Large-scale Pretraining for Visual Dialog: A Simple State-of-the-Art Baseline (Supplementary)

Vishvak Murahari[1] Dhruv Batra[1,2] Devi Parikh[1,2] Abhishek Das[1→2]

[1] Georgia Institute of Technology [2] Facebook AI Research

{vishvak.murahari,dbatra,parikh,abhshkdz}@gatech.edu

1 Negative Results

To encourage the model to learn sentence level semantics, we tried a pretraining strategy which we refer to as the inconsistency loss. During pretraining, we randomly select answers at 3 different rounds and replace them randomly with one of the 100 answer options at those rounds. Similar to the masked language modeling loss, at each token, we predict the probability of the token being "inconsistent" in the dialog history and we assume that all the tokens in the randomly selected answer option are "inconsistent". We also only consider dialog sequences which have at least 6 rounds for creating samples. This is to make sure that the model has enough context to make accurate predictions. We hoped that this loss would encourage the model to capture sentence level semantics as the model would need to figure out which sentences fit in together in a dialog sequence.

A pitfall in this setting is that some of the 100 answer options at each round are often similar to the GT answer or are generic responses (*e.g.* "yes", "no", "maybe", *etc.*). Thus, there is a chance that swapping the GT answer with a randomly selected answer option might lead to a consistent dialog sequence. We instead try to create a new sample by randomly selecting a round and reordering/jumbling the answers at that round and the answers at the preceding and the following rounds. We hope that jumbling the order of answers would lead to an inconsistent dialog sequence. We call this variant "inconsistency loss (jumbled)".

We cannot use the batch of data used to calculate the NSP and MLM loss to calculate the inconsistency loss. We first try to use half the batch to calculate the inconsistency loss and the other half to calculate the NSP and MLM losses. We then try to calculate the inconsistency loss and the NSP and MLM losses in alternating batches. We present results for models trained with multiple variants of the inconsistency loss in the language-only setting in Table 1.

We do not see any significant improvement by training with the inconsistency loss. We note that creating samples by reordering answers in different rounds does not improve performance. We also note that optimizing for the inconsistency loss and the NSP and MLM losses in alternating batches leads to improvements.

Table 1: Performance of language-only models on VisDial v1.0 val, trained with different sets of pretraining losses including the inconsistency loss.

Variant	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	MR ↓
NSP + MLM	**57.22**	64.10	**50.05**	81.09	90.00	4.16
NSP + MLM + Inconsistency	56.22	64.13	49.88	81.35	**90.11**	**4.01**
NSP + Inconsistency	55.66	63.00	48.63	80.61	89.36	4.20
NSP + Inconsistency (jumbled)	55.83	63.08	48.71	80.47	89.49	4.19
NSP + MLM + Inconsistency (alternating batches)	57.17	**64.21**	50.04	**81.48**	90.02	4.03

2 Additional results: Dense annotation finetuning

We vary the coefficients for CE and NSP losses during dense annotation finetuning (α for CE loss and $1-\alpha$ for NSP loss) and report results in Table 2. α acts as a lever balancing metrics based on dense annotations (NDCG) and metrics based on sparse annotations (MRR, R@1, R@5, R@10).

Table 2: Performance on VisDial v1.0 val for variants finetuned on dense annotations with a coefficient of α for the CE loss and $1-\alpha$ for the NSP loss. α balances NDCG and MRR performance with higher α leading to higher NDCG and lower MRR.

α	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	MR ↓
0.25	68.29	**65.96**	**53.34**	**81.42**	**90.37**	**4.12**
0.5	69.04	65.58	53.04	80.64	90.01	4.40
0.75	71.17	62.88	49.49	78.91	88.83	4.39
1.0 (CE)	**75.10**	52.12	39.84	64.93	80.47	6.26

3 Qualitative examples

We present some additional qualitative samples in Fig. 1 and Fig. 2.

Image + Caption	GT Dialog	Base Model	Base Model + CE	Base Model + CE + NSP
 the intersection had a red light and wet streets	**Q:** is the light lit up? **A:** yes **Q:** are there cars? **A:** yes **Q:** is this a city area? **A:** maybe , i'm not sure **Q:** is there people? **A:** no **Q:** is it daytime? **A:** no **Q:** is there a street sign? **A:** no **Q:** is it dark? **A:** yes **Q:** is the light metal? **A:** yes **Q:** can you see buildings? **A:** yes **Q:** can you see any trees? **A:** yes	**Q:** is the light lit up? **A:** yes **Q:** are there cars? **A:** yes **Q:** is this a city area? **A:** yes **Q:** is there people? **A:** no **Q:** is it daytime? **A:** no **Q:** is there a street sign? **A:** no **Q:** is it dark? **A:** yes **Q:** is the light metal? **A:** yes **Q:** can you see buildings? **A:** yes **Q:** can you see any trees? **A:** yes	**Q:** is the light lit up? **A:** yes **Q:** are there cars? **A:** yes **Q:** is this a city area? **A:** yes **Q:** is there people? **A:** no **Q:** is it daytime? **A:** yes **Q:** is there a street sign? **A:** yes **Q:** is it dark? **A:** yes **Q:** is the light metal? **A:** no **Q:** can you see buildings? **A:** yes **Q:** can you see any trees? **A:** yes	**Q:** is the light lit up? **A:** yes **Q:** are there cars? **A:** yes **Q:** is this a city area? **A:** yes **Q:** is there people? **A:** yes, at least 1 **Q:** is it daytime? **A:** yes **Q:** is there a street sign? **A:** yes **Q:** is it dark? **A:** yes **Q:** is the light metal? **A:** yes **Q:** can you see buildings? **A:** yes **Q:** can you see any trees? **A:** yes
 2 toothbrushes sit in a blue cup in a bathroom	**Q:** are there any people? **A:** no **Q:** what color is the cup? **A:** blue **Q:** what color are the toothbrushes? **A:** red and blue **Q:** is the sink visible? **A:** no **Q:** is a cabinet visible? **A:** no **Q:** are they adult toothbrushes? **A:** they look like it **Q:** is the ground visible? **A:** no **Q:** is there a mirror there? **A:** yes, i can see the corner of it **Q:** it a close up? **A:** yes **Q:** is the counter visible? **A:** no	**Q:** are there any people? **A:** no **Q:** what color is the cup? **A:** black **Q:** what color are the toothbrushes? **A:** orange **Q:** is the sink visible? **A:** no **Q:** is a cabinet visible? **A:** no **Q:** are they adult toothbrushes? **A:** yes **Q:** is the ground visible? **A:** no **Q:** is there a mirror there? **A:** no **Q:** it a close up? **A:** yes **Q:** is the counter visible? **A:** no	**Q:** are there any people? **A:** no **Q:** what color is the cup? **A:** black **Q:** what color are the toothbrushes? **A:** white **Q:** is the sink visible? **A:** no **Q:** is a cabinet visible? **A:** no **Q:** are they adult toothbrushes? **A:** yes **Q:** is the ground visible? **A:** no **Q:** is there a mirror there? **A:** no **Q:** it a close up? **A:** yes **Q:** is the counter visible? **A:** no, it is not	**Q:** are there any people? **A:** no **Q:** what color is the cup? **A:** black **Q:** what color are the toothbrushes? **A:** orange **Q:** is the sink visible? **A:** no **Q:** is a cabinet visible? **A:** no **Q:** are they adult toothbrushes? **A:** yes **Q:** is the ground visible? **A:** no **Q:** is there a mirror there? **A:** no **Q:** it a close up? **A:** yes **Q:** is the counter visible? **A:** yes
 an elephant eats large amounts of foliage as another elephant stands nearby	**Q:** is the elephant a baby? **A:** no **Q:** is he eating from a tree? **A:** no the ground **Q:** are they outside? **A:** yes **Q:** is there grass? **A:** yes **Q:** are there trees around? **A:** yes **Q:** is it sunny? **A:** it is hard to tell **Q:** is the food in his mouth? **A:** yes **Q:** do the leaves look fresh? **A:** yes **Q:** is the other elephant watching him? **A:** no **Q:** what color is the grass? **A:** green	**Q:** is the elephant a baby? **A:** i think it's probably a young elephant it doesn't appear to be very big **Q:** is he eating from a tree? **A:** yes **Q:** are they outside? **A:** yes **Q:** is there grass? **A:** yes **Q:** are there trees around? **A:** yes **Q:** is it sunny? **A:** yes **Q:** is the food in his mouth? **A:** yes **Q:** do the leaves look fresh? **A:** yes **Q:** is the other elephant watching him? **A:** no **Q:** what color is the grass? **A:** brown	**Q:** is the elephant a baby? **A:** i think so **Q:** is he eating from a tree? **A:** yes **Q:** are they outside? **A:** yes **Q:** is there grass? **A:** yes **Q:** are there trees around? **A:** yes **Q:** is it sunny? **A:** yes **Q:** is the food in his mouth? **A:** yes **Q:** do the leaves look fresh? **A:** yes **Q:** is the other elephant watching him? **A:** i can't tell **Q:** what color is the grass? **A:** brownish	**Q:** is the elephant a baby? **A:** yes **Q:** is he eating from a tree? **A:** yes **Q:** are they outside? **A:** yes **Q:** is there grass? **A:** yes **Q:** are there trees around? **A:** yes **Q:** is it sunny? **A:** yes **Q:** is the food in his mouth? **A:** yes **Q:** do the leaves look fresh? **A:** yes **Q:** is the other elephant watching him? **A:** no **Q:** what color is the grass? **A:** brown and green
 a lemon, peach, and another piece of fruit all sitting next to each other	**Q:** any people? **A:** no **Q:** how many lemons? **A:** 1 **Q:** is it large? **A:** no **Q:** is the fruit ripe? **A:** yes **Q:** has it been sliced? **A:** no **Q:** is there a plate? **A:** can't tell **Q:** what is the other fruit? **A:** apple peach **Q:** is this on a table? **A:** can't tell it is a very close up shot **Q:** is this in a kitchen? **A:** can't tell **Q:** indoors? **A:** no	**Q:** any people? **A:** no **Q:** how many lemons? **A:** 1 **Q:** is it large? **A:** yes **Q:** is the fruit ripe? **A:** yes **Q:** has it been sliced? **A:** no **Q:** is there a plate? **A:** no **Q:** what is the other fruit? **A:** apple peach **Q:** is this on a table? **A:** yes **Q:** is this in a kitchen? **A:** no **Q:** indoors? **A:** yes	**Q:** any people? **A:** no **Q:** how many lemons? **A:** 2 **Q:** is it large? **A:** yes **Q:** is the fruit ripe? **A:** yes **Q:** has it been sliced? **A:** no **Q:** is there a plate? **A:** no **Q:** what is the other fruit? **A:** apple peach **Q:** is this on a table? **A:** can't tell **Q:** is this in a kitchen? **A:** can't tell **Q:** indoors? **A:** can't tell	**Q:** any people? **A:** no **Q:** how many lemons? **A:** 2 **Q:** is it large? **A:** no **Q:** is the fruit ripe? **A:** yes **Q:** has it been sliced? **A:** no **Q:** is there a plate? **A:** no **Q:** what is the other fruit? **A:** apple peach **Q:** is this on a table? **A:** can't tell **Q:** is this in a kitchen? **A:** can't tell **Q:** indoors? **A:** yes

Fig. 1: Qualitative samples for three model variants – ViLBERT w/ CC + VQA (called 'Base'), Base + CE, and Base + CE + NSP.

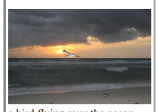
 the door of an orange, white, and black train car is open, with steps leading up to it	**Q:** how many steps are there? **A:** 3 **Q:** is it a passenger train? **A:** yes **Q:** what color are the steps? **A:** black ladder **Q:** are there people on the steps? **A:** no **Q:** can you see windows? **A:** yes **Q:** how many cars are there? **A:** only see 1, too close up to see much **Q:** are there windows? **A:** just the 1 **Q:** what does the weather look like? **A:** hazy **Q:** what can you see through the window? **A:** sky **Q:** what color is the sky? **A:** whitish pale blue	**Q:** how many steps are there? **A:** 2 **Q:** is it a passenger train? **A:** yes **Q:** what color are the steps? **A:** black **Q:** are there people on the steps? **A:** no **Q:** can you see windows? **A:** yes **Q:** how many cars are there? **A:** 1 **Q:** are there windows? **A:** yes **Q:** what does the weather look like? **A:** sunny **Q:** what can you see through the window? **A:** no **Q:** what color is the sky? **A:** white	**Q:** how many steps are there? **A:** i cannot tell **Q:** is it a passenger train? **A:** yes, it is **Q:** what color are the steps? **A:** white **Q:** are there people on the steps? **A:** no **Q:** can you see windows? **A:** yes **Q:** how many cars are there? **A:** just 1 **Q:** are there windows? **A:** yes **Q:** what does the weather look like? **A:** cloudy **Q:** what can you see through the window? **A:** no **Q:** what color is the sky? **A:** white	**Q:** how many steps are there? **A:** 2 **Q:** is it a passenger train? **A:** yes **Q:** what color are the steps? **A:** gray **Q:** are there people on the steps? **A:** no **Q:** can you see windows? **A:** yes **Q:** how many cars are there? **A:** 1 **Q:** are there windows? **A:** yes **Q:** what does the weather look like? **A:** cloudy **Q:** what can you see through the window? **A:** nothing **Q:** what color is the sky? **A:** blue
 a bunch of vegetables on a shelf at the market	**Q:** is this an outdoor or indoor market? **A:** you can only see the table so i am unsure **Q:** does the shelf have many layers? **A:** just 2 that i can see **Q:** is the picture in color? **A:** yes **Q:** is the shelf full? **A:** yes **Q:** do the vegetables look fresh? **A:** yes **Q:** are there carrots? **A:** not that i see **Q:** are there cabbages? **A:** no **Q:** what kind of vegetable has the most number present? **A:** sugar snap beans **Q:** are there labels on the shelf? **A:** yes **Q:** are there prices? **A:** yes	**Q:** is this an outdoor or indoor market? **A:** outdoors **Q:** does the shelf have many layers? **A:** can't see the top **Q:** is the picture in color? **A:** yes **Q:** is the shelf full? **A:** yes **Q:** do the vegetables look fresh? **A:** yes **Q:** are there carrots? **A:** no mostly veggies **Q:** are there cabbages? **A:** no **Q:** what kind of vegetable has the most number present? **A:** squash, potatoes **Q:** are there labels on the shelf? **A:** yes **Q:** are there prices? **A:** no	**Q:** is this an outdoor or indoor market? **A:** can't really tell **Q:** does the shelf have many layers? **A:** no, it does not **Q:** is the picture in color? **A:** yes **Q:** is the shelf full? **A:** yes **Q:** do the vegetables look fresh? **A:** yes **Q:** are there carrots? **A:** yes **Q:** are there cabbages? **A:** no **Q:** what kind of vegetable has the most number present? **A:** i don't know **Q:** are there labels on the shelf? **A:** yes **Q:** are there prices? **A:** yes	**Q:** is this an outdoor or indoor market? **A:** outdoors **Q:** does the shelf have many layers? **A:** no **Q:** is the picture in color? **A:** yes **Q:** is the shelf full? **A:** yes **Q:** do the vegetables look fresh? **A:** yes **Q:** are there carrots? **A:** yes **Q:** are there cabbages? **A:** no **Q:** what kind of vegetable has the most number present? **A:** sugar snap beans **Q:** are there labels on the shelf? **A:** yes **Q:** are there prices? **A:** yes
 a bird flies across a lake next to some buildings	**Q:** is the photo in color? **A:** yes **Q:** is there any people? **A:** no **Q:** how many birds? **A:** 1 **Q:** is it night? **A:** no **Q:** is the bird brown? **A:** no it's white **Q:** is it day? **A:** yes **Q:** are there trees? **A:** yes **Q:** is the sky visible? **A:** no **Q:** are there many buildings? **A:** a lot of them **Q:** are the trees lush? **A:** yes	**Q:** is the photo in color? **A:** yes **Q:** is there any people? **A:** no **Q:** how many birds? **A:** 1 **Q:** is it night? **A:** no **Q:** is the bird brown? **A:** no **Q:** is it day? **A:** yes **Q:** are there trees? **A:** yes **Q:** is the sky visible? **A:** no **Q:** are there many buildings? **A:** yes **Q:** are the trees lush? **A:** yes	**Q:** is the photo in color? **A:** yes **Q:** is there any people? **A:** no **Q:** how many birds? **A:** 1 **Q:** is it night? **A:** no it is daytime **Q:** is the bird brown? **A:** nope **Q:** is it day? **A:** yes **Q:** are there trees? **A:** yes **Q:** is the sky visible? **A:** yes **Q:** are there many buildings? **A:** yes **Q:** are the trees lush? **A:** yes	**Q:** is the photo in color? **A:** yes **Q:** is there any people? **A:** no **Q:** how many birds? **A:** 1 **Q:** is it night? **A:** no **Q:** is the bird brown? **A:** no **Q:** is it day? **A:** yes **Q:** are there trees? **A:** yes **Q:** is the sky visible? **A:** yes **Q:** are there many buildings? **A:** yes **Q:** are the trees lush? **A:** yes
 a bird flying over the ocean beach at sunrise	**Q:** is the photo in color? **A:** yes **Q:** can you see clouds? **A:** yes **Q:** what color is the ocean? **A:** grey **Q:** can you see the sand on the beach? **A:** yes **Q:** any people around? **A:** no **Q:** what color is the bird? **A:** white **Q:** do you see any boats? **A:** no **Q:** any other birds? **A:** 1 other **Q:** any jets or balloons? **A:** no **Q:** do you see buildings? **A:** no	**Q:** is the photo in color? **A:** yes **Q:** can you see clouds? **A:** yes **Q:** what color is the ocean? **A:** blue **Q:** can you see the sand on the beach? **A:** yes **Q:** any people around? **A:** no **Q:** what color is the bird? **A:** white **Q:** do you see any boats? **A:** no **Q:** any other birds? **A:** no **Q:** any jets or balloons? **A:** no **Q:** do you see buildings? **A:** no	**Q:** is the photo in color? **A:** yes it is **Q:** can you see clouds? **A:** yes **Q:** what color is the ocean? **A:** it is brownish gray **Q:** can you see the sand on the beach? **A:** yes **Q:** any people around? **A:** no **Q:** what color is the bird? **A:** white **Q:** do you see any boats? **A:** no **Q:** any other birds? **A:** no **Q:** any jets or balloons? **A:** no **Q:** do you see buildings? **A:** no	**Q:** is the photo in color? **A:** yes **Q:** can you see clouds? **A:** yes **Q:** what color is the ocean? **A:** greenish gray **Q:** can you see the sand on the beach? **A:** yes **Q:** any people around? **A:** no **Q:** what color is the bird? **A:** white **Q:** do you see any boats? **A:** no **Q:** any other birds? **A:** no **Q:** any jets or balloons? **A:** no **Q:** do you see buildings? **A:** no

Fig. 2: Qualitative samples for three model variants – ViLBERT w/ CC + VQA (called 'Base'), Base + CE, and Base + CE + NSP.