POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Transfer Learning Analysis of Fashion Image Captioning Systems

Master of Science in Computer Science and Engineering - Ingegneria Informatica

Author: **Filippo Colombo**

Advisor: **Prof. Paolo Cremonesi**

Co-advisors: **Federico Sallemi**, **Umberto Pietroni**

Academic year: **2020-2021**

Abstract

The performance of image captioning systems and the quality of the generated texts drop when the input samples to process depart from the distribution of the training data: we analyze the generalization capabilities of systems able to automatically provide captions of images, trying to overcome variations and perturbations in the input samples and still achieve high-quality descriptions. Specifically, we tackle this problem in the fashion domain, where clothing samples have several details, making the task of describing garments expensive and only feasible for experts. We design a pre-training procedure of fashion image captioners that exploit a noise generation strategy to improve the performance on unseen distributions of data. We observe that a pre-trained model fine-tuned using a very narrow set of target samples is competitive with a fashion image captioner extensively trained on the complete target source. Moreover, we propose a novel Transformer-based approach that leverages the generative performance of the GPT-2 language model along with the Vision Transformer and BERT encoders to generate text from an image of a garment and its metadata. Besides using automatic metrics, we perform a user study to evaluate the quality of the description of clothing samples.

1. Introduction

Modern deep learning technologies generate text samples of outstanding quality, and when combined with a visual feature extractor, they accurately describe the subjects or the scenes depicted in images at the cost of time-consuming training procedures over a large number of data samples. However, the performance of these models and the quality of the generated texts drop when they process input samples that depart from the distribution of the training data. We tackle this problem in the fashion domain, where clothing samples have a large number of details, and it is crucial to have high-quality descriptions of the products a fashion firm wants to sell online to attract more effectively the attention of customers. Moreover, online catalogues continuously increase and change when new releases of fashion items enter the market: it would be beneficial to have a robust model able to overcome the variations in new clothing samples, saving the time, energy, and resources required to train a new model from scratch that describes the last releases of fashion items.
Our analysis aims to find a way to design image captioning systems applied to the fashion domain that achieve competitive performance without explicit training on the target dataset or with little adaptation through fine-tuning. The training samples in input to the fashion image captioner are triples structured as `<image, metadata, caption>`, where `metadata` is a set of attributes and classes related to the fashion item. We study and analyze the application of

Transfer Learning techniques to Image Captioning systems to define a training procedure to improve the generalization capabilities of said models by using several input sources, and a model architecture to implement fashion image captioners.

The challenge of Image Captioning is widely studied by the research community, and related state-of-the-art solutions usually use an encoder-decoder architecture. In [1], the model consists of a CNN encoder and an LSTM decoder, and an additional attention mechanism makes the model focus on salient regions of the input image. More recent solutions are based on the Transformer architecture, as done in [2], which uses a cross-modal pre-training method that processes the textual caption related to the input image along with the object tags and regions of the input images extracted by an object detector. In [3], the authors tackle the Image Captioning challenge specifically applied to the fashion domain. They propose a model that exploits GPT-2 to generate text by processing image features extracted by a CNN module and additional textual information related to the input item.

When dealing with deep learning models, the usual Transfer Learning procedure consists of two phases: the *pretraining*, where the model learns a representation of the input samples related to a source task, and the *adaptation*, where the representation learned by the model is transferred and applied to a new task. In our analysis, we consider both stages: first, the design of a pre-train methodology specific for fashion image captioners to improve the generalization capabilities of such models; afterward, the adaptation through *finetuning* of the knowledge held in the weights of the model to the final task, using a limited number of samples belonging to the target data source. Moreover, we propose a Transformer-based architecture that processes a multi-modal input (visual and text) through two distinct encoder stacks to generate the description of a fashion item having as input its image and metadata. Besides language modeling, we consider the contrastive training objective used by CLIP [4] to align the visual and textual modalities: in [4], the authors propose a learning framework where a model is trained to minimize a contrastive loss to prompt paired image and text embeddings to be similar in a shared representation space and, on the other hand, push away non-paired images and texts.

2. Approach

In this section, we describe our proposed pre-training methodology and model architecture.

2.1. Pre-Training for Fashion Image Captioning

We propose a pre-training procedure that leverages various data sources that highly differs in several aspects, like the metadata labels associated with clothes and the poses or the surrounding in which garments are pictured, and a noise generation strategy.

Batches and Loss The data sources employed in the pre-training can be very different in size and structure of textual metadata: we apply a set of transformations to obtain a common textual format across datasets and design a *stratified batch sampler* algorithm to avoid biases towards the dominant sources in the overall training set. We force the training batches to include at least one sample per data source to ease the learning process and avoid single dataset specialization. While oversampling introduces redundant data that can cause the model to overfit and increases the training time, and undersampling can discard plenty of useful information, the batch sampling algorithm uses all the data available and determine the optimal assignment of samples belonging to a data source so that the overall number of batches, thus the training time, is minimized.

Our goal is to learn a language model, which translates into estimating the conditional probability distribution over labels of the model vocabulary; therefore, the optimization criterion we use is the *cross-entropy* loss. When updating the parameters of the model, the data samples are weighted to ensure an equal contribution of each data source. The weight of a sample is the reciprocal of the frequency of samples belonging to the same data source in the batch.

Noise generation Leveraging multiple data sources during training makes the fashion image captioner more robust in identifying the details in the input samples. Nevertheless, when generating the final caption of a fashion item, the model is prone to recognize the data source that most probably includes the current input sam-

ple and leverage the knowledge related only to that source. To force an even greater generalization, we define a strategy to combine existing clothing samples belonging to different sources to prevent this unintended behavior and process clothing samples without considering not relevant features like the background of the image, having or not a human wearing the garment, or the structure of the metadata.

Given two *similar* clothing products, a hybrid sample consists of the `image` and the `metadata` of the first one and the `caption` of the second one. Hybrid samples do not replace the original ones; they are new clothing products added to the overall training set.

Mixing clothing items is a challenging task: combining products at *random*, despite the ease of development, confuses the model as there is no semantic relationship between the two samples merged; contrary a *one-to-one* mapping between products of different sources would be very effective, but too complex to achieve. Our approach uses a reference *taxonomy* that guides the matching between products of various sources by starting from the metadata associated with each clothing sample. First, we extract the taxonomies $\mathcal{T}_1, \ldots, \mathcal{T}_N$, related to each data source: in this way, the challenge of matching similar products across sources translates into finding pairings between leaves of the resulting trees. Assignments between taxonomies are manually designed by inspecting the trees, and having N taxonomies to match together with connections that are not always symmetric implies at least $\frac{N \times (N-1)}{2}$ set of rules to be coded. This process is expensive and time-consuming, so we define an additional reference taxonomy $\mathcal{T}^*$ that acts as a proxy: each dataset taxonomy $\mathcal{T}_n$ is mapped to the reference $\mathcal{T}^*$, reducing the total number of mappings to be coded to N.

Before training, iterating over a source dataset X allows the conversion of the metadata of each clothing sample into the structure of the reference $\mathcal{T}^*$ through the mapping rules $\mathcal{T}_X \rightarrow \mathcal{T}^*$. This procedure is repeated for all the datasets. At training time with probability p, if clothing products of various sources share a leaf of the tree of $\mathcal{T}^*$, then they are *similar* because they have the same classification according to $\mathcal{T}^*$, thus suitable to forge a hybrid sample with the `image` and `metadata` of the *first* product and the `caption` of the *second* one.

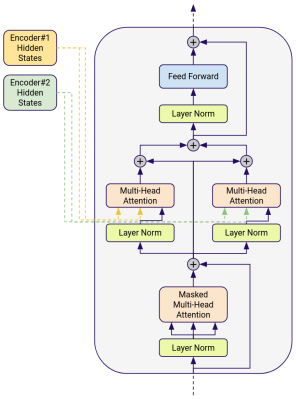


Figure 1: Vision-Text MED decoder block.

2.2. Vision-Text MED

We propose a Transformer-based architecture named *Vision-Text Multi-Encoder Decoder* that learns the visual and the metadata hidden representations using two different encoding stacks that keep the two modalities apart. The Transformer decoder block has a *cross-attention* layer that allows focusing on relevant parts of the input during the generation process. Our model otherwise requires that the decoder attends not only the output of a single encoding stack but two of them, as the visual and textual modalities are treated separately. To do so, we extend the standard decoder block by adding a *multi-head attention layer* and an additional *normalization layer* as reported in figure 1.

Contrastive objective The image and metadata are complementary ways to describe a clothing product, and ideally, regardless of the modality, their embeddings are nearby in shared representation spaces. To promote this behavior, when training our architecture we consider the contrastive pre-training objective proposed in CLIP [4], in addition to the language modeling objective. Given a batch containing N training samples, the goal is the prediction of the correct pairings between images and metadata: overall, there are $N \times N$ possible pairings, of which $N^2 - N$ are incorrect matchings. In practice, the embeddings of the image encoder and the text encoder are projected to a *multi-modal* representation space, and a similarity measure (*cosine*) is maximized for the N correct pairs of image and metadata embeddings while minimized for the other $N^2 - N$. The optimization involves the computation of a symmetric cross-entropy loss over the similarities scores.

Unlike CLIP, which is used for Image Classifi-

cation, our goal is the generation of captions of clothing products having as input their images and metadata. We explore two different approaches to take advantage of the contrastive alignment of vision and text embeddings together with the language modeling objective: using a *multi-objective loss* function or adding a *pre-alignment* stage. The former option involves a single learning stage in which the model is trained according to a joint contrastive and language modeling objective, ruled by the hyperparameter λ:

$$L = \lambda L_{contrastive} + (1 - \lambda) L_{c.e.}, \ \lambda \in [0, 1]$$

The latter instead requires splitting the overall training procedure into two stages: first, the self-supervised alignment of the encoders embeddings through the contrastive loss; then, the update of the model parameters through the language modeling objective. The model input in the first stage consists of `<image, metadata>` pairs, while in the second one the `caption` of clothing products is also included. In this second option, if available, we add the possibility to include less informative images when performing the initial contrastive alignment: data sources may have multiple images per fashion item organized according to their importance (for example, first the frontal picture, then the complete outfit, and lastly, the cropped details of the garment); we associate a probability to each image proportional to its importance and choose the image at training time through a *weighted random selection*.

3. Experiments

In the following sections, we analyze the performance of our pre-training method on two different deep learning models for Fashion Image Captioning.

3.1. Datasets

In our experiments, we use one public dataset and three private ones.

Fashiongen Fashiongen [5] is a large-scale dataset used in the Generative Fashion Challenge[1]. It consists of fashion images annotated with descriptions and categories provided by professional stylists. Each clothing product belongs to a main category and a more detailed subcategory, which are mutually exclusive.

[1]The competition is part of the workshop of Computer Vision for Fashion, Art and Design at ECCV, 2018.

Dataset	Train	Valid	Test	$\overline{w}$	$\overline{s}$
Fashiongen	54132	6015	7519	30.9	6.5
ID1	1235	153	138	54.1	2.9
ID2	39282	4365	5921	23.0	2.2
ID3	1881	233	209	18.0	1.9

Table 1: Number of items in train, validation, and test splits along with the average number of words ($\overline{w}$) and sentences ($\overline{s}$) per caption.

Industrial datasets Thanks to the collaboration with industrial partners, we can use three private sources that we cannot share, but we describe their features and statistics. The datasets consist of high-resolution fashion images annotated with descriptions and metadata. Two out of three datasets contain clothing samples that may have multiple pictures, each of them representing either the garment worn by a fashion model through different perspectives or the clothing product alone. Metadata associated with a fashion item are either categorical data (classes) or attributes. Depending on the dataset, captions of the clothing products have different styles and lengths; all the data sources have descriptions written by fashion experts. We refer to this datasets as *ID1*, *ID2*, and *ID3*.

Dataset comparison The datasets differ in their sizes and the structures of the captions of fashion items. *Fashiongen* and *ID2* are the datasets with the highest number of samples among the four available. Considering the average number of words and sentences per caption, *ID2* and *ID3* are similar, while *Fashiongen* has captions that contain several short sentences. On the contrary, *ID1* contains clothing samples with very long captions characterized by refined words and few sentences. Table 1 provides how we split the samples of each dataset along with the average number of words per caption $\overline{w}$, and the average number of sentences per caption $\overline{s}$.

3.2. Evaluation metrics

The scores we consider in the performance evaluations are BLEU-n, GLEU, METEOR, ROUGE-n, and ROUGE-L [6]: each generated caption of the model under analysis is compared with its reference, computing a single score; then they are averaged over the whole test partition

to provide an estimate of the overall quality of the image captioner.

3.3. Multimodal GPT-2

We test our pre-training procedure using the *Multimodal GPT-2* architecture proposed in [3]. We choose to use this model after comparing its performance on the task of Fashion Image Captioning with other works.
The first step of our analysis is the selection of a *source* and a *target* dataset. The models trained on only one of these two datasets determine a gap in performance that we want to cover through our pre-training and fine-tuning approach.

***Target-only* experiments** The *target-only* experiments use the model that has the best achievable performance on the *target* dataset; therefore, the train and test partitions lean on the same data distribution, i.e., belong to the same dataset. We analyze the transfer learning performance of models using as target datasets *Fashiongen* and *ID2*.

***Source-only* experiments** The *source-only* experiments represent the simple condition in which an already available captioning system, trained on one *source* dataset, is used with clothing samples of a different dataset without adopting any pre-training or adaptation technique that eases the transfer of learned representations. In our scenario, the source-only model is trained using the train partition of *ID1*. It struggles even in determining details of clothing products that seem easy to catch: changing the distribution of the input samples makes the model unable to distinguish the main features of the garment like its category or the target gender. This wrong behavior is due to the less relevant details in the input sample (the background, the presence or not of a person wearing the clothing product, . . .) that the model learns as critical for the final description of the garment.

Pre-Training experiments The models learn from multiple sources but are tested on a new distribution, so we leverage all the datasets available except the one used for testing, as reported in table 2. Moreover, we add 5% of noise to the train samples following our strategy to combine clothing products of different sources. Pre-training a fashion image captioner through

		Target dataset	
		Fashiongen	*ID2*
Train datasets	*target-only*	Fashiongen	ID2
	source-only	ID1	ID1
	pre-training	ID1	Fashiongen
		ID2	ID1
		ID3	ID3
	fine-tuning	Fashiongen (5%)	ID2 (5%)

Table 2: In the *pre-training* stage, we leverage the train partitions of all the datasets available except the *target* one.

our approach allows achieving higher scores than the *source-only* model and reduces the gap in performance between the two reference models, as the main features of the clothes are predicted correctly, but still, it can struggle with close colors or finer details.
Besides catching the correct details in the input, an important aspect is the structure of the generated caption: each data source uses a particular style in describing products, which is the result of brand-related design choices such as the type of narrative, or which kind of information should or should not be in a textual description. A fashion image captioner does not learn these stylistic choices when trained only with sources that adopt a structure of captions different from the target one, and metrics that rely only on n-gram matching can get low-level scores even with semantically correct output captions.

Fine-tuning experiments To learn the source-dependent properties, like the ones identified in the previous section, we fine-tune the pre-trained model on a target dataset by using a limited amount of training samples (5%) for a fixed number of epochs (5). In this way, besides producing a semantically correct description, the output sentences of the fashion image captioner follow the stylistic decisions of the target data source. Fine-tuning allows covering a large portion of the gap in the performance between the two reference models, and these results underline the benefits of leveraging a pre-trained architecture when dealing with constraints such as costly or limited training samples and computing resources with bounded training time or memory. Table 3 shows the data and time requirements in detail: fine-tuning the pre-trained model on the *target* new dataset requires 3.5% of the time necessary to train the *target-only* model.

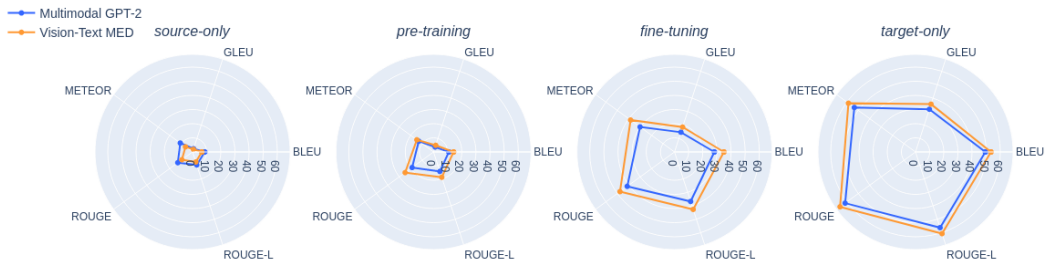


Figure 2: Comparison of the performance achieved in each step of our transfer learning methodology applied to *Multimodal GPT-2* and *Vision-Text Multi-Encoder Decoder* when tested on *ID2*. In each step, the datasets used as train sources are reported in table 2.

	#Train Samples	GPU Time
Fashiongen		
target-only	54132	12169 sec (1.55 days)
fine-tuning	**2706**	**4275 sec (1.18 h)**
Industrial dataset#2		
target-only	39282	8456 sec (1.08 days)
fine-tuning	**1964**	**3556 sec (0.98 h)**

Table 3: Time and data requirements to fine-tune the *pre-trained* models or training them from scratch (*target-only*).

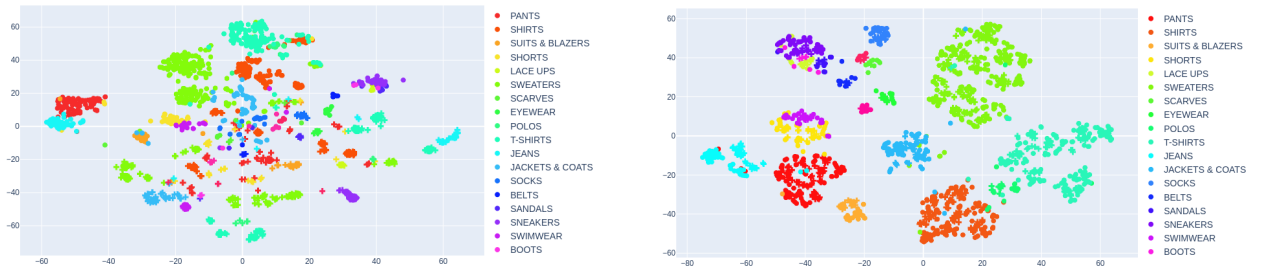
3.4. Vision Text MED

Our training methodology does not depend on the particular choice of deep learning model, so we carry out the same analysis done with *Multimodal GPT-2* to our architecture *Vision-Text Multi-Encoder Decoder*. The architecture is initialized using the weights of a pre-trained model available in the Hugging Face Transformers[2] library. The vision encoder is a ViT model pre-trained using the BEiT masked image modeling task [7], the text encoder is a pre-trained uncased BERT, and the decoder uses the weights of the pre-trained GPT-2 model.

Pre-training options Our new architecture introduces the possibility to consider an additional contrastive learning objective in the pre-training stage to align image and metadata embeddings. We identify two options to perform this alignment together with the language modeling objective: follow a *multi-objective* learning approach or add a *pre-alignment* stage. The hyperparameter that rules the contribution of the two losses in the *multi-objective* optimization option is equal to $\lambda = 0.3$. We find that the *pre-alignment* of image and metadata embeddings alone is less effective than the single pre-training stage that leverages the *multi-objective* learning (MOL) task, and combining the two approaches does not guarantee an improvement of the performance: the conditioning of the embeddings through the *pre-alignment* step before the MOL pre-training improves the performance on datasets that make extensive use of additional metadata related to clothing samples.
Notice that we analyze three potential settings, but in principle, there could be others according to the order of the pre-training steps. Besides, a deep optimization of the hyperparameter λ could provide additional insights.
Our model outperforms *Multimodal GPT-2* in all the stages except for the *source-only* model: using an additional encoder stack that processes the metadata related to the clothing samples increases the data requirements, so it becomes effective when the overall architecture is pre-trained over multiple sources. In that case, the fine-tuned *Vision-Text Multi-Encoder Decoder* outperforms *Multimodal GPT-2* by a significant margin. Figure 2 provides a comparison of the scores achieved by the two architectures in all the stages of our analysis.

Contrastive alignment Regardless of the training objective, the image and metadata embeddings related to a clothing sample should be close in a shared representation space. We train two models using the data of the *industrial dataset#2*: for the first one, we use only the language modeling objective without forcing the alignment between the two modalities; differently, we train the second one using the multi-objective approach with the hyperparameter $\lambda = 0.3$. For both the models, we extract the image and metadata embeddings of the clothing samples of the test partition and use the *t-distributed stochastic neighbor embedding* (t-SNE) dimensionality reduction technique to project the embeddings in a 2-dimensional space and visualize them, as in figure 3. Using the contrastive objective, the embeddings of the two

[2] `https://github.com/huggingface/transformers`



(a) Training without the contrastive alignment.

(b) Training with the contrastive alignment.

Figure 3: Visualization of image and metadata embeddings of clothing samples through t-SNE whether the training procedure of *Vision-Text Multi-Encoder Decoder* consists of the language modeling objective either alone (a) or with the additional contrastive alignment (b). We use $\cdot$ and $+$ to represent the embeddings of images and metadata, respectively.

modalities are close; otherwise, the model differentiates between image and metadata embeddings even though the clothing category is the same. Besides, leveraging the contrastive objective helps the model in understanding the clothing category of embeddings of the same modality more clearly, as the image embeddings belonging to the same cluster become more cohesive and the distances between clusters are more marked, showing that the alignment between image and textual embeddings improves the ability of the model to differentiate among clothing categories.

3.5. User study

Automatic metrics designed for Natural Language Generation tasks are inexpensive, deterministic, and quick to compute ways to approximate the quality of automatically generated text samples. Their main drawback is that the words in both candidate and reference sentences are equally weighted, so missing out on content-bearing pieces of sentences instead of less significant ones is valued the same rather than being penalized more.

We design a survey to elicit judgments from users about the quality of descriptions of clothing products: the goal is to compare the captions generated by models trained through our pre-train methodology with the corresponding ground truths. The models we consider in the survey are *Multimodal GPT-2* and our *Vision-Text Multi-Encoder Decoder*.

Evaluation Evaluating the quality of a text sample is not easy to assess, especially when dealing with the description of fashion items characterized by several details; thus, we identified three criteria that jointly determine the overall quality of the caption of a clothing product. First, a *precision* measure to evaluate the correctness of the information provided in the description; second, a *recall* oriented score stating the number of relevant details mentioned in the description concerning the garment; finally, we ask to give feedback on the *syntactic correctness* of the text sample under evaluation.

We provide the image and metadata related to a garment, and the users express their level of agreement or answer to the following sentences using 5-point Likert scales:

1. *The description contains only correct information about the garment in question.* (*precision* oriented)
2. *The description contains all relevant information to describe the garment.* (*recall* oriented)
3. *How do you evaluate the syntactic correctness of the sentence?*

Data The clothing items evaluated in the survey belong to *Fashiongen* and the *ID2*. We extract a random subset of 15 samples of the test partitions of each dataset and generate their descriptions using the fine-tuned *Multimodal GPT-2* and *Vision-Text Multi-Encoder Decoder*, making the overall number of captions of the survey equal to 90.

Overall, we collect 786 ratings of descriptions of items by 33 users. Notice that users answering our survey usually do not rate all the 90 captions: on average, a user provides 24.1 ratings of descriptions of clothing items. Additionally, we check the distribution of answers per user and item and filter out the ratings of users that answered to a number of descriptions below a threshold ($<$ 5 items rated) and the descriptions of clothing items that received few ratings ($<$ 5 ratings of users) in at least one of the

	Q1	Q2	Q3
TOST $[\theta = 1.0]$			
Multimodal GPT-2	**0.04211**	**0.00055**	**1.27e-13**
Vision-Text MED	**9.51e-08**	**0.00001**	**4.16e-09**
TOST $[\theta = 0.5]$			
Multimodal GPT-2	0.88141	0.87791	**0.00001**
Vision-Text MED	**0.00888**	**0.03543**	**0.00017**

Table 4: The resulting p-values of the statistical tests to compare the distributions of observed data relative to generated captions and the ground truths.

three scenarios (ground truth, *Multimodal GPT-2*, *Vision-Text Multi-Encoder Decoder*). Finally, we standardize the remaining rates by removing the user biases: in this way, each rate does not depend on the user preference.

Results Given independent samples of the average rating of items in the three scenarios (*ground truth*, *Multimodal GPT-2*, and *Vision-Text Multi-Encoder Decoder*), we perform an equivalence test for each question of the survey to compare whether the mean ratings related to descriptions generated by our models differ by a small amount to the average ground truth rating. The statistical test we use is the *two one-sided t-test* (TOSTs). Given a margin of equivalence θ on a 5-point Likert scale, TOST considers the null and alternative hypotheses defined as:

$$\begin{aligned} H_0 &: \mu_2 - \mu_1 \leq -\theta \text{ or } \mu_2 - \mu_1 \geq \theta \\ H_A &: -\theta < \mu_2 - \mu_1 < \theta \end{aligned}$$

We analyze the mean values of the collected ratings according to different values of tolerance (equivalence margins). Table 4 provides the results of the statistical tests we perform. We find that the mean syntactical quality of generated captions is indistinguishable from the one related to the ground truths independently by the automated model. Differently, considering the question 1 and 2, according to the samples collected and the fixed significance level $\alpha = 0.05$, their mean values of ratings related to the captions generated by *Multimodal GPT-2* are statistically equivalent to the mean values related to the ground truths within a 1.0 rating margin. For *Vision-Text Multi-Encoder Decoder*, the same equivalence properties hold even within a 0.5 rating margin.

4. Conclusions

We study and analyze the performance of fashion image captioning systems when the distribution of the data in input to the model changes. We show how the performance of trained fashion image captioners varies according to the input distribution: the use of a model which is simply trained over a different source reflects a drastic reduction in the performance and the quality of the generated captions, not being able to recognize the clothing samples in input anymore. By leveraging our pre-training method and noise generation strategy, we can improve the performance over unseen distributions of data, making the model generalize better. Through a final adaptation stage of the pre-trained model using a very narrow set of target samples, the fashion image captioner achieves competitive performance and high-quality captions compared to the model extensively trained on the target source. Additionally, we design a novel Transformer-based approach that leverages two encoder stacks to process a multimodal input and an additional contrastive objective that aligns the embeddings of the two input modalities. We carry out a performance study showing how the contrastive alignment between embeddings reflects on the representation learned by the model and how it improves the performance of fashion image captioners. Finally, we perform a user study to demonstrate that the mean values of the distributions of the ratings given to the ground truths and to the captions generated by fashion captioners are statistically equivalent, assuming a pre-defined tolerance value.

References

[1] Kelvin Xu et al. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. 2015. URL: https://proceedings.mlr.press/v37/xuc15.html.

[2] Xiujun Li et al. *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*. 2020. arXiv: 2004.06165 [cs.CV].

[3] Umberto Pietroni, Federico Sallemi, and Paolo Cremonesi. *Image tagging and captioning for fashion catalogues enrichment*. 2020. URL: https://www.politesi.polimi.it/handle/10589/169410?mode=complete.

[4] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[5] Negar Rostamzadeh et al. *Fashion-gen: The generative fashion dataset and challenge*. 2018.

[6] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. *A Survey of Evaluation Metrics Used for NLG Systems*. 2020. URL: https://arxiv.org/abs/2008.12009.

[7] Hangbo Bao, Li Dong, and Furu Wei. *BEiT: BERT Pre-Training of Image Transformers*. 2021. arXiv: 2106.08254. URL: https://arxiv.org/abs/2106.08254.