is implement using pytorch[6].

System	Precision	Recall	F1	Rank
UNITOR-R2	**0.8522**	**0.8480**	**0.8501**	1
SNK-R1	0.8515	0.8431	0.8473	2
UNITOR-R1	0.8390	0.8431	0.8411	4
Baseline	0.5250	0.5147	0.5198	-

Table 1: UNITOR Results in Task 1.

Task 1 - MEME Detection. For the subtask 1, the training dataset counts 1,600 examples, equally labelled as "`MEME`" and "`NotMEME`". Results of UNITOR is reported in Table 1, where results are evaluated in terms of Precision, Recall and F1-measure, calculated over the binary classification task (this last used to rank systems). The last row reports a baseline model which randomly assigns labels to images. MEMEs generally adhere to specific visual conventions, where the meaning of text is secondary: as a consequence, our first model (UNITOR-R1) only relies on an image classifier. In particular, it corresponds to the fine-tuning of EfficientNet-B3 over the official dataset. In order to improve the robustness of such a CNN, we adopted a simple data augmentation technique, by duplicating the training material and horizontally mirroring it. UNITOR-R1 ranked forth (over 10 submissions) in the competition. This clearly confirms the effectiveness of EfficientNet, combined with the adopted Ensemble technique. We also investigated larger variants of EfficientNet but they did not outperform the B3 variant: we suppose these larger architectures are more exposed to over-fitting, also considering the dataset size.

Moreover, we adopted a model that combines the output of EfficientNet-B3 with a Transformer-based architecture. Among all the investigated architecture, AlBERTo achieved the highest classification accuracy. Once tuned (in the same 10-cross fold evaluation schema) it is used to encode the entire dataset and the embeddings are concatenated to the ones from EfficientNet-B3. This enables the training of 10 MLPs (one per fold) whose Ensemble defines UNITOR-R2, which ranked first in the task, with a F1 of 0.8501. The overall results thus confirm also the beneficial (although limited) impact of textual information in this subtask.

Task2 - Hate Speech Identification. The training dataset available for the subtask 2 contains 800 training examples, labelled as "`Hate`" and "`NotHate`", while the test dataset counts 200 ex-

[5] In a simplified English: *"Are you seriously asking who I am? Mr President, it's me without Photoshop effects!"*

[6] `https://pytorch.org/`

amples. In Table 2 the results obtained by UNITOR are reported, according to the same metrics adopted in Task 1. Unlike the first subtask, Hate Speech is more related to the textual information. Even the baseline is given by the performance of a classifier labelling a MEME as offensive whenever it includes at least a swear word (resulting in a system with a high Precision and a very low Recall).

System	Precision	Recall	F1	Rank
UNITOR-R2	**0.7845**	0.8667	**0.8235**	1
UNITOR-R1	0.7686	**0.8857**	0.8230	2
UPB	0.8056	0.8286	0.8169	3
Baseline	0.8958	0.4095	0.5621	-

Table 2: UNITOR Results in Task 2.

In this task, we adopted UmBERTo (pre-trained over Wikipedia), fine-tuned for 3 epochs over the HaSpeeDe dataset and then for 3 epochs over the DANKMEMES dataset. Again, a 10-cross fold schema is adopted and the final ensemble of such UmBERTo models originated UNITOR-R1, which ranked 2 over 5 submissions. The improvements with respect to the first competitive system confirms the robustness of the adopted Transformer-based architecture combined with the adopted auxiliary training step. We thus combined this model with a CNN (here ResNET152) to exploit also visual information as for the previous subtask. This combination originated UNITOR-R2, which again provided the best results in the competition, even though a very little margin is obtained w.r.t. UNITOR-R1.

Task3 - Event Clustering. The training dataset available for the subtask 3 contains 800 training examples for the 5 targeted topics and a test dataset made of 200 examples. In Table 3 the performances of UNITOR are reported, as for the previous subtask. Since it is a multi-class classification task, each system is evaluated with respect to each of the 5 labels in a binary setting and then the macro-average is applied to Precision, Recall and F1. Here, the baseline is given by a classifier labelling every MEME as belonging to the most represented class (i.e. topic 0, containing miscellaneous examples). Its results, i.e. a F1 of 0.1297, suggest this is a very challenging task, where the dataset is quite limited, especially considering the overlap that exists among all political topics. In the first row, the run UNITOR-R1 is reported: it corresponds to a model that combines the embeddings from ResNET152 and those obtained by AlBERTo, both achieving best accuracy in our initial tuning within this subtask. UNITOR-R1 ranked first (among three submissions) in this competition with a F1 of 0.2657, which doubles the result obtained from the baseline. It must be said that the Transformer achieves significantly better results with respect to the CNN, suggesting that the visual information is negligible also in this subtask[7]. We thus evaluated a model which considers only text, by fine-tuning an AlBERTo model adopting the pair-based approach presented in Section 2, where each text is associated with the description of the topic. Unfortunately, this model, namely UNITOR-R2, under-performed the first submission, with a F1 of 0.2183.

System	Precision	Recall	F1	Rank
UNITOR-R1	**0.2683**	**0.2851**	**0.2657**	1
UNITOR-R2	0.2096	0.2548	0.2183	2
Baseline	0.0960	0.2000	0.1297	-

Table 3: UNITOR Results in Task 3.

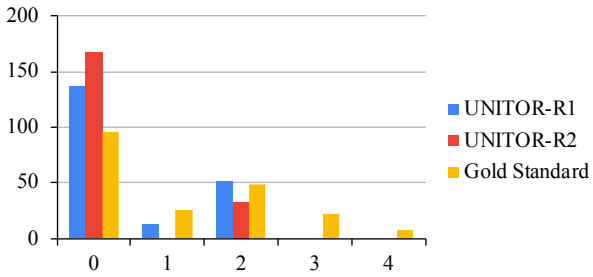

Figure 1: Distribution of labels and classifications in Task 3.

For an error analysis, we compared the assignments provided in the test set and the ones derived from UNITOR, as shown in Figure 1. First, it is clear that the dataset is highly unbalanced, with half of the examples assigned to the class with uncertain topics. Moreover, it can be seen that the combination of textual and visual information makes UNITOR-R1 more robust in detecting topic 2, and most importantly, topic 1, which is ignored from UNITOR-R2. Topics 3 and 4 are ignored by UNITOR but they are also under-represented in the training material. UNITOR-R2 seems more conservative with respect to the largest class (topic 0): it is clear that the repetition of the same topic over many examples introduced a bias. Future work will consider the adoption of more expressive and varied topic descriptions to be paired with texts: for examples, we will select headline news that can be retrieved using Retrieval Engines (e.g.,

[7]These results are not reported for lack of space.

by querying with the topic description) to have a more expressive representation of the topics.

4 Conclusions

This work presented the UNITOR system participating to DANKMEMES task at EVALITA 2020. UNITOR merges visual and textual evidences by combining state-of-the-art deep neural architectures and ranked first in all subtasks defined in the competition. These results confirm the beneficial impact of the adopted Convolutional and Transformer-based architecture in the automatic recognition of MEMEs as well as in Hate Speech Identification or Event Clustering. Future work will investigate multi-task learning approaches to combine the adopted architectures in a more principled way.

References

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of EVALITA 2018, Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020, Online, July 5-10, 2020*, pages 8440–8451.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*, pages 4171–4186, Minneapolis, Minnesota, June.

R. Gomez, J. Gibert, L. Gomez, and D. Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA, USA.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

L. Jiao and J. Zhao. 2019. A survey on the new generation of deep learning in image processing. *IEEE Access*, 7:172231–172263.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Martina Miliani, Giulia Giorgi, Ilir Rama, Guido Anselmi, and Gianluca E. Lebani. 2020. Dankmemes @ evalita2020: The memeing of life: memes, multimodality and politics). In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Benet Oriol, Cristian Canton-Ferrer, and Xavier Giró i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. In *NeurIPS 2019 Workshop on AI for Social Good*, Vancouver, Canada, 09/2019.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Andrew Ross and Damian J. Rivers. 2017. Digital cultures of political participation: Internet memes and the discursive delgitimization of the 2016 u.s. presidential candidates. *Discourse, Context and Media*, 16:1–11, 01.

Limor Shifman. 2013. Memes in a digital world: Reconciling with a conceptual troublemaker. *J. Comput. Mediat. Commun.*, 18:362–377.

Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv e-prints*, page arXiv:1905.11946, May.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated Residual Transformations for Deep Neural Networks. *arXiv e-prints*, page arXiv:1611.05431, November.