Multimodal Conditionality for Natural Language Generation

Anonymous ACL submission

Abstract

Large scale pretrained language models have demonstrated state-of-the-art performance in language understanding tasks. Their application has recently expanded into multimodality learning, leading to improved representations combining vision and language. However, progress in adapting language models towards conditional Natural Language Generation (NLG) has been limited to a single modality, generally text. We propose MAnTiS, Multimodal Adaptation for Text Synthesis, a general approach for multimodal conditionality in transformer-based NLG models. In this method, we pass inputs from each modality through modality-specific encoders, project to textual token space, and finally join to form a conditionality prefix. We fine-tune the pretrained language model and encoders with the conditionality prefix guiding the generation. We apply MAnTiS to the task of product description generation, conditioning a network on both product images and titles to generate descriptive text. We demonstrate that MAnTiS outperforms strong baseline approaches on standard NLG scoring metrics. Furthermore, qualitative assessments demonstrate that MAnTiS can generate human quality descriptions consistent with given multimodal inputs.

1 Introduction

The use of transfer learning techniques in Natural Language Processing (NLP) significantly improves previous state of the art methods across a wide range of NLP tasks (Dai and Le, 2015; Devlin et al., 2018; Howard and Ruder, 2018; Radford et al., 2019; Brown et al., 2020). In this setting a transformer-based language model is pretrained on large unlabelled corpra and then fine-tuned on supervised data together with a task-related head (Devlin et al., 2018). Such approaches are prominent in Natural Language Understanding (NLU) tasks, but remain less explored for text generation.

Transfer learning methods have recently been applied to the joint learning of multiple modalities, where both image and text based inputs are pretrained together (Lu et al., 2019; Li et al., 2020; Su et al., 2019b; Chen et al., 2020; Li et al., 2019). In these approaches, learning combined representations of visual and textual data during pretraining instead of task specific training, leads to better semantic representations. Due to state-of-the-art performance and straightforward downstream training, it is fast becoming the default method for multimodal tasks like visual question answering, visual entailment, and caption-based image retrieval.

A natural extension to this approach would adapt pretrained language models for conditional Natural Language Generation (NLG) with multimodal conditionality. This can be achieved in an encoder-decoder framework where the encoder learns to embed conditionality while the pretrained decoder would modify the generation based on this encoding. Earlier work suggests this works well for tasks where generation depends on purely textual information (Golovanov et al., 2019; Zhang et al., 2019; Song et al., 2019). Recent work used other modalities like image or class information Ziegler et al. (2019) to guide the generation of pretrained models. However, that work considered only a single modality and required the introduction of new parameters within the pretrained model that could adversely affect generation capability.

In this work, we propose MAnTiS, a general approach to adapt transformer-based language models into multimodal conditional NLG models. We encode each modality type using specific encoders, joined to form a conditionality prefix, with separator tokens delimiting each modality. During fine-tuning the decoder uses the prefix as history and predicts outputs in a continuous fashion. Because the prefix is decoder independent, the generation can be conditionalized towards any modality. We drew inspiration from Kiela

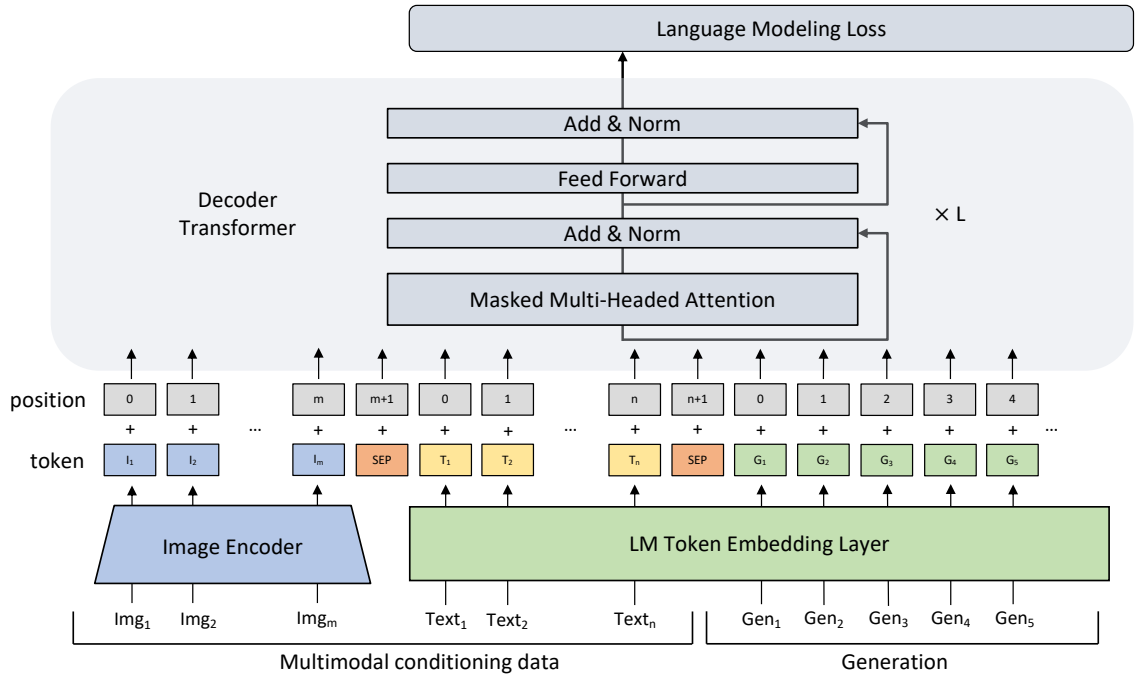

Figure 1: An overview of the MAnTiS architecture. Conditioning images are passed as input through an image encoder and mapped to textual token space of language model. Input text is encoded using the language model's encoder and together with image tokens form the conditionality prefix. The language modeling loss is computed only for the text tokens. Here m and n represent the number of input images and text tokens respectively and L is the number of decoder transformer layers.

et al. (2019) which shows that self-supervised unimodal transformer models are capable of learning context between different modalities through supervised learning for classification tasks.

We demonstrate the effectiveness of this approach on a fashion captioning dataset (Yang et al., 2020), where given a product's name and image, the model generates an e-commerce relevant description. We compare generations against competing approaches that rely on injecting conditionality vectors into pretrained language models. Through this, we found that MAnTiS outperforms other models. We perform both quantitative and qualitative experiments and show the effectiveness of this approach without requiring any complex model stitching. Extension of MAnTiS to any modality type is straightforward to implement for any transformer-based pretrained model. We thus provide a strong baseline approach for future transfer learning in NLG.

2 Related Work

2.1 Transfer Learning in Multimodal Models

Language representation model BERT (Devlin et al., 2018) demonstrated that transformer models trained with masked language modeling and next sentence prediction objective can lead to state-of-the-art performance for a variety of NLU tasks. VilBERT (Lu et al., 2019) extended the approach towards multimodality with separate transformer streams for image and text with cross-modality interaction though co-attention between the two streams. Other methods (Li et al., 2020; Su et al., 2019b; Chen et al., 2020; Li et al., 2019) showed that single stream transformer models can learn the relationship between image and text. These models are pretrained on vision and language data, however Kiela et al. (2019) proposed a different approach where a pretrained unimodal (text) BERT model is fine-tuned together with a different modality (image), skipping the multimodal pretraining step. These methods are effective for understanding tasks like classification, but have not been studied

Product Title	Product Images	Product Description
Denim Parka with Genuine Fox Fur Trim		*Plush fox fur lining the hood and sparkling embellishments across the front bring luxe detail to a utilitarian-chic parka cut from pure-cotton Italian denim.*

Table 1: A sample entry from the FACAD dataset of fashion products (metadata not shown).

for multimodal conditional generative tasks.

2.2 Finetuning Natural Language Models for Controllability

Unconditional language models can be adopted for text generation tasks such as language translation (Edunov et al., 2019), question answering (Su et al., 2019a), and summarization (Zhang et al., 2019). Other work demonstrates the sufficiency of providing the context text as the prefix in guiding text generation for these tasks (Brown et al., 2020; Radford et al., 2019). For example, translation models generate translated sentences given the source language input presented as prefix. Moreover, Golovanov et al. (2019) showed that concatenation of multiple textual contexts may form the guiding prefix. Keskar et al. (2019) added controllability in language model during training by appending training corpus with different control codes, resulting in impressive generations for existing codes. However, these approaches are limited to text-based controllability.

2.3 Conditionalizing Pretrained Language Models for Generation

Models conditioned with pretrained language include noisy channel modeling (Yee et al., 2019) and fusion approaches (Sriram et al., 2017; Gulcehre et al., 2015) that concatenate hidden states of the conditional model with that of language model to predict the next word. Recently, Ziegler et al. (2019) proposed a modality invariant conditionalization approach for any transformer-based language model through pseudo self-attention. There, in every pretrained transformer layer the encoding vectors are considered as history and allowed to be attended over, leading to conditioning during self-attention. They also tested context attention where the decoder transformer layer is converted to a encoder-decoder, using pretrained weights for the decoding part. They demonstrated that pseudo self-attention is effective for even non-textual conditioning like image-based paragraph generation and class-based review generation. However, their work considers single modality conditioning whereas we treat the problem of multimodality conditioning.

In this work, we use the approaches of Ziegler et al. (2019) as a comparative baseline. However, introducing new parameters within the whole network may hinder generative capabilities of pretrained language models. In addition, they require cumbersome manipulation of pretrained model architectures. MAnTiS addresses these issues with a simple approach requiring fewer additional parameters.

3 Method

Given a sequence of token vectors $x = (x_1, \ldots, x_n)$, language models learn the probability $p(x)$,

$$p(x) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1})$$

Here, we adapt a pretrained language model into a multimodal conditional model that learns the conditional probability distribution $p(x|y)$, where $y = (y_1, \ldots, y_n)$ consists of tokens of any modality.

$$p(x|y) = \prod_{i=1}^{n} p(x_i \mid y, x_1, \ldots, x_{i-1})$$

The goal is to learn $p(x|y)$ given supervised dataset of x, y pairs. To achieve this, we frame the problem using an encoder-decoder architecture. We encode conditional modalities using modality specific encoders and then project to the textual token space of the language model. Between the different modality types we add separator tokens, allowing the model to distinguish between them. We prepend

Model	BLEU4	CIDEr	METEOR	ROUGE-L
CONTEXT-ATTN	3.9	30.4	10.2	17.2
PSEUDO-SELF	4.2	32.3	10.3	17.3
MANTIS	**4.8**	**36.8**	**10.8**	**17.9**
MANTIS-SCRATCH	3.9	30.8	10.1	17.3
MANTIS-MULTI	4.9	39.0	11.1	18.2
MANTIS-MULTI + TEXT DROPOUT	**5.0**	**39.5**	**11.1**	**18.2**

Table 2: Comparison of generator performance scores

Model	Grammar	Non-Redundancy	Consistency	Attractiveness	Overall
CONTEXT-ATTN	0.74	0.89	0.75	0.51	0.603
PSEUDO-SELF	0.76	0.93	0.72	0.54	0.578
MANTIS	**0.82**	**0.96**	**0.81**	**0.62**	**0.665**

Table 3: Qualitative evaluation of generated descriptions

these conditional tokens y to the input guiding the generation x. We illustrate the overall architecture of our approach in Figure 1.

The following subsections describe the encoding strategy, details of input construction, and the fine-tuning procedure.

3.1 Encoder Mapping

During the encoding stage we use both image and text modalities to condition the generation. To encode images we extract the embedding form of the last fully connected layer of a pretrained ResNet-152 model (He et al., 2016). This can be regarded as a single dense token per image whose dimension N depends on the ResNet model. Transformation of the input image uses the same setting as during the pretraining process, which includes resizing, center cropping and normalization. Next, we project the token into the language model embedding space D through a linear layer with learnable weight matrix $W \in \mathcal{R}^{N \times D}$.

The embedding function of the decoder language model encodes the text. For the language model, we use the transformer-based pretrained model GPT-2 (Radford et al., 2019), an auto-regressive model whose self-attention module can attend only on previous tokens, with Byte Pair Encoding (BPE) for text tokenization.

This approach can easily be extended towards any modality because we map the encoding to the textual space. The encoder and decoder are jointly fine-tuned end-to-end during supervised learning. Allowing the encoder, specifically, the image encoder to be fine-tuned will contribute towards effective learning of image token mapping.

3.2 Multimodal Fine-tuning

In the GPT-2 language model the input consists of a sum of token and position embeddings, with the position encoding zero-indexed. For each conditional modality, we start the position encoding from zero as well. Between each modality token we add a separator token [SEP] whose position is one plus the previous token position. The first conditional token is prepended with a beginning of sentence [BOS] token and the generation ends with a end of sentence [EOS] token.

During fine-tuning, the model is trained using the same loss function (cross-entropy) as GPT-2, between the next predicted word of the language model head and the ground truth word. No loss is computed for the image tokens because they have no exact vocabulary.

3.3 Modality Dropout

Fine-tuning in this manner forces the pretrained language model to learn cross-modality correlations between image and text. Naturally, this can cause text tokens to influence generation more than

other modality tokens. In our approach, we pretrain the decoder language model and image embedding model while we randomly initialize the image mapping layer for training. Neverova et al. (2015) proposed ModDrop, arguing that randomly dropping different modality channels during training could help learn cross-modality representations and reduce false co-adaptions.

Because image representations must be fused into the text-only model, we randomly dropped out text conditionality paths with a probability p tuned during training. We speculate that this could provide improved image conditioning and lead to better overall performance. This was performed in addition to the standard dropout within the transformer decoder layers.

4 Experimental Setting

This section includes detailed information on the datasets, metrics, and baselines used during training and evaluation.

4.1 Datasets

We used the initially released version of the fashion captioning dataset FACAD (Yang et al., 2020). This dataset consists of fashion articles and their names, images from different perspectives, e-commerce relevant descriptions, colors, and other pieces of metadata. In this work, we want to generate product descriptions given the title and various images of a product. An example of the dataset is shown in Table 1. There are total 55,959 descriptions. We removed entries with empty description, name or images, as well as duplicated descriptions, reducing the size to 45,748. Out of these 40,748 were used for training, 2,500 for validation, and the remaining 2,500 for testing.

Yang et al. (2020) used this dataset for the image captioning problem where the generated caption depends only on a single given image. We used this dataset for multimodal conditioned NLG, where multiple instances of each modality may be provided as input. However, using multiple images per description significantly reduces the total number of training samples.

4.2 Evaluation Metrics

To perform qualitative evaluation we report model performance on the most commonly used NLG metrics, which include BLEU4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Denkowski and Lavie, 2014), and ROUGE-L (Lin, 2004) scores.

4.3 Training Details

GPT-2 is a large transformer-based model trained on the WebText dataset ($\sim$40 GB), consisting of text from 8 million non-Wikipedia webpages (Radford et al., 2019). It shows excellent performance with coherent text generation; thus, we use it as the base unconditional pretrained language model. In particular, we use GPT-2 medium, possessing an embedding size of 1024, comprising 24 layers with 16 heads per layer and including a total of 345M parameters. It is publicly available from the HuggingFace repository (Wolf et al., 2019). We use the same vocabulary with an addition of three tokens: BOS, SEP and PAD (padding token). For encoding images we use ResNet-152 trained on the ImageNet dataset (Deng et al., 2009), which is publicly available in PyTorch's torchvision package (Paszke et al., 2019).

Additionally, we tuned the learning rates for each model between $1e{-}5$ to $5e{-}5$. We tuned text modality dropout between 0.3 to 0.7 and set all other dropout values to 0.1. Training was done using the AdamW optimizer and a linear scheduler with warmup.

4.4 Baseline Methods

We compare MAnTiS against the current most advanced approaches for conditioning language model. In comparing approaches we used the latest available code published by the authors.

CONTEXT-ATTN: Context attention adds a randomly initialized encoder-decoder layer on top of every pretrained decoder layer of GPT-2 (Ziegler et al., 2019). Multimodal conditionality tokens are used as the encoder tokens.

PSEUDO-SELF: Pseudo self-attention prepends additional multimodality conditioning tokens to every self-attention layer of GPT-2 (Ziegler et al., 2019). This achieved the best performance for unimodal conditioning, forming the strongest baseline.

5 Results

A single image for each fashion item was chosen (the first image in the dataset sample) to fairly compare Context-Attn, Pseudo-Self, MAnTiS, and MAnTiS-scratch. In MAnTiS-scratch we randomly initialized the internal language model. In MAnTiS-multi we used up to five images if avail-

Input	Model	Generated Text
Beverly Skinny Flare Jeans 	CONTEXT-ATTN	*Tonal stitching and subtle fading add worn in character to dark wash jeans cut with a flattering figure flattering flare.*
	PSEUDO-SELF	*These stretchy bootcut jeans are inspired by the '70s super-sweet denim look that has you feeling like the real thing.*
	MANTIS	*Figure-flattering flared jeans are made from soft denim with shape-retaining stretch and a clean front for a modern silhouette.*
Azur Tassel Hem Cotton Blouse 	CONTEXT-ATTN	*With a twirl that's T-shirt cut to the natural waist, this gauzy blouse is ready to be fancy or just fun.*
	PSEUDO-SELF	*From a collaboration with fashion/lifestyle blogger Lindsey Schuster, cute pieces like this top prove that looking good can be a breeze—even on crazy-busy days.*
	MANTIS	*Shine through your work-to-play look in this gauzy blouse trimmed with embroidered tassels for a free-spirited vibe.*

Table 4: Sample outputs on the FACAD dataset.

Input	Generated Text
Cover-Up Dress 	*If you don't already have a beach vacation planned, this long-sleeve cover-up dress in a vintage floral print would like you to reconsider that.*
Velour-Hooded Jumpsuit 	*Get a luxe look in this one-and-done jumpsuit designed in sumptuous velour with a dramatic high/low hem.*
Hudson Holly High Waist Distressed Deconstructed Crop Flare Jeans 	*Essential white jeans get dashed with destruction, from the ripped knee to the slashed hem, and the end result delivers some drama for denim days and nights.*
Davis Feather Trim Cami 	*A feather-trimmed hem and spaghetti straps add dynamic finishing touches to this streamlined cami.*

Table 5: MAnTiS generated descriptions with colored annotations (green: higher quality; blue: information unique to the image; purple: textual information; orange: incorrect information).

able and added conditioning text dropout with a probability 0.3. We used the full product name of each fashion item to train all models.

5.1 Fashion Description Generations

The main quantitative results for the fashion description generation task are summarized in Table 2. MAnTiS significantly outperformed the baseline approaches in all the evaluation metrics. MAnTiS improved the BLUE4 score by 0.6, CIDEr by 4.5, METEOR by 0.5, and ROUGE-L by 0.6 respectively compared to Pseudo-Self, which shows the effectiveness of our approach.

Context-Attn adds a new layer in each transformer block, and jointly optimizing pretrained weights along with the newly initialized weights in every layer adds difficulty. We suspect this hindered the information gain from images, or even negatively impacted the generation capability of the pretrained model. Similar observations were made in uni-modal settings by Ziegler et al. (2019). We believe our approach outperformed the strong baselines because it entailed minimal interference with the pretrained language model. We introduced new parameters only at the input level, compared to Pseudo-Self which alters the self-attention module.

We next analyzed whether using multiple images improved text generation. The dataset provides several images of each fashion article. Providing MAnTiS-multi with up to 5 images increased performance over MAnTiS in all metrics, with improvements of 0.1, 2.2, 0.3 and 0.3 in BLEU4, CIDEr, METEOR and ROUGE-L respectively. This shows that the model combines information from different visual inputs. We further studied the effect of incorporating text modality dropout. Dropping out product name information randomly improved the performance slightly over MAnTiS-multi in BLEU4 and CIDEr, with no change in METEOR or ROUGE-L. This indicates that modality dropout can provide small benefits and no negative effect during optimization. As expected, MAnTiS-scratch greatly underperformed MAnTiS-single, indicating the benefit of large pretrained models.

5.2 Human Rating

Product descriptions geared towards e-commerce should entice customers using appealing phrases. Difficulty arises when analyzing such properties using automatic metrics, so we performed human evaluations to rate aspects of generated descriptions. Two random judges of different genders were tasked to score 200 product descriptions. Inspired by Dang (2005) we asked the raters to measure five linguistic qualities including grammar, non-redundancy, consistency, attractiveness, and overall scores. For the first four qualities, the task demanded only a yes/no answer. A consistent description is coherent and correct given a product's name and image, and is attractive if it is interesting or attention-grabbing. For the last category "overall", judges scored descriptions between 1 (worst) and 5 (best) from an e-commerce perspective.

The normalized average scores are shown in Table 3. MAnTiS outperformed the baseline approaches in all five categories. This shows that our conditional adaptation approach is significantly better than the previous approaches. This is likely due to the fact that MAnTiS does not introduce new parameters within the pretrained language model, unlike other approaches.

5.3 Qualitative Analysis

In Table 4 we show example generations from different models. We see that Context-Attn has repetition and incorrect information like "subtle fading", while no fading is seen in the image. MAnTiS generated higher quality descriptions highlighting image features like "trimmed with embroidered tassels".

To illustrate our results, we give a few representative MAnTiS generations from the test dataset in Table 5. We color-coded important parts of the generated text with green to indicate a high-quality phrase, blue to indicate attributes present only in the image, purple to indicate attributes from product name and orange to indicate possibly incoherent information. The blue highlighted phrases demonstrate that MAnTiS generations are guided in part by image content. In the first row, the generated description aptly connects the Cover-Up Dress to a beach setting. The model may sometimes fail to pick up on image-based cues correctly, as seen in third example where color was pronounced as "white" instead of denim blue, although this confusion is understandable as the faded regions are white. Overall, the examples show that MAnTiS can generate diverse coherent descriptions conditioned on both modalities.

6 Conclusion

In this work, we introduce MAnTiS, a novel approach for adapting pretrained language models

into multimodal conditional NLG models. We showed that our approach significantly outperforms strong baselines methods on several common NLG evaluation metrics. For a qualitative analysis, we perform human evaluations and show that our approach generates high-quality text that agrees with the conditional input. Based on several qualitative measures we show that conditionalizing a pretrained language model through new modalities does not hamper its generative capabilities.

Our approach is straightforward, easy to implement, and extendable to any modality and provides an effective way to conditionalize any pretrained language model. We believe this study will set a strong baseline in the field of multimodal NLG.

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. *arXiv preprint arXiv:1511.01432*.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. *arXiv preprint arXiv:1903.09722*.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. 2015. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style,

high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2017. Cold fusion: Training seq2seq models together with language models. *arXiv preprint arXiv:1708.06426*.

Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu, and Pascale Fung. 2019a. Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 203–211.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019b. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. 2020. Fashion captioning: Towards generating accurate descriptions with semantic rewards. *arXiv preprint arXiv:2008.02693*.

Kyra Yee, Nathan Ng, Yann N Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. *arXiv preprint arXiv:1908.05731*.

Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*.

Zachary M Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M Rush. 2019. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*.