Modelling the interplay of metaphor and emotion through multitask learning

Verna Dankers[1] **Marek Rei**[2,3] **Martha Lewis**[1] **Ekaterina Shutova**[1]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]The ALTA Institute, Computer Laboratory, University of Cambridge
[3]Department of Computing, Imperial College London
vernadankers@gmail.com, marek.rei@cl.cam.ac.uk, {m.a.f.lewis,e.shutova}@uva.nl

Abstract

Metaphors allow us to convey emotion by connecting physical experiences and abstract concepts. The results of previous research in linguistics and psychology suggest that metaphorical phrases tend to be more emotionally evocative than their literal counterparts. In this paper, we investigate the relationship between metaphor and emotion within a computational framework, by proposing the first joint model of these phenomena. We experiment with several multitask learning architectures for this purpose, involving both hard and soft parameter sharing. Our results demonstrate that metaphor identification and emotion prediction mutually benefit from joint learning and our models advance the state of the art in both of these tasks.

1 Introduction

Metaphors allow us to reason about abstract concepts by linking them to our physical experiences (Lakoff and Johnson, 1983). Metaphorical language arises through systematic association between two distinct semantic domains — the source and the target — as illustrated by the sentence "The news *leaked out* despite the secrecy", where a term from the source domain of *liquids* is used to describe *information* (the target domain). This metaphorical association widely manifests itself in language, e.g. we can similarly talk about "being *engulfed* by a *stream* of bad news". Metaphorical associations allow us to project knowledge from the source domain to the target, inviting new reasoning frameworks and connotations to emerge.

Much previous research on metaphorical language in fields such as linguistics (Blanchette et al., 2001; Kövecses, 2003), cognitive psychology (Crawford, 2009; Thibodeau and Boroditsky, 2011) and neuroscience (Aziz-Zadeh and Damasio, 2008; Jabbi et al., 2008) points to its prevalent affective content. Linguistic expressions describing one's emotional state have a relatively high incidence of figurative language and metaphor in particular (Fainsilber and Ortony, 1987; Fussell and Moss, 1998; Gibbs Jr et al., 2002), as illustrated by the phrase "My mind was *seething* and *boiling*". On the other hand, a stronger emotion appears to be conveyed through the association of source and target domains more generally. Mohammad et al. (2016) found that metaphorical phrases are consistently perceived as carrying more emotion than their literal paraphrases and the literal uses of the same source domain words. For instance, "*leaking* information" conveys an implicit judgement, as compared to the more neutral paraphrase "*disclosing* information". Their results also suggest that the emotional content of the metaphor is not due to the properties of individual source and target domains, but rather arises compositionally through their interaction. These findings are supported through a range of psycholinguistic studies: Citron and Goldberg (2014) find taste metaphors to be more emotionally evocative than their literal counterparts. Citron et al. (2016) show that conventional metaphorical language in short stories from various domains elicits more activation in brain regions involved in emotional processing, compared to literal language.

Computational modelling of metaphor (Shutova et al., 2016; Rei et al., 2017; Gao et al., 2018) and emotion (Wang et al., 2016; Zhang et al., 2018; Wu et al., 2019) are tasks widely addressed in natural language processing (NLP), with a range of applications from machine translation (Fadaee et al., 2018) to opinion mining (Yadollahi et al., 2017). However, the two phenomena have been typically modelled independently. Exceptions include the use of hand-engineered emotion features when training a classifier for metaphor identification (Strzalkowski et al., 2014) and auto-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2218–2229, Hong Kong, China, November 3–7, 2019.

matic identification of affect carried by metaphors (Kozareva, 2013; Strzalkowski et al., 2014). However, none of this research has attempted to model metaphor and emotion within a unified model of semantic composition. In this paper, we present the first joint model of metaphor and emotion, trained to learn the patterns of their interaction via flexible parameter sharing techniques offered by multitask learning (MTL). Our model is compositional, building meaning representations of words and phrases in context. The intuition is that the meaning of a word is not intrinsically metaphorical or emotional, but both of these phenomena may manifest when the word is used in a particular context.

Specifically, we train deep learning architectures on metaphor identification and emotion prediction tasks jointly. Metaphor identification is performed at word level and sentence level, while emotion prediction is modelled as a regression task, predicting numerical scores for the valence, arousal and dominance dimensions of emotion. We experiment with MTL architectures employing both hard and soft parameter sharing methods. Models employing hard parameter sharing jointly encode the lower-level word representations using layers shared among the tasks. The soft parameter sharing methods have two task-specific networks connected through linear units or gates.

Our models outperform existing approaches to both metaphor identification and emotion prediction tasks, advancing the state of the art in these areas. Moreover, we show that jointly learning both tasks within one model provides stable performance improvements across architectures.

2 Related work

2.1 Computational models of metaphor

Early approaches to metaphor identification used hand-engineered features and a trained classifier, such as logistic regression (Dunn, 2013), random forests (Tsvetkov et al., 2014), decision trees (Mohler et al., 2013; Gargett and Barnden, 2015) or support vector machines (Hovy et al., 2013; Mohler et al., 2013). Examples of linguistic features used are POS tags (Hovy et al., 2013), concreteness or imageability ratings (Turney et al., 2011; Broadwell et al., 2013; Gargett and Barnden, 2015), ontological concepts (Dunn, 2013) and WordNet super-senses (Hovy et al., 2013) and synsets (Mohler et al., 2013). To become less reliant on hand-crafted features, corpus-driven approaches emerged, using sparse (Shutova et al., 2010; Gutierrez et al., 2016) or dense word embeddings (Shutova et al., 2016; Bulat et al., 2017).

Recently, the use of deep neural networks for metaphor identification has gained popularity. Rei et al. (2017) presented a network designed to predict the metaphoricity of a word pair, by modelling the words' interaction using a gating function. Other approaches treated metaphor identification as a sequence labelling task. Do Dinh and Gurevych (2016) proposed a multi-layer perceptron acting on word embeddings. Do Dinh et al. (2018) present a MTL approach combining multiple metaphor identification tasks using two architectures: a hard parameter sharing recurrent network and the recurrent Sluice network of Ruder et al. (2019). During a recent shared task on metaphor identification, various deep neural architectures were presented (Leong et al., 2018), among which were several hybrid approaches that incorporated linguistic features in recurrent networks. Gao et al. (2018) presented the current best-performing model for metaphor sequence labelling. They employed GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018) embeddings as input to a bidirectional LSTM (Bi-LSTM) followed by a classification layer.

2.2 Computational models of emotion

The vast majority of NLP research on affective language analysis has focused on the prediction of emotion categories and sentiment analysis. Early work on emotion prediction assumed categorical models of emotion, such as Ekman's model of six emotions (Ekman, 1992) (anger, disgust, fear, joy, sadness and surprise). A variety of computational models have been proposed for emotion classification, ranging from vector space models (Danisman and Alpkocak, 2008), to machine learning classifiers (Perikos and Hatzilygeroudis, 2016) and deep learning architectures (Zhang et al., 2018).

Recently, multi-dimensional emotion analysis has gained popularity: it represents emotion through a more fine-grained and psychologically-motivated model (Buechel and Hahn, 2017). We employ the Valence-Arousal-Dominance (VAD) model (Mehrabian, 1996) that describes affective states relative to these emotional dimensions. Valence represents the polarity, arousal the degree of excitement, and dominance the perceived degree

of control over a situation.

Existing methods for dimensional emotion analysis are either lexicon-based or use supervised learning. Lexicon-based methods assume the emotional value of a sentence to be a composition of per-word values. These values are extracted from an affect lexicon (Warriner et al., 2013) and combined using their mean (Kim et al., 2010), a weighted mean, or a Gaussian Mixture Model (Paltoglou et al., 2013). Other approaches train classifiers using n-gram and sentiment features (Malandrakis et al., 2013; Buechel and Hahn, 2016), and deep learning models.

Wang et al. (2016) were among the first to present a deep learning architecture for dimensional emotion analysis using the VA model: they proposed a convolutional network operating on regions within the input, and a LSTM layer acting on the region encodings. Wang et al. (2018) used the VAD labelled corpus of Buechel and Hahn (2017), but considered only valence, effectively reducing the task to sentiment analysis. They present a deep network of stacked Bi-LSTM layers with residual connections. Akhtar et al. (2018) performed regression for all three dimensions using a convolutional and two recurrent networks combined in an ensemble extended with hand-crafted features. The emotion dimensions were considered separately and in a MTL setup. Most recently, Wu et al. (2019) proposed a variational auto-encoder model including a recurrent module trained to perform emotion prediction. The model was trained in a semi-supervised way, using only the labels of 40% of the training samples.

2.3 Metaphor and emotion

Existing work combining metaphor and emotion either focuses on the inclusion of emotion features in metaphor identification or on the automatic identification of affect carried by metaphors.

Kozareva (2013) and Strzalkowski et al. (2014) modelled the affect carried by metaphors and evaluate their approaches on a metaphor-rich corpus containing data from four languages. Kozareva (2013) performs polarity classification and valence regression using the AdaBoost classifier and support vector regression trained on information from the sentence, its context, and source and target domain annotations. Strapparava and Mihalcea (2007) proposed an affect calculus to estimate the affect expressed by a linguistic metaphor as positive, negative, or neutral. The affect calculus takes into account the metaphor target, the source relation, the relation's arguments and type, and the prior affect of the target.

Gargett and Barnden (2015) considered metaphor identification on nouns, verbs and prepositions using hand-engineered features, including lexicon-based VAD emotion features. The emotion features proved most beneficial for metaphor identification for nouns and verbs.

3 Tasks and datasets

VUA metaphor corpus The VUA metaphor corpus[1] (Steen, 2010) is a subset of the British National Corpus (Leech, 1992) in which each word is annotated as literally- or metaphorically-used. The corpus contains over ten thousand sentences, sampled from four genres: academic writing, news, conversation and fiction. The reported inter-annotator agreement is 0.84 in terms of Fleiss's κ. For comparability reasons, we use a preprocessed variant of the corpus as provided by Gao et al. (2018), who use 25% of the sentences for testing. We perform metaphor identification at word level, experimenting in a sequence labelling paradigm.

LCC metaphor corpus The Language Computer Corporation (LCC) metaphor dataset (Mohler et al., 2016) is a metaphor-rich corpus containing data in English, Farsi, Spanish and Russian.[2] We use the English portion of this dataset that consists of data from the ClueWeb corpus and the Debate Politics online forum. Annotators rated the metaphoricity of sentences from zero (i.e. literal) to three (i.e. clearly metaphorical). Mohler et al. (2016) considered agreement between annotators to be a difference of ≤ 1 (on a range from 0 to 3). With this definition, the inter-annotator agreement on metaphoricity is 92.8%.

We extract nine thousand samples from the freely available portion of the dataset, average the scores assigned by individual annotators and normalise them to the scale from zero to one. We use the data to perform sentence-level regression, employing ten-fold cross-validation using 70-10-20 splits for training, validating and testing, respectively.

[1] Preprocessed variant available at: https://github.com/gao-g/metaphor-in-context.

[2] Available upon request from Mohler et al. (2016).

Sentence	Val.	Arous.	Dom.
"Tell her I love her."	.94	.88	.83
Tell me, or I'll kill –	.35	.69	.83
What did you say?	.50	.54	.50
This is torture.	.14	.72	.27

Table 1: EmoBank examples with normalised scores, illustrating the differences among the dimensions.

EmoBank corpus EmoBank,[3] (Buechel and Hahn, 2017) is one of the most recent corpora developed based on the VAD model. EmoBank contains ten thousand sentences from the manually annotated sub-corpus of American English (Ide et al., 2008) and the Affective Text corpus (Strapparava and Mihalcea, 2007). The corpus balances many genres: news headlines, blogs, essays, fiction, letters, newspapers and travel guides. Each sentence is rated on a scale from one to five for each dimension, from the perspective of the writer and the reader. The inter-annotator agreement rates are 0.61 and 0.63 in terms of Pearson's r for the two perspectives, respectively. We combine the scores of readers and writers, normalised to the scale from zero to one. We use EmoBank to perform sentence-level regression for each of the V, A, and D dimensions separately, using ten-fold cross-validation with 70-10-20 splits for training, validating and testing, respectively. Table 1 lists examples exhibiting a range of VAD values.

Since we focus on the interaction of metaphor and emotion, we pair up the word-level and sentence-level metaphor tasks with the regression tasks for each separate dimension of V, A, or D one by one, in a MTL setup.

4 Methods

We construct a recurrent neural architecture operating at two different levels. Based on the VUA metaphor corpus, the model learns to detect metaphor at word level in a sequence labelling paradigm. When optimised on the LCC metaphor corpus, the architecture is adapted to predict a metaphoricity score at sentence level. For emotion prediction, the architecture is the same as for the sentence-level metaphor prediction task.

The system receives a tokenised sentence as input and maps it to word embeddings, by concatenating representations from pre-trained GloVe and ELMo models. Next, these embeddings are passed through a Bi-LSTM, building task- and context-dependent representations for each word. For token labelling, the hidden states from each direction are concatenated and passed through a feedforward layer, followed by a sigmoid activation. We model metaphor detection as a binary task for comparability to the literature. Gao et al. (2018) use a similar model for metaphor detection at word level, but employ a softmax activation.

For the sentence-level score prediction, the concatenated Bi-LSTM hidden states are passed through the attention function, which includes a linear layer and softmax normalisation, in order to construct a sentence representation. The resulting vector is passed through a feedforward layer, then used to predict a sentence-level score with sigmoid activation. Since we used a sigmoid ac-

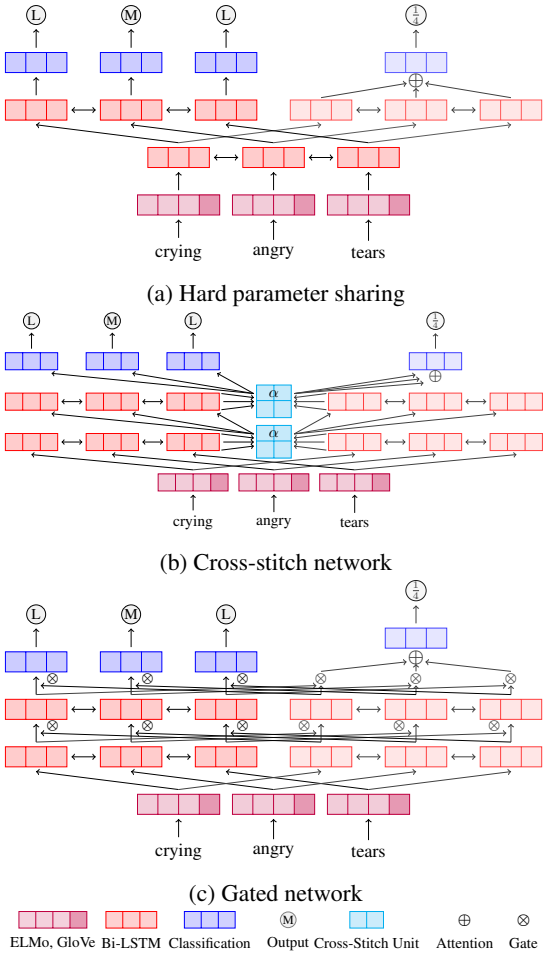


Figure 1: Overview of the MTL architectures. The parameters of the main and auxiliary task are indicated through highlighting, where the metaphor detection task is considered the main task in this setup. For compactness, two of the three Bi-LSTM layers are shown.

[3] https://github.com/JULIELab/EmoBank.

tivation function in the token labelling task, both the metaphor and emotion tasks are structurally very similar. We train for metaphor detection using the binary cross-entropy loss function and for regression using the root mean squared error loss function.

We also experiment with fine-tuning a pre-trained BERT architecture (Devlin et al., 2019) for each of the tasks. This validates that the performance differences are due to the task interactions and not specific to the recurrent architecture. The inputs to this network consist of the BERT-specific word and position embeddings. For the word-level sequence labelling task, the outputs of the last Transformer layer are fed to the classification layer. For the sentence-level tasks, an additional attention module is again used to construct sentence representations. An alternative way of using BERT would be to provide contextualised embeddings. We do not consider this in the present work but leave it as an area to be explored. BERT performs labelling on subword units called WordPieces; we consider a word metaphorical if any of these subword units is labelled metaphorical. This choice is motivated by the fact that although a metaphorical prefix or suffix could result in an incorrect metaphorical label, this is unlikely: what is much more likely is that a common prefix or suffix is not considered metaphorical while the main piece is.

In the following sections we describe three different approaches to optimising these networks using MTL. We experiment with four setups: one recurrent and one BERT hard parameter sharing setup, one recurrent cross-stitch network and one gated recurrent network. In the MTL setups, the models are trained on two tasks at once, but we distinguish between a main and auxiliary task by down-weighting the loss of the auxiliary task to allow the networks to specialise most in one direction. For example, when seeking performance improvements in the word-level metaphor identification task, metaphor identification is the main task and emotion regression is the auxiliary task.

4.1 Hard parameter sharing

We customise the models to jointly perform metaphor detection and emotion prediction. By training the model to identify emotional states in text, the system learns to recognise emotion-related features which can be useful for the task of metaphor detection. In addition, optimising for two different but related tasks helps prevent the model from overfitting to either of them.

Following established work on MTL (Caruana, 1993), we first experiment with hard parameter sharing. In this setting, the architecture shares the word embeddings and lower Bi-LSTM layers between the two tasks, as shown in Figure 1a. On top of these shared components, each task has one separate Bi-LSTM layer, followed by a task-specific output layer. For sentence scoring, the attention function for constructing sentence representations is also learned individually for every task. This setup allows the model to learn shared feature detectors in the lower layers, while top layers are still able to learn task-specific features.

The hard parameter sharing setup using BERT shares all of BERT's Transformer layers among the tasks, apart from the last layer to allow for specialisation. Furthermore, the output and attention layers are task-specific as well.

4.2 Cross-stitch network

As an alternative to hard parameter sharing, soft sharing provides parallel models with dedicated parameters for each task, while also connecting them together to allow for information transfer. In the cross-stitch network, the soft sharing mechanism is a cross-stitch unit (Misra et al., 2016). These units contain α-parameters which regulate the information flow in each direction and are optimised during training. We apply cross-stitch sharing after each recurrent layer, computing the updated hidden states as:

$$\widetilde{\mathbf{h}}_A = \alpha_{AA}\mathbf{h}_A + \alpha_{BA}\mathbf{h}_B \quad (1)$$

$$\widetilde{\mathbf{h}}_B = \alpha_{BB}\mathbf{h}_B + \alpha_{AB}\mathbf{h}_A \quad (2)$$

where $\mathbf{h}_A$ and $\mathbf{h}_B$ are the concatenated Bi-LSTM hidden states, from parallel networks for tasks A and B, while $\widetilde{\mathbf{h}}_A$ and $\widetilde{\mathbf{h}}_B$ are the updated hidden states. Note that the α-parameters are specific to each layer. The α-parameters control the directions of information flow; for example, α_{AB} scales the information passed from network A to network B. The cross-stitch network is shown in Figure 1b.

If both tasks operate at sentence level, an additional cross-stitch sharing unit is placed after the attention module. The α-parameters are initialised with a bias towards favouring the information in the same network, with $\alpha_{AA} = \alpha_{BB} = 0.9$ and $\alpha_{AB} = \alpha_{BA} = 0.1$. These values are optimised

during training but remain static during testing.

4.3 Gated network

The cross-stitch network learns a single set of shared values for the α-parameters during optimisation. As an alternative, we can construct a network that calculates these values dynamically for each input sentence, even at testing time. This allows the model greater flexibility and modulates the information flow depending on the particular input sentence.

In this architecture, shown in Figure 1c, the α-parameters are replaced with gates (Liu et al., 2016). Each pair of parallel layers has two gates, where one modulates the information flow from the main to the auxiliary task, while the other controls the information flow in the opposite direction. For two jointly learned sentence-level tasks, two more gates are placed before the classification layer, operating on the sentence representations.

Equations (3)-(6) detail the gating mechanisms:

$$\mathbf{g}_A = \sigma(\mathbf{W}_A[\mathbf{h}_A; \mathbf{h}_B] + \mathbf{b}_A) \quad (3)$$

$$\widetilde{\mathbf{h}}_A = (1 - \mathbf{g}_A) \odot \mathbf{h}_A + \mathbf{g}_A \odot \mathbf{h}_B \quad (4)$$

$$\mathbf{g}_B = \sigma(\mathbf{W}_B[\mathbf{h}_A; \mathbf{h}_B] + \mathbf{b}_B) \quad (5)$$

$$\widetilde{\mathbf{h}}_B = (1 - \mathbf{g}_B) \odot \mathbf{h}_B + \mathbf{g}_B \odot \mathbf{h}_A \quad (6)$$

where $\mathbf{g}_A$ and $\mathbf{g}_B$ are the gates for task A and B, $\mathbf{W}_A$ and $\mathbf{W}_B$ are weight matrices, $\mathbf{b}_A$ and $\mathbf{b}_B$ are bias vectors. The bias parameters of the gates are initialised with a bias towards one task.

5 Experiments and results

5.1 Experimental setup

MTL training procedure We apply pairwise joint learning, where at each step in the training process one of the two tasks is selected at random and a batch is sampled from that task. To distinguish main tasks from auxiliary tasks the loss of the auxiliary task is down-weighted by a factor λ such that it comprises 10% of the loss of the main task. λ is initialised with $\frac{1}{10}$ and computed dynamically as training progresses.

Hyperparameters The input to the recurrent network consists of concatenated ELMo and GloVe embeddings, with $1,024$ and 300 dimensions respectively. The recurrent encoder contains three Bi-LSTM layers with a dimensionality of 200. The models are trained using a batch size of 64 for $2,000$ steps and the Adam optimiser with initial learning rates of $4\text{e}{-3}$, $1\text{e}{-3}$ and $0.5\text{e}{-3}$ for metaphor detection, metaphor regression and emotion regression respectively. Models are selected based on validation data.

Approach	Metaphor Task	
	Word (F_1)	Sent. (r)
Gao et al. (2018)	.726	-
LSTM (single task)	.737	.544
Hard Sharing		
+ Valence	.740	**.559**
+ Arousal	.740	**.558**
+ Dominance	**.743**	**.560**
Cross-Stitch Network		
+ Valence	.741	**.556**
+ Arousal	.740	**.558**
+ Dominance	**.743**	**.563**
Gated Network		
+ Valence	**.742**	**.561**
+ Arousal	.741	**.558**
+ Dominance	**.745**	**.560**
BERT (single task)	.763	.604
Hard Sharing		
+ Valence	**.769**	**.614**
+ Arousal	.765	.610
+ Dominance	**.768**	**.614**

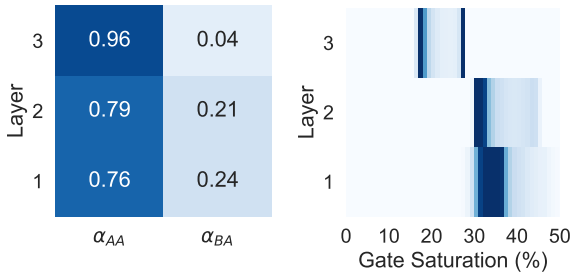
Table 2: System performance for the word- and sentence-level metaphor tasks using the F_1-score and Pearson's r respectively. Statistically significant ($p < 0.05$) differences to the single task models are shown in boldface.

We employ the pretrained BERT Base model, whose inputs and hidden states have 768 dimensions. The model contains 12 Transformer layers and is fine-tuned as described by Devlin et al. (2019), using the Adam optimiser with an initial learning rate of $5\text{e}{-5}$ and a batch size of 32. BERT is fine-tuned for $3,000$ steps for the regression tasks and for $8,000$ steps for the word-level metaphor detection task. The difference is compensated for through down-scaling λ.

Significance testing We test for statistical significance using the one-sided approximate randomisation test (Edgington, 1969) for metaphor detection, and Williams's test (Williams, 1959) for regression tasks. For Williams's test we consider the number of samples to be the number of unique samples in the dataset. All performance measures reported are averages from models initialised with ten random seeds.

Approach	Emotion Task		
	Val.	Arous.	Dom.
Akhtar et al. (2018)	.616	.355	.237
+ Val., Arous., Dom.	.635	.375	.277
Wu et al. (2019)†	.620	.508	.333
LSTM (single task)	.728	.557	.373
Hard Sharing			
+ Metaphor (Token)	**.734**	**.564**	**.384**
+ Metaphor (Sent.)	**.734**	.558	**.388**
Cross-Stitch Network			
+ Metaphor (Token)	**.737**	**.564**	**.388**
+ Metaphor (Sent.)	**.735**	.558	**.384**
Gated Network			
+ Metaphor (Token)	**.738**	**.563**	**.389**
+ Metaphor (Sent.)	**.735**	.560	**.384**
BERT (single task)	.771	.565	.403
Hard Sharing			
+ Metaphor (Token)	**.779**	**.572**	**.420**
+ Metaphor (Sent.)	**.778**	**.570**	**.417**

Table 3: System performance for emotion regression tasks according to Pearson's r. Statistically significant ($p < 0.05$) differences to the single task model are shown in boldface. †Used 40% of the gold labels.



(a) Cross-Stitch Units (b) Gates

Figure 2: Illustration of the information flow in between the Bi-LSTM layers, for the dominance regression (B) and metaphor identification (A) tasks. Gate saturation % is calculated by averaging across the hidden dimensionality for every word in the test set.

5.2 Results

Table 2 presents the results for the two metaphor tasks. The STL setup already provides improvements over the current state of the art, but moreover, we see further improvements when MTL is introduced. Each MTL setup should be compared to the corresponding STL setup, which involves training the model on the metaphor task only. Regardless of the MTL architecture, the auxiliary task of dominance regression provides statistically significant ($p < 0.05$) improvements over the STL setup. Furthermore, valence regression provides significant improvements as well in a select number of setups. The largest improvement is achieved by replacing the recurrent encoder with BERT. This indicates that the rich contextual information learned by BERT in the pre-training phase is highly relevant for metaphor identification. For sentence-level metaphor regression, MTL setups consistently improve upon STL setups, indicating that the effect is not specific to the VUA metaphor corpus. Our MTL models outperform the previous state of the art in metaphor identification (F_1 of 0.726 on the VUA corpus Gao et al. (2018)) with both LSTM (F_1 of 0.745) and BERT (F_1 of 0.769) encoders. These results lend support to the hypothesis on the interaction of metaphor and emotion in semantic composition.

Table 3 presents the results for the emotion regression tasks. Again, STL setups perform strongly, and MTL architectures improve this even further. While the valence and dominance tasks consistently improve with the addition of the metaphor task, the improvements achieved on arousal regression are less stable. Once again, our MTL models outperform the best-performing existing approaches to dimensional emotion modelling (Akhtar et al., 2018; Wu et al., 2019), advancing the state of the art in this task. These results suggest that it may be beneficial to include information about metaphor into emotion analysis and, more broadly, sentiment analysis systems.

The differences between hard and soft parameter sharing manifest most with metaphor identification. This is possibly due to the explicit per-word sharing mechanisms in the top layer. While the bottom layers capture more general information, the top layer captures task-specific information. This can be seen through analysing the models' behaviour: the gating mechanism is most selective at the top layer and is the most active for emotion-laden words. The cross-stitch units gradually share less information from the bottom to the top layer. This behaviour is illustrated in Figure 2.

6 Data analysis and discussion

Metaphor identification Most improvements in word-level metaphor identification are achieved by corrections from a literal prediction in STL to metaphorical in MTL. To establish this fact,

Auxiliary Task	Sentence	STL	MTL	Gold
Valence	It is sad, and somewhat ominous, that so little of that should have been reflected in the **sombre** statement (. . .)	L	M	M
Arousal	There is still endless dithering on how broad a **safety** net Britain will extend to its citizens.	L	M	M
Dominance	In a **bare**, mud-walled cell, sitting on the floor, is Tepilit.	L	M	M

Table 4: Examples of how MTL improves over STL on metaphor identification, with gold metaphors underlined and corrections by MTL bolded. L refers to literal and M to metaphorical.

Main Task	Sentence	STL	MTL	Gold
Valence	Scam lures victims with free puppy offer.	.48	.41	.40
Valence	(. . .) looking at me like I was breaking her poor, sweet heart.	.52	.42	.24
Arousal	(. . .) the authors depict her as bewitchingly beautiful.	.51	.57	.61
Dominance	In fact, I've never felt so out of place in all my life.	.73	.54	.17
Dominance	Frustration rises as North Korea nuclear talks stall.	.46	.41	.30

Table 5: Examples of how MTL using metaphor identification improves over STL emotion prediction, with predicted metaphors underlined.

we pooled predictions on test data for ten models trained with different random seeds. The STL outputs were compared to MTL outputs to determine whether corrections changed from literal to metaphorical or the other way around. Examining this set of corrected predictions gives us insight into the behaviour of the MTL model.

While some improvements hold across all emotion dimensions, others are unique to each. Among the corrections unique to each dimension we find multiple terms indicative of the dimension: for valence regression we find improvements for attractiveness (*wise*, *attractive*) and averseness indicators (*severity*, *drain*) and for arousal excitement (*flame*, *crisis*) or calmness indicators (*empty*, *rest*). Dominance corrects various terms related to control (*capable*, *courtesy*) and submissiveness (*owe*, *puny*). We selected the presented example terms from the set of corrections described previously, and established the scores using the ANEW lexicon (Warriner et al., 2013).

Table 4 illustrates model decisions corrected through joint learning. Examples for valence and arousal illustrate how emotion-laden words participate in metaphors, such as "a *sombre* statement" or "a *safety* net". The example for dominance illustrates that control indicators may seem less related to emotion (e.g. *bare*), but carry affect through the contexts in which they are embedded.

Overall, introducing emotion improves performance of metaphor identification, as we expected. Dominance appears to be most beneficial, while arousal contributes the least. This might be due to the fact that arousal (and to some extent valence) predictions rely strongly on explicit sentiment markers in text. In contrast, dominance prediction requires the model to learn richer semantic representations. In our models, this manifests in the sparseness of the attention distribution: models trained using arousal have the most sparse attention patterns as measured through the Gini index sparsity measure (Hurley and Rickard, 2009), while models trained using dominance have the least sparse attention patterns.

Dominance regression is the most complex task and yet the most beneficial performance-wise, despite it sometimes being discarded by previous research in favour of the VA emotion model. Several studies argue for the inclusion of dominance in emotion analysis (Stamps III, 2005; Bakker et al., 2014). Bakker et al. (2014) emphasise that while valence and arousal highlight the affective and cognitive aspects of emotion, dominance is related to environmental factors and social influences. This relates to the role of metaphorical framing in the social world, e.g. in politics. Metaphor allows us to highlight certain aspects of a target domain and mask others, encouraging specific inferences (Lakoff, 1991; Entman, 1993). These in turn activate emotional considerations (Boeynaems et al.,

2017), allowing the metaphor to steer the emotions recalled and affecting the evaluation of the argument made and persuasiveness of the speaker (Marcus, 2000).

Emotion prediction Table 5 lists examples for which including metaphor identification improved performance on emotion regression. The examples for valence show that metaphor can be used to describe an emotion ("to *break* one's *heart*") explicitly or to convey an implicit judgement (e.g. *luring*). For arousal, example improvements include applying excitement indicators to objects or concepts, such as "being *bewitchingly* beautiful". Examples for dominance regression indicate the importance of power; one has no control over *stalling* or "feeling *out of place*".

While joint learning improves the estimations overall, it also introduces some new errors. Although generally the presence of metaphor makes phrases more emotionally evocative (Mohammad et al., 2016), this does not always hold. Lexicalised metaphors – e.g. *up* in "grades going *up*" – are no longer viewed as metaphorical by lay language users and throw the models off. Other common errors introduced are related to misinterpreting the perspective – e.g. confusing "to be *knocked out*" and "to *knock out*" – and the direction in which words contribute to emotion – e.g. negative metaphorical terms can contribute to the positive sentiment and vice versa, such as in "Stop cancer with a *shot*".

7 Conclusion

In this paper, we introduced the first compositional deep learning model to jointly capture the phenomena of metaphor and emotion. We considered metaphor tasks at word and sentence level and modelled emotion through the dimensional model of valence, arousal and dominance. We experimented with multiple MTL techniques, regulating the information flow between the two tasks.

We demonstrated that the proposed methods advance the state of the art for the tasks of metaphor identification and emotion regression. Both tasks benefit from joint learning, with the emotion dimension of dominance contributing most to metaphor and benefiting most from metaphor. Our results support the hypothesis on the interaction of metaphor and emotion, and suggest that it may be beneficial to incorporate a model of metaphor into emotion- and sentiment-related NLP applications in the future.

Acknowledgments

Martha Lewis gratefully acknowledges funding from NWO Veni grant 'Metaphorical Meanings for Artificial Agents'. Marek Rei's research is supported by Cambridge English via the ALTA Institute.

References

Md Shad Akhtar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2018. A multi-task ensemble framework for emotion, sentiment and intensity prediction. *arXiv preprint arXiv:1808.01216*.

Lisa Aziz-Zadeh and Antonio Damasio. 2008. Embodied semantics for actions: Findings from functional brain imaging. *Journal of Physiology - Paris*, 102(1-3).

Iris Bakker, Theo van der Voordt, Peter Vink, and Jan de Boon. 2014. Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Current Psychology*, 33(3):405–421.

Isabelle Blanchette, Kevin Dunbar, John Hummel, and Richard Marsh. 2001. Analogy use in naturalistic settings: The influence of audience, emotion and goals. *Memory and Cognition*, 29(5).

Amber Boeynaems, Christian Burgers, Elly A Konijn, and Gerard J Steen. 2017. The effects of metaphorical framing on political persuasion: A systematic literature review. *Metaphor and Symbol*, 32(2):118–134.

George Aaron Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 102–110. Springer.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem—dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 1114–1122. IOS Press.

Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 578–585.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 523–528.

Rich Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the International Conference on Machine Learning*.

Francesca MM Citron and Adele E Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of cognitive neuroscience*, 26(11):2585–2595.

Francesca MM Citron, Jeremie Güsten, Nora Michaelis, and Adele E Goldberg. 2016. Conventional metaphors in longer passages evoke affective brain response. *NeuroImage*, 139:218–230.

Elizabeth Crawford. 2009. Conceptual Metaphors of Affect. *Emotion Review*, 1(2).

Taner Danisman and Adil Alpkocak. 2008. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 53.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Erik-Lân Do Dinh, Steffen Eger, and Iryna Gurevych. 2018. Killing four birds with two stones: Multi-task learning for non-literal language detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1558–1569.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33.

Jonathan Dunn. 2013. Evaluating the premises and results of four metaphor identification systems. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 471–486. Springer.

Eugene S Edgington. 1969. Approximate randomization tests. *The Journal of Psychology*, 72(2):143–149.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 925–929.

Lynn Fainsilber and Andrew Ortony. 1987. Metaphorical uses of language in the expression of emotions. *Metaphor and Symbol*, 2(4):239–250.

Susan R Fussell and Mallie M Moss. 1998. Figurative language in emotional communication. *Social and cognitive approaches to interpersonal communication*, pages 113–141.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural Metaphor Detection in Context. In *Conference on Empirical Methods in Natural Language Processing*.

Andrew Gargett and John Barnden. 2015. Modeling the interaction between sensory and affective meanings for detecting metaphor. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 21–30.

Raymond W Gibbs Jr, John S Leggitt, and Elizabeth A Turner. 2002. What's special about figurative language in emotional communication? In *The verbal communication of emotions*, pages 133–158. Psychology Press.

E Dario Gutierrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 183–193.

Dirk Hovy, Shashank Shrivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57.

Niall Hurley and Scott Rickard. 2009. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. Masc: The manually annotated sub-corpus of American English. In *6th International Conference on Language Resources and Evaluation*, pages 2455–2460. European Language Resources Association (ELRA).

Mbemba Jabbi, Jojanneke Bastiaansen, and Christian Keysers. 2008. A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PLoS ONE*, 3(8):e2939.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70. Association for Computational Linguistics.

Zoltán Kövecses. 2003. *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press, Cambridge.

Zornitsa Kozareva. 2013. Multilingual affect polarity and valence prediction in metaphor-rich texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 682–691.

George Lakoff. 1991. Metaphor and war: The metaphor system used to justify war in the Gulf. *Peace Research*, 23(2/3):25–32.

George Lakoff and Mark Johnson. 1983. Metaphors we live by. *University of Chicago Press*, 59.

Geoffrey Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Chee Wee Ben Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2873–2879. AAAI Press.

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392.

George E Marcus. 2000. Emotions in politics. *Annual Review of Political Science*, 3(1):221–250.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 4221–4227.

Georgios Paltoglou, Mathias Theunis, Arvid Kappas, and Mike Thelwall. 2013. Predicting emotional responses to long informal text. *IEEE Transactions on Affective Computing*, 1(4):106–115.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Isidoros Perikos and Ioannis Hatzilygeroudis. 2016. Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence*, 51:191–201.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1537–1546.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 4822–4829.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics.

Arthur E Stamps III. 2005. In search of dominance: The case of the missing dimension. *Perceptual and motor skills*, 100(2):559–566.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Tomek Strzalkowski, Samira Shaikh, Kit Cho, George Aaron Broadwell, Laurie Feldman, Sarah Taylor, Boris Yamrom, Ting Liu, Ignacio Cases, Yuliya Peshkova, et al. 2014. Computing affect in metaphors. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 42–51.

Paul H. Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE*, 6(2):e16782.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 248–258.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690. Association for Computational Linguistics.

Jin Wang, Bo Peng, and Xuejie Zhang. 2018. Using a stacked residual LSTM model for sentiment intensity prediction. *Neurocomputing*, 322:93–101.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 225–230.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4):1191–1207.

Evan James Williams. 1959. *Regression Analysis*. Wiley publication in applied statistics. Wiley.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Zhigang Yuan, Junxin Liu, and Yongfeng Huang. 2019. Semi-supervised dimensional sentiment analysis with variational autoencoder. *Knowledge-Based Systems*, 165:30–39.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):25.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018. Text emotion distribution learning via multi-task convolutional neural network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4595–4601. AAAI Press.